Protein Function Prediction from Genetic and Protein-Protein Interactions via Regularized Multi-Label Canonical Discriminant Analysis

Mu-Fen Hsieh[1], Chang-Fu Hsieh[3] and Sing-Hoi Sze[1,2]

[1]Department of Computer Science and Engineering,
[2]Department of Biochemistry & Biophysics,
Texas A&M University, College Station, TX 77843, USA

[3]Institute of Ecology and Evolutionary Biology,
National Taiwan University, Taipei, Taiwan

Abstract

We introduce multi-label canonical discriminant analysis (MCDA) and regularization techniques to tackle the problem of multi-label protein function prediction. Classical canonical discriminant analysis (CDA) is a variant of linear discriminant analysis (LDA) that has been shown to have optimality close to Bayes classifier for single-label classification. We generalize CDA to handle multi-label classification and show that it outperforms previous approaches on multi-label protein function prediction. In order to handle the singularity problem of the estimated covariance matrix, we utilize a two-step regularization technique that produces better results than utilizing pseudo-inverse or a single regularization. We transform distance matrices from genome-wide comparisons of protein sequences, genetic and protein-protein interactions into a reduced feature space by applying multi-dimensional scaling (MDS). With further dimension reduction from MCDA, proteins are positioned at canonical space that maximizes group separation. We show that our algorithm outperforms previous approaches in predicting function of yeast proteins by mapping them to categories in the MIPS Functional Catalogue (FunCat) and to Gene Ontology (GO) Slim terms. To facilitate biological investigations, protein variables can be projected and visualized on a two-dimensional plot of MDS feature space and of MCDA canonical space.

1 Introduction

Proteins are often annotated with one or more functions (i.e., multi-labeled), as seen in gene ontology (GO) (Ashburner *et al.*, 2000), enzyme classifications (Kanehisa *et al.*, 2004) and the MIPS Functional Catalogue (FunCat) (Mewes *et al.*, 2004). Many previous algorithms have been developed that utilize interaction network data to predict protein function (Vazquez *et al.*, 2003; Karaoz *et al.*, 2004; Nabieva *et al.*, 2005), while integrated approaches have also been developed to utilize more than one source of data (Lanckriet *et al.*, 2004; Tsuda and Noble, 2004; Mostafavi *et al.*, 2008). Multi-label classification methods consider association between groups and have been shown to outperform classical binary classifier (Wang *et al.*, 2015), which assumes that each group is independent and classifies one single group at a time.

We introduce multi-label canonical discriminant analysis (MCDA), which is a generalization of classical canonical discriminant analysis (CDA). Since CDA has the capability to handle distance

matrices when combined with multi-dimensional scaling (MDS), it has been applied extensively in many areas, including ecological land classification (Pouyat *et al.*, 2007) and biogeography (Mallet *et al.*, 2014). Observations include typical numerical features, such as soil properties, weather conditions, terrain types, and similarity or distance matrices that represent geographical correlations between species. Despite its popularity in these areas, CDA has not yet been applied to protein function prediction.

CDA is a variant of linear discriminant analysis (LDA) that takes a set of m multivariate Gaussian population groups $g_1, g_2, \ldots, g_m$ with means $\mu_1, \mu_2, \ldots, \mu_m$ and a common covariance matrix Σ as input, and predicts the probability of x belonging to each group g_j, where each variable x can belong to only one group g_j. Discriminant functions are derived through eigen-decomposition of the estimated common covariance matrix by considering the product of the between-class and within-class covariance matrices to maximize the separation of population groups at transformed space, where a p-dimensional variable x is transformed to a q-dimensional variable z for $q \leq p$. The probability of a variable z belonging to a group g_j is predicted based on its Mahalanobis distance to group centroids. CDA differs from LDA on how the common covariance matrix is estimated, with CDA relying on multivariate analysis of variance (MANOVA).

Our multi-label CDA takes the same input as single-label CDA, except that each variable may belong to one or more groups among $g_1, g_2, \ldots, g_m$. Given n variables $x_1, x_2, \ldots, x_n$, group membership is represented by a design matrix D, where $D_{ij} = 1$ indicates that variable x_i belongs to group g_j. While a previous multi-label LDA approach (Wang *et al.*, 2010) computes a between-class covariance matrix and a pooled within-class covariance matrix by summing up the corresponding covariance matrices of each class in both cases, our multi-label CDA computes a single within-class covariance matrix by MANOVA as in single-label CDA.

One difficulty of the approach is that when the number of classification features is larger than the cardinality of any group or when the correlation between features is high, collinearity problem may occur that results in a singular covariance matrix. In addition to applying pseudo-inverse to the covariance matrix, a commonly used regularization technique can be applied to stablize the covariance matrix and reduce the bias of discriminant functions (Guo *et al.*, 2007). In contrast to regularized LDA, we apply a two-step regularization in MCDA, which includes ridge regression (Hoerl and Kennard, 1970) and adjustment of the estimated within-class covariance matrix.

We apply MCDA to the multi-label protein function prediction problem and compare our performance to a previous multi-label protein function prediction algorithm that utilizes a maximization of data-knowledge consistency (MDKC) approach to consider functional group correlation by solving a non-negativity matrix factorization problem (Wang *et al.*, 2015).

2 Methods

2.1 Multi-label protein function prediction

We define the multi-label protein function prediction problem as follows. Given m protein functional groups $g_1, g_2, \ldots, g_m$ and a set S of n proteins, each of which belongs to one or more of these groups, predict the probability of $x \in S$ belonging to each group g_j. Here S contains both annotated and unannotated proteins, and the function of unannotated proteins is predicted by grouping them together with annotated proteins. Observations include protein sequences, genetic interaction networks and protein-protein interaction networks.

2.2 Distance matrices and multi-dimensional scaling (MDS)

While various pairwise distance measures have been developed for genes, proteins or species in phylogeny, the adjacency matrix can be viewed as a basic similarity matrix in a genetic or protein-protein interaction network. A measure of topological similarity was also developed in networks (Pei and Zhang, 2005).

In our protein function prediction algorithm, we define distance matrices from sequence and interaction network comparisons. Sequence distances are computed from Smith-Waterman local alignment (Smith and Waterman, 1981), while interaction distances are defined as Sørensen-Dice dissimilarity or Jaccard distance of interaction partners, following the observation in Bandyopadhyay *et al.* (2006) that two proteins are likely to have the same function if their interaction partners are conserved.

In order to apply discriminant analysis, we perform multi-dimensional scaling (MDS) to transform a distance matrix into feature vectors in a low-dimensional space. Given a double-centered distance matrix Δ that is decomposed into a matrix Λ of eigenvalues at the diagonal and eigenvectors V, MDS variables $X_p = V_p \Lambda_p^{1/2} R$ of the first p axes are chosen by minimizing $\mathrm{tr}((\Delta - X_p X_p^T)^2)$ subject to $RR^T = I$ (Mardia, 1978).

2.3 Canonical discriminant analysis (CDA)

Given m population groups $g_1, g_2, \ldots, g_m$ and n variables $x_1, x_2, \ldots, x_n$ each of dimension p, CDA finds a transformation matrix C of canonical coefficients so that $Z = XC^T$, where X represents the original variables x_i of dimension p and Z represents the transformed variables z_i of dimension q (Seal, 1964). The transformation matrix C is defined to be the one that maximizes

$$C\Sigma_B C^T \qquad \text{subject to} \qquad C\Sigma_W C^T = I$$

and

$$(C\Sigma_W^{\frac{1}{2}})\Sigma_W^{-\frac{1}{2}}\Sigma_B\Sigma_W^{-\frac{1}{2}}(C\Sigma_W^{\frac{1}{2}})^T = F\Sigma_W^{-\frac{1}{2}}\Sigma_B\Sigma_W^{-\frac{1}{2}}F^T \qquad \text{subject to} \qquad FF^T = I$$

where Σ_B is the between-class covariance matrix and Σ_W is the within-class covariance matrix. The common covariance matrix is $\Sigma = F\Sigma_W^{-\frac{1}{2}}F^T$ at feature space and $\Sigma_C = I$ at canonical space.

By solving $\Sigma_B\Sigma_W^{-1}$, eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_q$ and eigenvectors $f_1, f_2, \ldots, f_q$ can be computed. The canonical coefficients can be computed by $C = \Sigma_W^{-\frac{1}{2}}F$ and canonical correlation for the kth axis is defined by $\rho_k = \sqrt{\lambda_k/(1+\lambda_k)}$.

Given prior probabilities $q_1, q_2, \ldots, q_m$ for the m population groups, the probability that the canonical variable z_i belongs to group g_j can be expressed as a normalized exponential function of the Mahalanobis distance d^2:

$$\Big(q_j \exp\Big(-\frac{1}{2}d^2(z_i, \hat{\mu}_{C,j})\Big)\Big) \Big/ \sum_{k=1}^{m}\Big(q_k \exp\Big(-\frac{1}{2}d^2(z_i, \hat{\mu}_{C,k})\Big)\Big)$$

where $d^2(z_i, \hat{\mu}_{C,k}) = (z_i - \hat{\mu}_{C,k})^T\hat{\Sigma}_C^{-1}(z_i - \hat{\mu}_{C,k})$, and the canonical mean $\hat{\mu}_{C,k}$ is estimated by the centroid of group g_k at canonical space, with the assumption that $\hat{\Sigma}_C^{-1} = I$.

2.4 Covariance matrix estimation from MANOVA

Instead of solving $\Sigma_B\Sigma_W^{-1}$ directly, CDA considers the multivariate analog of the F statistic $SS_H SS_E^{-1}$, where SS_H and SS_E are the multivariate analogs of the sum of squares so that $\Sigma_B =$

$SS_H/(m-1)$ and $\Sigma_W = SS_E/(n-m)$. SS_H and SS_E can be computed from MANOVA through the multivariate test of equal means.

By assuming a multivariate general linear model (GLM) $X = D^T\beta + \epsilon$ for our observations X along with a design matrix D, where $D_{ij} = 1$ indicates that x_i belongs to group g_j, the null hypothesis can be written as $H_0 : H^{(k)}\beta = 0$, where $H^{(k)}$ is the hypothesis matrix of 0's and 1's that results in equal means for group g_k (Friendly, 2007).

We can compute the estimates $\hat{SS}_E = \hat{\epsilon}^T\hat{\epsilon}$ and

$$\hat{SS}_H^{(k)} = (H^{(k)}\hat{\beta})^T\Big(H^{(k)}(D^TD)^{-1}(H^{(k)})^T\Big)^{-1}(H^{(k)}\hat{\beta})$$

for group g_k (Timm, 1975). We follow multi-label LDA (Wang *et al.*, 2010) and define the multi-label sum of squares for hypothesis as $\hat{SS}_H = \sum_{k=1}^m \hat{SS}_H^{(k)}$.

2.5 Two-step regularization for multi-label canonical discriminant analysis

With least-squares estimation, the model coefficients can be computed as $\hat{\beta} = (D^TD)^{-1}D^TX$. Since there is often a singularity problem for $(D^TD)^{-1}$, ridge regression (Hoerl and Kennard, 1970) is applied to regularize it:

$$\hat{\beta}^{\text{ridge}} = (D^TD + \lambda^{\text{ridge}}I)^{-1}D^TX = (D_r^TD_r)^{-1}D^TX$$

where $\hat{\beta}^{\text{ridge}}$ is computed from ridge regression by choosing λ^{ridge} to minimize the generalized cross validation (GCV) error (Wahba, 1990), and $(D_r^TD_r)^{-1}$ is obtained by right-matrix division. The resulting sum of squares for hypothesis is

$$(\hat{SS}_H^{\text{ridge}})^{(k)} = (H^{(k)}\hat{\beta}^{\text{ridge}})^T\Big(H^{(k)}(D_r^TD_r)^{-1}(H^{(k)})^T\Big)^{-1}(H^{(k)}\hat{\beta}^{\text{ridge}}).$$

We follow Guo *et al.* (2007) to reduce bias and apply regularization on $\hat{SS}_E$ by modifying the sample correlation matrix $\hat{R} = \hat{d}^{-\frac{1}{2}}\hat{SS}_E\hat{d}^{-\frac{1}{2}}$, where $\hat{d} = \text{diag}(\hat{SS}_E)$, by the formula $\hat{R}_r = \lambda\hat{R} + (1-\lambda)I$ to obtain $(\hat{SS}_E)_r = \hat{d}^{\frac{1}{2}}\hat{R}_r\hat{d}^{\frac{1}{2}}$. We choose λ to maximize the first canonical correlation ρ_1 since it often results in better prediction. The eigen-decomposition of $\hat{SS}_H^{\text{ridge}}(\hat{SS}_E)_r^{-1}$ is solved as in CDA.

3 Results

3.1 Evaluation

Yeast is among one of the species that currently provides the most abundant genetic and protein-protein interaction data. For the purpose of comparisons, we used the same BioGRID interaction data (Stark *et al.*, 2006) and MIPS Functional Catalogue (FunCat) (Mewes *et al.*, 2004) as in Wang *et al.* (2015) for classification of yeast proteins. The first level of FunCat was used, which includes 17 high-level protein functions. On average, each protein belongs to 2.39 FunCat categories.

We also utilized Gene Ontology (GO) Slim terms created by the *Saccharomyces* Genome Database (SGD) (Cherry *et al.*, 1998), which is a high-level subset of Gene Ontology. In total, 91 GO Slim terms were chosen after removing the terms that annotate fewer than 30 or more than 500 of the yeast proteins.

We followed Wang *et al.* (2015) and removed proteins with no functional annotation or with only one interaction. For FunCat classification, there are 4292 proteins with 47492 genetic interactions

Table 1: Effect of using different number of MDS variables during the transformation on five-fold prediction performance of regularized MCDA on *S. cerevisiae* integrated data over 17 MIPS FunCat level-1 categories. (a) Average of best F1 scores (%). (b) Average AUC-ROC (%).

#MDS variables	(a) F1 score	(b) AUC-ROC
10	35.36±0.51	69.20±0.42
20	37.58±0.52	70.97±0.42
50	42.47±0.75	73.40±0.52
100	45.58±0.77	74.36±0.42
200	46.93±0.62	75.56±0.54
500	**48.85±0.55**	**77.02±0.38**
1000	47.73±0.60	75.57±0.46
2000	42.94±0.40	72.51±0.42

and 53393 protein-protein interactions, and 98142 interactions when the two types of interactions are mixed together. For GO Slim classifications, there are 3698 proteins with 43893 genetic interactions and 32446 protein-protein interactions, and 73512 interactions when the two types of interactions are mixed together. Protein sequences were retrieved from SGD (Cherry *et al.*, 1998).

To assess the performance of our protein function prediction algorithm, we perform five-fold cross validation on each set of proteins. To create integrated data that combine the effect of different types of data, we consider a linear combination of sequence, genetic and protein-protein interaction distance matrices with weights. Possible ranges of the regularization parameters are $10^{-2} \leq \lambda^{\text{ridge}} \leq 10^{10}$ and $0 \leq \lambda \leq 1$, and linear searches can be performed over a range of values to determine the best values. Table 1 shows that a good value to use for the number of MDS variables is 500 on the MIPS FunCat dataset. For the SGD GO Slim dataset, MCDA is applied twice to further separate the functional groups to obtain improved performance.

In order to extract a set of predicted proteins from the output of MCDA, we impose a threshold on the probability estimate of each protein belonging to each functional category. We compute the best F1 score and the receiver operating characteristic curve (ROC) over a range of thresholds and obtain the average of best F1 scores and the average area under the ROC curve (AUC-ROC) over the functional categories. We report each performance value as a range within one standard error over five-fold prediction.

3.2 Predictions from protein sequences and interaction networks

Table 2 shows that prediction based on interaction partner dissimilarity has better performance than sequence alignment distances for both FunCat categories and GO Slim terms, with the protein-protein interaction (PPI) results better than the genetic interaction (GI) results. A similar trend is observed based on the average of alignment distance and interaction dissimilarity, with the results of combining sequence and PPI matrices better than the results of combining sequence and GI matrices.

Among interaction dissimilarity measures, Jaccard distance produces better prediction than Sørensen-Dice dissimilarity for both GI and PPI networks, but Sørensen-Dice dissimilarity is better on combined sequence and interaction distances.

While performance of MDKC on Kullback-Leibler divergence of amino acid distributions is significantly better than MCDA on sequence alignment distances, MCDA has better performance

Table 2: Comparison of distance/similarity features computed from *S. cerevisiae* protein sequences and interaction networks, and five-fold prediction performance of regularized MCDA and MDKC. (a) Average of best F1 scores over 17 MIPS FunCat level-1 categories (%). (b) Average AUC-ROC over 17 MIPS FunCat level-1 categories (%). (c) Average of best F1 scores over 91 SGD GO slim terms (%). (d) Average AUC-ROC over 91 SGD GO slim terms (%).

	MIPS FunCat		SGD GO Slim	
Distance/similarity	(a) F1 score	(b) AUC-ROC	(c) F1 score	(d) AUC-ROC
SW	33.11±0.33	62.47±0.30	16.21±0.14	61.08±0.35
KL (by MDKC)	**42.17**[1]			
SD-GI	**38.83±0.86**	69.94±0.56	30.85±0.52	76.47±0.44
Jc-GI	38.73±0.90	**70.25±0.61**		
SD-GI-1hop	35.22±0.32	66.31±0.33		
SD-PPI	41.70±0.39	72.52±0.34	36.61±0.57	77.62±0.39
Jc-PPI	**42.38±0.53**	**72.96±0.28**		
SD-PPI-1hop	37.55±0.28	69.84±0.64		
SD-GI/PPI	45.27±0.74	75.07±0.62	39.48±0.48	81.78±0.45
Jc-GI/PPI	**45.75±0.64**	**75.22±0.63**		
TM-GI/PPI (by MDKC)	40.03[1,2]			
SW + SD-GI[3]	**43.40±0.58**	**72.85±0.59**	32.31±0.49	78.22±0.64
SW + Jc-GI[3]	42.42±0.46	72.23±0.47		
SW + SD-PPI[3]	**43.75±0.56**	**73.68±0.12**	36.98±0.56	78.52±0.34
SW + Jc-PPI[3]	42.91±0.46	73.26±0.12		
SW + SD-GI/PPI[3]	47.04±0.34	75.85±0.42	40.35±0.43	**82.33±0.54**
SW + Jc-GI/PPI[3]	45.90±0.44	74.97±0.22		
SW + SD-GI + SD-PPI[4]	**49.50±0.42**	**77.00±0.51**	**40.92±0.37**	81.28±0.49
SW + Jc-GI + Jc-PPI[4]	46.93±0.39	76.18±0.26		
KL + SD-GI/PPI[4] (by MDKC)	≤45[5]		40.11[1,6]	

SW: One minus log odd score from Smith-Waterman sequence alignment
KL: Kullback-Leibler divergence of 3-amino-acid distributions
SD: Sørensen-Dice dissimilarity
Jc: Jaccard distance
GI: Genetic interactions
PPI: Protein-protein interactions
GI/PPI: Mix of genetic and protein-protein interactions
1hop: Interaction partners at one hop distance are included
TM: Topological Measurement similarity (Pei and Zhang, 2005)
1: Reported by Wang *et al.* (2015)
2: Prediction by randomly splitting the data into two halves (Wang *et al.*, 2015)
3: Average of sequence and interaction distance matrices
4: Linear combination of sequence and interaction distance matrices with weights
5: Performance estimated from Figure 3b in Wang *et al.* (2015) for MDKC and other algorithms
6: Classification over 90 random GO terms across three domains (Wang *et al.*, 2015)

Table 3: Five-fold prediction performance of regularized and non-regularized MCDA on *S. cerevisiae* integrated data over 17 MIPS FunCat level-1 categories. (a) Average of best F1 scores (%). (b) Average AUC-ROC (%).

Ridge	RegCo	MCDA	Prior	Corr	(a) F1 score	(b) AUC-ROC
		×			47.18	75.04
×		×			47.43	75.43
	×	×			49.00	76.48
×	×	×			**49.50**	**77.00**
×	×	×	×		45.87	73.76
×	×	×		×	37.96	67.54

Ridge: Ridge regression
RegCo: Regularization of within-class covariance matrix
MCDA: Multi-label canonical discriminant analysis
Prior: Prior of function distribution
Corr: Cosine function-function correlation (Wang *et al.*, 2015)

based on dissimilarity of interaction partners than MDKC on topological similarity. When sequence and interaction distances are combined, MCDA has better performance than MDKC.

While MDKC was tested on mixed GI and PPI networks, we tried two options: using one single matrix on mixed GI and PPI networks, and computing distance matrices separately for GI and PPI networks. While prediction based on a GI or PPI network alone does not perform as well probably due to less information, separate distance matrices are preferred when both GI and PPI networks are used together for both FunCat categories and GO Slim terms. In contrast, better AUC-ROC is obtained over GO Slim terms when combining sequence and mixed GI and PPI networks.

3.3 Effect of regularization and MCDA

Table 3 shows that with each step of the two-step regularization, the F1 score improves further. The use of prior information of protein function distribution to MCDA does not give improved performance. Wang *et al.* (2015) introduced function correlation to the multi-label protein function prediction problem. We applied their cosine product to the design matrix before running MCDA, but the performance drops significantly. Thus the cosine function-function correlation is not a suitable measure for MCDA.

3.4 Performance by category

Table 4 shows that regularized MCDA outperforms MDKC for most FunCat categories, especially the ones with more than 500 proteins. There are two categories for which MCDA shows significantly lower F1 score, including 34 (environment) and 41 (development). Figure 1 further shows that MCDA generally does not perform as well in those categories X that have large protein overlap with some other categories, especially for 41 (development).

Through dimension reduction of MDS, protein groups can be drawn on a two-dimensional plot that retains the original distance information. When transformed with the canonical coefficients, groups are further separated on different axes. Figure 2 shows that on the first two axes of canonical space, FunCat groups 10, 11, 12 and 38 stand out when compared to the same groups at MDS

Table 4: Average of best F1 scores (%) of five-fold prediction on *S. cerevisiae* integrated data over 17 MIPS FunCat level-1 categories by (a) regularized MCDA; (b) MDKC and other algorithms in Wang *et al.* (2015).

ID	Description	#proteins	(a) MCDA	(b) MDKC[1]
01	metabolism	1390	**60.14**	≤40
02	energy	327	44.53	≤**45**
10	cell cycle and DNA processing	965	**62.85**	≤∼60
11	transcription	995	**68.37**	≤60
12	protein synthesis	460	**71.52**	≤70
14	protein fate	1106	**53.33**	≤40
16	protein with binding function	1008	**44.15**	≤30
18	regulation of metabolism	240	**32.41**	≤20
20	cellular transport	974	**63.47**	≤∼60
30	cellular communication	230	**51.17**	≤∼50
32	cell rescue, defense and virulence	509	**38.37**	≤∼30
34	interaction with the environment	443	33.45	≤∼**45**
38	transposable elements	29	**65.59**	≤∼45
40	cell fate	264	44.08	≤∼**45**
41	development	66	16.14	≤**30**
42	biogenesis of cellular components	822	**45.79**	≤∼45
43	cell type differentiation	430	**46.16**	≤∼45
All		10258	**49.50**	≤45

1: Performance estimated from Figure 3b in Wang *et al.* (2015)

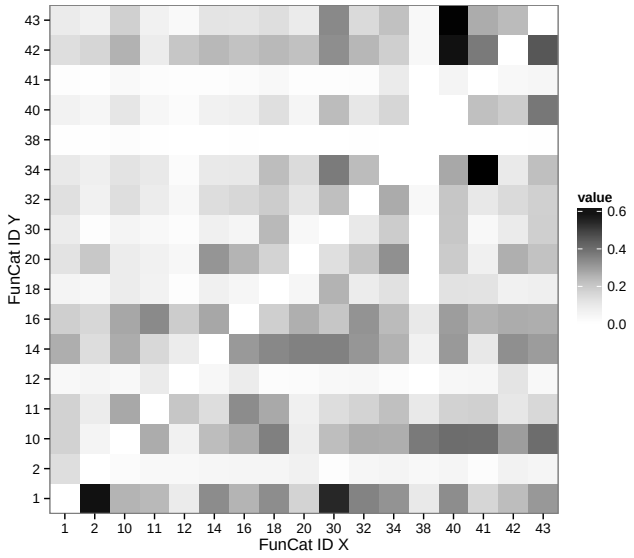


Figure 1: Protein overlap between category X and category Y relative to the size of category X among MIPS FunCat level-1 categories (%). Statistics are not shown at the lower left to upper right diagonal with identical category pairs.

space, and MCDA outperforms MDKC in these categories. Group 38 is among one of the outlier groups that MCDA outperforms the other approaches the most (this is also true for group 01).

Using MANOVA Wilks' test on normality, only the MDS variables from genetic interaction network that belong to or not belong to group 38 (transposable elements) are Gaussian. Since many other groups with non-Gaussian variables are predicted well, MCDA demonstrates robustness to assumption violation.

When looking closely at Figure 2, while groups 01 and 14 are not separated on the first axis at canonical space, they are well separated on the second axis and are predicted well. Also note that the centroids of groups 34 and 41 are close to the middle region with respect to the other groups at canonical space. Since they are not predicted well with F1 score smaller than 35%, they are probably not separated on subsequent axes.

4 Discussion

We have developed a generalized version of canonical discriminant analysis (CDA) that solves the multi-label problem in protein function prediction. We show that by applying dimension reduction techniques MDS and MCDA to transform the distance matrices, very good performance can be obtained. By choosing suitable distance measures such as Sørensen-Dice dissimilarity, prediction performance can be further improved. More accurate predictions can be obtained when genetic and protein-protein interaction data are considered separately.

By applying regularization to two covariance matrices to correct bias, we solve the singularity problem and further improve the performance. Regularized MCDA outperforms other approaches, especially for functional categories that contain a large number of proteins. In contrast, the performance of MCDA can be sensitive to large protein overlap among functional groups. By visualizing distance information on two-dimensional plots through the MDS and MCDA techniques, it is possible to identify important outlier groups.

5 Acknowledgments

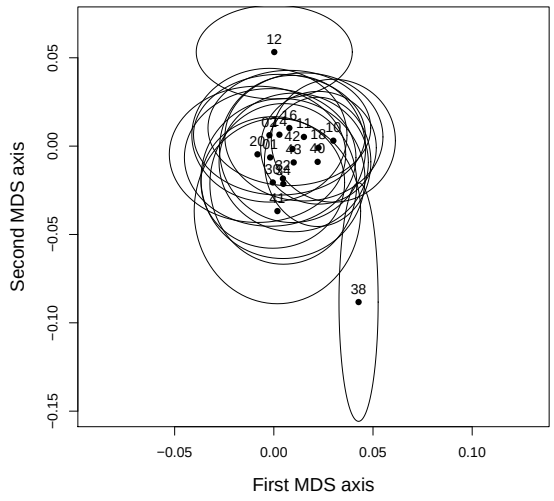
We thank Kuang-Yu Chang and Huei-Fang Yang from Institute of Information Science, Academia Sinica, Taiwan for providing literature. Computations were performed on the Brazos Cluster at Texas A&M University. This work was supported in part by the National Institute of Justice (2012-DN-BX-K024). Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice.

References

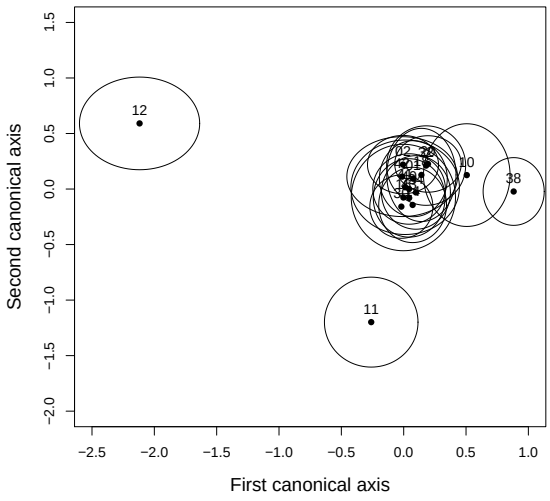
[1] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

[2] Bandyopadhyay, S., Sharan, R. and Ideker, T. (2006) Systematic identification of functional orthologs based on protein network comparison. *Genome Res.*, **16**, 428–435.

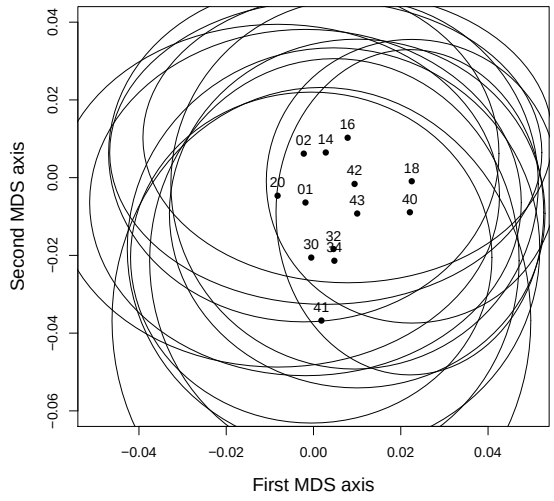
[3] Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S. and Botstein, D. (1998) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.*, **26**, 73–79.



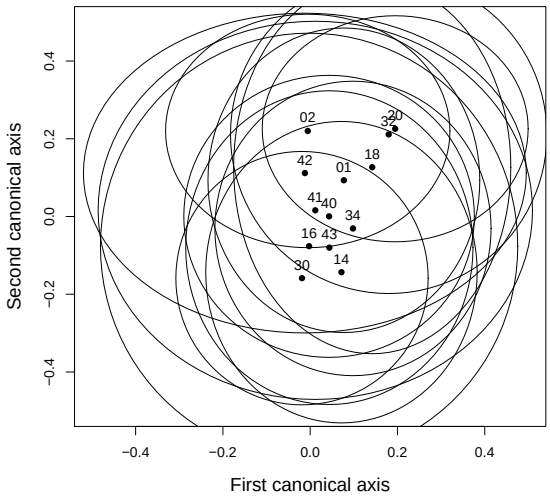
(a) FunCat groups at MDS space



(b) FunCat groups at canonical space



(c) Partial FunCat groups at MDS space



(d) Partial FunCat groups at canonical space

Figure 2: Visualization of 17 FunCat groups on *S. cerevisiae* integrated data, showing group centroids and 90% confidence circle on the first two axes of (a) MDS feature space; (b) MCDA canonical space; (c) MDS feature space showing a subset of categories that exclude groups 10, 11, 12 and 38; (d) MCDA canonical space showing a subset of categories that exclude groups 10, 11, 12 and 38.

[4] Friendly, M. (2007) HE plots for multivariate linear models. *J. Comput. Graph. Statist.*, **16**, 421–444.

[5] Guo, Y., Hastie, T. and Tibshirani, R. (2007) Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, **8**, 86–100.

[6] Hoerl, A.E. and Kennard, R.W. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

[7] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–280.

[8] Karaoz, U., Murali, T.M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C.R. and Kasif, S. (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl. Acad. Sci. USA*, **101**, 2888–2893.

[9] Lanckriet, G.R.G., Deng, M., Cristianini, N., Jordan, M.I. and Noble, W.S. (2004) Kernel-based data fusion and its application to protein function prediction in yeast. *Pac. Symp. Biocomput.*, **9**, 300–311.

[10] Mallet, B., Martos, F., Blambert, L., Pailler, T. and Humeau, L. (2014) Evidence for isolation-by-habitat among populations of an epiphytic orchid species on a small oceanic island. *PLoS One*, **9**, e87469.

[11] Mardia, K.V. (1978) Some properties of classical multi-dimensional scaling. *Commun. Statist. Theor. Meth.*, **7**, 1233–1241.

[12] Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Güldener, U., Mannhaupt, G., Münsterkötter, M., Pagel, P., Strack, N., Stümpflen, V., Warfsmann, J. and Ruepp, A. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–44.

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

[14] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. and Singh, M. (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, **21(Suppl. 1)**, i302–310.

[15] Pei, P. and Zhang, A. (2005) A topological measurement for weighted protein interaction network. *Proc. IEEE Comput. Syst. Bioinform. Conf.*, 268–278.

[16] Pouyat, R.V., Yesilonis, I.D., Russell-Anelli, J. and Neerchal, N.K. (2007) Soil chemical and physical properties that differentiate urban land-use and cover types. *Soil Sci. Soc. Am. J.*, **71**, 1010–1019.

[17] Seal, H.L. (1964) *Multivariate Statistical Analysis for Biologists*. Methuen.

[18] Smith, T.F., and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

[19] Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–539.

[20] Timm, N.H. (1975) *Multivariate Analysis, with Applications in Education and Psychology.* Brooks/Cole Pub. Co.

[21] Tsuda, K. and Noble, W.S. (2004) Learning kernels from biological networks by maximizing entropy. *Bioinformatics*, **20(Suppl. 1)**, i326–333.

[22] Vazquez, A., Flammini, A., Maritan, A. and Vespignani, A. (2003) Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.*, **21**, 697–700.

[23] Wahba, G. (1990) *Spline Models for Observational Data*. SIAM.

[24] Wang, H., Huang, H. and Ding, C. (2010) Multi-label linear discriminant analysis. *Proc. Euro. Conf. Comput. Vision*, 126–139.

[25] Wang, H., Huang, H. and Ding, C. (2015) Correlated protein function prediction via maximization of data-knowledge consistency. *J. Comput. Biol.*, **22**, 546–562.