Whole-Genome Functional Classification of Genes by Latent Semantic Analysis on Microarray Data

See-Kiong Ng* **Zexuan Zhu†** **Yew-Soon Ong†**

*Knowledge Discovery Department, Institute for Infocomm Research,
21 Heng Mui Keng Terrace, Singapore 119613.
†School of Computer Engineering, Nanyang Technological University,
Blk N4, #02a-32 Nanyang Avenue, Singapore 639798.
Email: sknq@i2r.a-star.edu.sg, zhuzexuan@pmail.ntu.edu.sg, asysong@ntu.edu.sg

Abstract

Quantitative simultaneous monitoring of the expression levels of thousands of genes under various experimental conditions is now possible using microarray experiments. The resulting microarray data are very useful for elucidating the functional relationships among genes in the genomes. However, due to the experimental and biological nature of the data, whole-genome functional classification of genes on microarray data remains a challenging machine learning problem. In this paper, we introduce the application of latent semantic analysis (LSA) to microarray expression data for systematic, genome-wide functional classification of genes.

In the LSA approach considered here, singular value decomposition is first applied as a dimension-reducing step on the gene expression data, followed by an unsupervised clustering procedure based on vector similarities in the truncated space. Functional classification is then conducted through calling by majority on each of the resulting gene clusters. Using this semi-supervised LSA approach on microarray data, we have performed systematic functional classification on the genes in the partially-annotated yeast genome, annotating more than 1,700 unknown genes into 40 distinct functional classes with promising results.

Keywords: whole-genome gene functional classification, microarray data analysis, latent semantic analysis, singular value decomposition.

1 Introduction

DNA microarray technology has enabled the quantitative measurement of thousands of gene expression levels simultaneously. Through the use of this technology, it is possible for molecular biologists to study the differential gene expression across a set of related assays. Being a high throughput technology, it allows whole genomes to be scanned, generating thousands of data points per microarray experiment. Analysis of these genomic data allows for new ways of looking at the biology of living organisms.

To reveal the various functions of the genes in a genome, the gene expression profiles of a series of experimental assays or conditions can be analyzed to group the genes into clusters based on the similarity in their patterns of expression. These co-expression clusters can then be inferred as biological functional groupings for the genes—each cluster containing genes that encode proteins required for a common function. The functions of unknown gene products can then be systematically identified through the *guilt-by-association* principle (Walker, Volkmuth, Sprinzak et al. 1999).

However, the learning of gene functional classes from whole-genome microarray expression data is not an easy one, even for sophisticated machine learning algorithms such as support vector machines (Brown, Grundy, Lin et al. 2000) and multi-layer perceptrons (Mateos, Dopazo, Jansen et al. 2002). Some inherent problems include:

1. For genome-wide functional analysis, the number of experimental conditions (the *"features"*) is easily out-numbered by the number of genes in the genome. Typically, there are several thousands of genes but only tens or hundreds of different experimental assays or conditions.

2. There is a large number of functional classes to be learned. For example, there are ∼100 functional classes cataloged in the MIPS database (Mewes, Frishman, Guldener et al. 2002). As a result, an inherent problem for *whole-genome* functional classification is the imbalance in the number of positive and negative examples with respect to each function class. The noise in the relatively large proportion of the negative training examples can easily outweigh the small number of positive examples in each class, making it difficult for machine learning. In fact, Mateos et al. (2002) found that only ∼10% of the gene functional classes are machine-learnable.

We propose here the application of latent semantic analysis (LSA) to the problem of systematic, genome-wide functional classification of genes from microarray expression data. In LSA, singular value decomposition (SVD) is first applied as a dimension-reducing step on the gene expression data, followed by an unsupervised clustering procedure to group genes with similar expression in the truncated gene expression space. Classification is then conducted with functional assignment by majority voting in each of the resulting gene clusters.

The use of dimension reduction by SVD helps to de-noise the data as well as enable the clustering algorithm to focus only on significant components present in the expression data. The unsupervised pre-classification clustering facilitates the grouping of all classes globally, making the procedure less susceptible to the imbalance of training examples by individual classes mentioned earlier. Using this semi-supervised LSA approach on microarray data, we have performed systematic functional classification on the genes in the partially-annotated yeast genome and annotated more than 1,700 under 40 distinct functional classes from the Comprehensive Yeast Genome

Database (CYGD) available from Munich Information Centre for Protein Sequences (MIPS) (Mewes et al. 2002).

The rest of this paper is organized as follows: in Section 2, we provide some background information on gene clustering and classification, and on latent semantic analysis. We describe the data used and our method in details in Section 3. Finally, in Section 4, we present the performance evaluation of our classification approach on whole-genome yeast gene classification, and we discuss, in Section 5, how the LSA approach handles the various difficulties in whole-genome functional classification.

2 Background

In this section, we provide the background information for the various key concepts in this paper. First, we consider the differences between gene clustering and classification using supervised and unsupervised approaches. We then provide some background information on latent semantic analysis and singular value decomposition, and then we describe related work on using LSA or SVD for the analysis of gene expression data.

2.1 Gene Clustering and Classification

The living cell is a complex system comprising multiple cellular pathways that performs different biological functions. Through genome-wide measurements of the mRNA expression levels across different experimental conditions, we can construct a global map of the functional associations of the various genes in the genome based on their differential expression patterns under different conditions. This approach is called *gene clustering*—it involves the process of organizing genes into different functional groups using a similarity (or distance) measure on the gene expression data, but using no prior knowledge of the true functional classes of the genes. Gene clustering employs unsupervised machine learning techniques such as self-organizing maps (Tamayo, Slonim, Mesirov et al. 1999, Toronen, Kolehmainen, Wong & Castren 1999) and hierarchical clustering (Eisen, Spellman, Brown & Botstein 1998) to learn directly from the expression data.

When we have prior information about the functions of some of the genes, supervised approaches may be used. In fact, biologists often already knew a subset of genes involved in a biological pathway of interest. Such domain knowledge can be used—in the form of training sets—in *gene classification* for supervised machine learning techniques such as support vector machines (Brown et al. 2000) and neural networks (Mateos et al. 2002). However, as mentioned earlier, for whole-genome functional classification, these supervised techniques can suffer from the inherent imbalance in the positive and negative training examples for each of the functional classes, as the total number of functional classes in a living cell is large. Researchers have attempted to combat this problem by refining the innards of the machine learning algorithms—for example, Brown *et al.* (2000) modified the kernel values for their support vector machines to adjust for the misproportions in the positive and negative training examples.

In this paper, we adopt a combination of unsupervised clustering approach followed by a calling-by-majority semi-supervised approach to perform genome-wide multi-class functional annotation of genes. By avoiding using prior classification information, the unsupervised clustering is unaffected by the positive-negative population imbalance in the predefined classes. However, in the absence of such prior information, an incompetent clustering algorithm would generate clusters of genes that do not correspond well to the true underlying functional classes. As such, we use singular value decomposition to mathematically pre-analyze the expression data. We will show in Section 4 that this unsupervised clustering approach based on the latent semantic analysis approach (that have been previously proven successful in text mining applications) can indeed cluster the data with accuracy without supervision.

2.2 Latent Semantic Analysis

Latent semantic analysis, or LSA, is a popular analysis method in the text mining community. LSA uses singular value decomposition to map the high-dimensional word-document frequency count matrix to lower-dimensional latent "semantic" space wherein text terms and documents that are closely associated are placed near one another (Deerwester, Dumais, Furnas et al. 1990, Landauer & Dumais 1997). Dimension reduction can then be carried out as a preprocessing step, followed by clustering in the resulting truncated space. In this paper, we map this text mining approach into gene expression analysis by noting the similarity between the word-document count matrix and gene-sample expression data matrix—with "text terms" corresponding to genes, and "documents" corresponding to a sample (or an expression experiment). We can therefore apply LSA in a similar fashion for microarray data analysis as in text mining.

2.2.1 Singular Value Decomposition

LSA uses singular value decomposition (SVD) as a dimensionality reduction technique. In SVD, any $m \times n$ gene expression matrix A (i.e., m genes and n experiment samples, where $m > n$ typically) can be decomposed into a product of three other component matrices in the relation $A = U \cdot W \cdot V^T$, where:

- The component $m \times n$ matrix U describes the original row entities in A—i.e. the genes—as vectors of derived orthogonal column values (called the *"eigensamples"*), while the $n \times n$ matrix V describes the experimental samples (the original column entities in A) in terms of the so-called *"eigengenes"*; and
- The third component matrix W is an $n \times n$ diagonal matrix containing n scaling values indicating the relative significance of the eigenvectors.

Using SVD, we can reduce the dimensionality of the problem space simply by ignoring the insignificant coefficients in the diagonal matrix W. The reconstructed matrix is a least-squares best fit that uses fewer than the number of components present in the original data.

2.2.2 LSA and Gene Expression Analysis

In text mining, LSA involves the application of SVD with dimension reduction in order to reveal the underlying semantic components. For gene expression analysis, Alter *et al.*(2000, 2001) have also shown that much of the expression information in the microarray data can be captured by several significant eigenvectors, indicating the suitability of dimensional reduction with SVD in gene expression data analysis. In fact, they have found that some of the significant eigenvectors indeed represented independent biological and experimental processes that contributed to

the overall expression. Their observation suggests that SVD can be useful for revealing the implicit higher-order structure—such as the functional structures of the genes—in the gene expression data. In this aspect, the dimension reduction step in LSA then constitutes an inductive step by which genes are represented by values on a smaller set of abstract features such as their functional classes—rather than their raw patterns of observed expression levels in the various samples. Their decoupling by SVD therefore allows us to uncover the underlying functional classification map as expressed by the genes.

2.3 Related Work

There has been a recent interest in the use of SVD and related approaches for analyzing microarray expression data. Most of the work have been focused on applying SVD to mathematically discover the underlying components in the microarray expression data that have corresponding biological significance. For example, using Principal Component Analysis (similar to SVD), Raychaudhuri *et al.* (2000) demonstrated that the mathematical components they discovered in the time series sporulation expression data corresponded to significant biological subprocesses. Holter *et al.* (2000) used SVD to uncover underlying patterns or the so-called "characteristic modes" from gene expression data. They showed that the essential features of a given set of expression profiles can be captured using just a small number of characteristic modes, thereby suggesting the viability of dimension reduction process in LSA mentioned earlier. In a similar work, Alter *et al.* (2000) analyzed yeast cell cycle data using SVD and also showed that the components revealed by the mathematical analysis can be assigned corresponding biological meanings. For example, they identified sinusoidal modes in the singular value decomposition of the expression data that corresponded to the various cell cycle modes.

In the application of SVD to to the task of *gene clustering*, Wall *et al.* (2001) described a method for generating gene groups directly from the *eigengenes*. The LSA approach that we will describe in this paper is different from theirs in that we do not perform clustering in the "eigenspace". In a more recent work, Horn et al. (2003) presented an LSA approach that also uses SVD followed by dimensional reduction, but they have applied a quantum-clustering algorithm in the reduced SVD space for clustering of the genes. We use LSA here for semi-supervised whole-genome functional classification of genes instead of merely clustering. We conduct post-clustering classification of genes by applying a majority voting scheme with the known annotations from a reference set in each of the resulting gene clusters. Using this semi-supervised LSA approach on microarray data, we were able to perform systematic functional classification on the partially-annotated yeast genome.

3 Materials and Methods

3.1 Data

For our study, we apply our LSA classification method on the 6,221 genes in the *Saccharomyces cerevisiae* genome using the yeast gene expression data from Eisen's Lab (Eisen et al. 1998) available at http://rana.lbl.gov/EisenData.htm. For each gene, there is a total of 80 data points generated from a variety of experimental studies such as spotted arrays using samples collected at various time points during sporulation (Chu, DeRisi, Eisen et al. 1998), diauxic shift (DeRisi, Iyer & Brown 1997), mitotic cell division cycle (Spellman, Sherlock, Zhang et al. 1998), and various other experimental conditions.

The microarray data are represented by a gene expression matrix A of dimension 6221×80. Each value a_{ij} in A contains the relative expression level of the i-th gene under the j-th assay or condition. Each row in A therefore represents the expression signature of a gene under the 80 experimental conditions and assays, while the columns represent genome-wide expression profiles of a particular assay or condition.

For a reference set of functional classification of the genes, we refer to the MIPS annotations given in the CYGD database (Mewes et al. 2002). This database has been compiled based on extensive knowledge in the literature, and it is available at http://mips.gsf.de/genre/proj/yeast.

3.2 Method

We describe our method as follows:

Step 1. *Singular Value Decomposition.*

- **1.1** Given a gene expression matrix A, perform singular value decomposition on it such that $A = U \cdot W \cdot V^T$;
- **1.2** If not already so, arrange the eigenvectors in the order of their relative significance as indicated by the diagonal scaling values in W.

Step 2. *Dimension Reduction by Coverage.*

- **2.1.** Compute r, the number of eigenvectors to be retained based on the desired fraction θ of expression coverage by the eigenvectors. The expression coverage C_i of the i-th eigenvector is defined as $C_i = w_i^2 / \sum_{k=1}^{n} w_k^2$, where w_k is the k-th scaling value in W (Alter, Brown & Botstein 2000). The number of eigenvectors to be retained is thus the smallest r such that $\sum_{i=1}^{r} C_i \geq \theta$.
- **2.2** Create a new scaling matrix W' by setting the $w_k = 0$ for $k = r+1, \ldots, n$.
- **2.3** Reconstruct the reduced gene expression matrix A' using $A' = U \cdot W' \cdot V^T$;

Step 3. *Clustering by Vector Similarity.*

- **3.1** Normalize each row in A' such that $\sum_{j=1}^{n} a_{ij}^2 = 1$.
- **3.2** For each normalized row r_i, generate its neighborhood set $\mathcal{F}_i = \{k | r_i \cdot r_k \geq \rho_1\}$ for a pre-set value of ρ_1, $0 \leq \rho_1 \leq 1$.
- **3.3** Iteratively, merge any neighborhood sets with average inter-cluster similarity $\geq \rho_2$, where $\rho_1 \geq \rho_2$.

Step 4. *Calling by Majority.*

- **4.1** Each resulting $\mathcal{F}_i$ is assigned a gene functional class label by majority voting on the annotated genes in the predicted set.
- **4.2** The function of the unannotated genes in each set is then predicted to be the corresponding functional class label.

In our current study, we used $\theta = 0.80, \rho_1 = 0.95, \rho_2 = 0.85$. Note that we have used the vector dot-product here as the measure of similarity rather than the proximity between vectors here—this is consistent to the standard application of LSA (Landauer & Dumais 1997). While other similarity measures such as the Pearson correlation—a common similarity measure for microarray data analysis—can also be

Table 1: Classification of the top 30 major functional classes in MIPS using our LSA method for whole-genome functional classification on yeast gene expression data.

Functional Class	Known MIPS Class ($\mathcal{M}$)	Predicted LSA Class ($\mathcal{L}$)	Known Genes in $\mathcal{L}$ ($\mathcal{M}_\mathcal{L}$)	Dominant Class in $\mathcal{L}$ ($\mathcal{D}_\mathcal{L}$)	Precision ($\frac{\lvert\mathcal{D}_\mathcal{L}\rvert}{\lvert\mathcal{M}_\mathcal{L}\rvert}$)	Recall ($\frac{\lvert\mathcal{D}_\mathcal{L}\rvert}{\lvert\mathcal{M}\rvert}$)
mRNA transcription	246	481	172	110	0.64	0.45
ribosome biogenesis	126	177	125	109	0.87	0.87
cell cycle	107	433	121	65	0.54	0.61
amino acid metabolism	97	100	47	35	0.75	0.36
lipid, etc. metabolism	91	74	34	24	0.71	0.26
C-compd and carbohydrate metabolism	78	107	36	24	0.67	0.31
nucleotide metabolism	77	129	46	29	0.63	0.38
DNA processing	71	137	49	32	0.65	0.45
vesicular transport	65	114	49	27	0.55	0.42
metabolism of vitamins, etc.	56	89	36	20	0.56	0.36
nucleus	55	71	28	17	0.61	0.31
stress response	54	58	29	24	0.83	0.44
proteolytic degradation	51	50	19	15	0.79	0.29
cell differentiation	48	64	27	19	0.70	0.40
protein modification	47	39	17	12	0.71	0.26
translation	42	42	17	11	0.65	0.26
other transcription activities	37	45	16	14	0.88	0.38
aminoacyl-tRNA-synthetases	33	35	17	12	0.71	0.36
protein targeting, sorting, etc.	32	93	39	20	0.51	0.63
tRNA transcription	32	34	14	11	0.79	0.34
respiration	30	56	25	17	0.68	0.57
protein folding and stabilization	28	65	26	12	0.46	0.43
cell wall	27	51	20	15	0.75	0.56
detoxification	25	24	12	9	0.75	0.36
ionic homeostasis	21	18	9	7	0.78	0.33
other transport facilitators	19	16	7	6	0.86	0.32
assembly of protein complexes	15	22	8	6	0.75	0.40
extracellular transport, etc.	15	18	5	2	0.40	0.13
intracellular signalling	15	23	5	5	1.00	0.33
transport mechanism	12	32	5	4	0.80	0.33

applied, we have found that the LSA usage of vector dot-product as a similarity measure performs equally well (data not shown).

4 Results

Using our LSA classification method, we performed genome-wide functional annotation on the 6,221 yeast genes in the partially annotated *Saccharomyces cerevisiae* genome using the 80-experiment yeast gene expression data from Eisen's Lab described in Section 3.1. The functional annotations from MIPS's CYGD database are used as reference—the CYGD database uses the standard Gene Ontology (GO)(Ashburner, Ball, Blake et al. 2000) for its functional annotation. As the GO is a hierarchical classification scheme, we normalize the reference genes' functional annotations by using only functional class labels up to GO level 2.

We include in our reference set only MIPS-annotated yeast genes from non-trivial functional classes—i.e., functional classes with more than three genes. Unlike previous similar studies such as the study by Mateos *et al.*, we have chosen in our analysis to exclude from our reference set genes with ambiguous functional assignments—namely, genes that belong to multiple functional classes. The inclusion of such genes in the classification process had been found to corrode classification results due to the so-called "Borges Effect" (Mateos et al. 2002); more on this will be discussed later in Section 5. 4,095 out of the 6,221 yeast genes studied in the microarray experiments contained MIPS functional annotation. After excluding multi-function genes, we have in our reference set 1,851 single-function genes covering 58 GO level-2 MIPS functional classes. This means that we are using only 30% of the 6,221 yeast genes as a reference set to predict the functional classification of the other 4,370 unknown yeast genes.

In a related work by Mateos *et al.* (2002), they defined functional classes with a true positive (or precision) rate $\geq 40\%$ as *"learnable"*. In our LSA analysis, we found that 40 of the 58 MIPS functional classes—that is, about 70% of the functional classes that have gene expression data—are learnable using our approach. Table 1 shows the detailed classification results of the first 30 major MIPS functional classes. Overall, the resulting precision rate ranges mostly between 0.6 and 0,8, with an average of 0.7; see Figure 1. This represents a significant improvement in whole-genome gene functional classification. In the work by Mateos *et al.* (2002), they found that only $\sim 10\%$ of the MIPS functional classes—or rather, 8 out of the 96 classes that they looked at—are learnable with the supervised learning algorithm (multilayer perceptrons) on the same set of yeast microarray data. They have attributed the cause of the poor learnability partly to the so-called "Borges Effect"—they did not exclude the multi-function genes from their training set as we have done for our reference

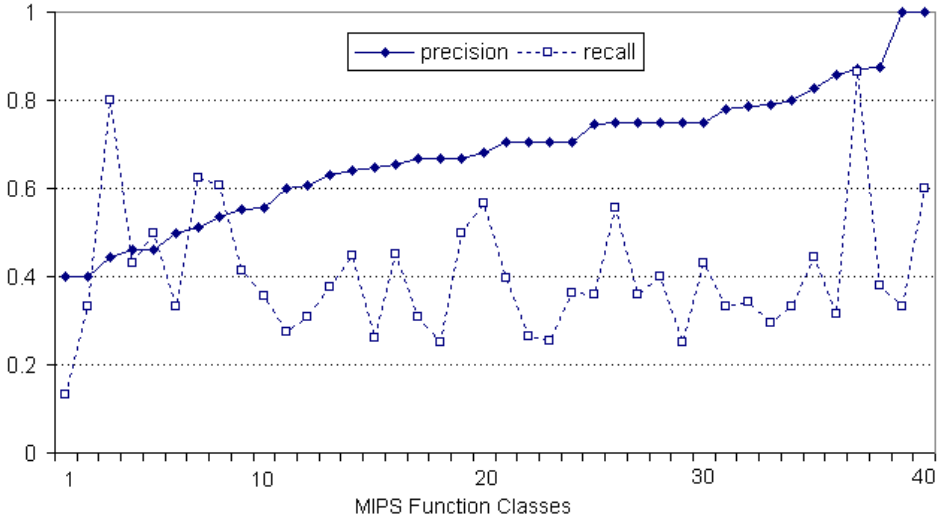

Figure 1: Classification performance of 40 non-trivial learnable MIPS yeast gene functional class using our LSA approach.

set, which explains why their reference set contains 96 MIPS classes while we only have half the number of MIPS classes in ours. We will show in Section 5 that our precision rates for whole-genome annotation approach is not affected even with the inclusion of multi-function genes in our reference set.

Prediction of functional class for unannotated genes is carried out in our LSA classification method using the "guilt-by-association" principle (Walker et al. 1999). By globally clustering the unknown genes together with the annotated genes, those unannotated genes that are clustered in successfully classified LSA groups will have their function class predicted as the function class of the cluster that they belong to. Using a reference set of only 1,851 annotated genes, we were able to classify 1,740 unknown yeast genes into 40 learnable function classes using our LSA approach.

5 Discussions

Previous researchers have noted that there are four main factors that influence the systematic learning of gene functional classes from DNA array expression data (Mateos et al. 2002): data noise, class size, class heterogeneity, and the internal structure of the catalog (or the so-called "Borges effect"). In this section, we discuss how our LSA approach deals with each of these by design.

5.1 Data Noise

The noisy nature of microarray data is one of the major complications in analyzing high-throughput gene expression data. In our LSA approach, we use SVD as a de-noising mechanism. We normalize our data in the dimension reduction step in our method by filtering out the various insignificant eigen components that may correspond to additive or multiplicative experimental noise and background signals from irrelevant biological processes. The decoupling of the various contributing components in the eigen-space by SVD ensures that they can be effectively filtered out from the data without eliminating any relevant information from the data and corroding the subsequent classification performance.

To show that the data normalization step by SVD improves the further analysis of expression data, we applied our classification method without the SVD dimension reduction steps (namely, skipping Steps 1 and 2 in Method). Instead of the 38 learnable classes that we have obtained previously, there were only 16 learnable classes resulted—each with a low recall rate, with the mean recall rate being only 14% instead of 40%. This result indicates that the dimension reduction step by SVD clearly benefits the processing of expression data for functional classification—the SVD-processed data had been effectively de-noised with the relevant signals enhanced for subsequent analysis.

5.2 Function Class Sizes

Another reported determinant of the learning rates of gene function classification is the size of the function class. Mateos *et al.* (2002) had showed that there is a clear trend for the true positive rate to increase with the class size—the more examples there are, the easier it is for a class to be learned. The larger class size also helps to offset the imbalance in the number of positive and negative examples with respect to each function class—a problem in multi-class whole-genome function classification. Supervised machine learning algorithms such as neural networks are known to be sensitive to function class sizes (Mateos et al. 2002).

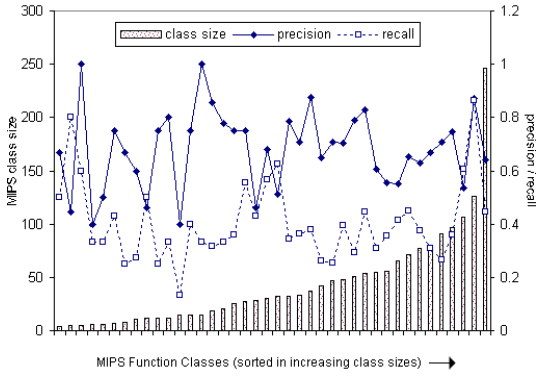

Figure 2: Classification performance versus MIPS function class size.

Figure 2 shows that unlike supervised classification approaches, there is no clear correlation between the function class size and the learning rates in our LSA approach. For a case in point, we were able to achieve a precision rate of 67% for the leftmost "`ion transporters`" class, which has only 4 genes. For

the rightmost "`mRNA transcription`" class, the precision rate is a similar 64% even though it has 246 genes. The use of unsupervised but sensitive SVD-based clustering has allowed for functional classes with original sizes from 4 to 246 genes to be automatically annotated with reasonable accuracies. In terms of recall rates, our approach is equally unaffected by the class sizes, as shown in the figure. As opposed to supervised machine learning approaches, our approach is clearly much less sensitive to function class sizes and the imbalance of positive and negative examples , and it is therefore more amenable to the challenges of the task of whole-genome gene classification.

5.3 Function Class Heterogeneity

Heterogeneity in the function classes are expected from a biological point of view, particularly for the larger classes. For example, genes in the function class for "`assembly of protein complexes`" are expected to be expressed in a heterogeneous fashion—since different complexes are compiled under different conditions, the genes are unlikely to be expressed in a coordinated fashion under different conditions.

Function classes with member genes that express heterogeneous profiles are clearly problematic for machine learning classifiers—the expression profiles of heterogeneous function classes would not contain clear-cut clusters for the classifiers to derive suitable decision boundaries. Our LSA clustering is no different from any other common classification methods—the LSA groupings can only capture genes that are homogeneous in expression. However, we handle class heterogeneity here by allowing for multiple groups to be called under the same function label, as long as there is a majority of reference genes in each group that have the function in question. For example, the "`assembly of protein complexes`" function class mentioned earlier is composed of six LSA subgroups, each with a majority of corresponding reference genes. Overall, the number of LSA subgroups for the various MIPS function classes in our analysis ranges from 2 to 85.

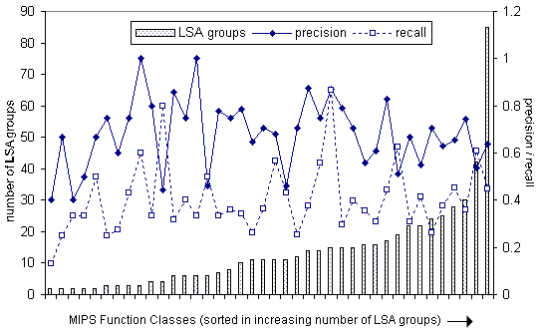

Figure 3: Classification performance (precision) versus number of LSA groups.

To verify whether our LSA approach is sensitive towards heterogeneity in the function classes, we use the number of LSA groups assigned to a function class here as a measure of class heterogeneity, and we compare the classification performance of function classes with varying heterogeneity. Figure 3 shows no direct correlation between the classification performance—precision and recall alike—and the differenece in heterogeneity in the underlying functional classes. The indicates that our LSA approach is robust against class heterogeneity for whole-genome gene classification.

5.4 Borges Effect

In Section 3.1, we had used a reduced reference set containing only single-annotation genes. In other words, our reference gene function set contained only equivalence classes. This may not be reasonable from the biological point of view, for most cellular processes are clearly not stand-alone as they are expected to interact with other processes. Thus, most functional classes are not equivalence classes. Mateos *et al.* (2002) have termed this inherent limitation of functional classification systems the "Borges effect". They have shown that the inherently cross-linking internal structure of the catalog of functional classes can be costly to the performance of supervised machine learning classifiers as they can easily be confused in distinguishing positive from negative examples. This was the reason why we had chosen the 40 non-overlapping MIPS classes as our reference set for our analysis.

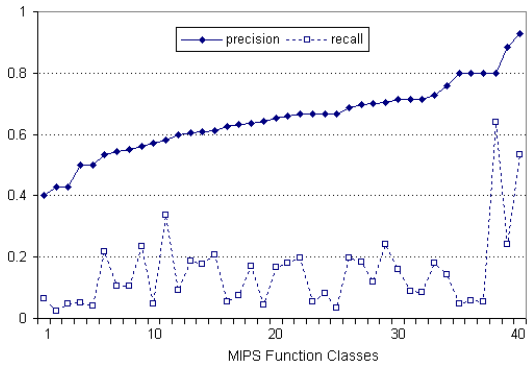

Figure 4: Classification performance using intersecting MIPS function group annotations.

In investigating the extent by which the Borges effect can affect our LSA approach in whole-genome functional annotation, we perform a similar classification experiment, including multi-function genes in our reference set this time. For the same 58 MIPS functional classes, the new reference set contains 8,674 functional annotation for 4,095 genes—in comparison, our previous reference set contained only 1,851 singly-annotated genes.

Figure 4 shows the resulting classification performance of our LSA approach—in terms of both precision and recall—for the same 40 MIPS classes that we have investigated previously. Comparing with the previous results shown in Figure 1, we observe that the main decline is in the recall rates. The Borges effect introduced in the reference set had caused a mean decrease of -0.25 in recall, despite the use of a much larger reference set. On the other hand, the LSA approach is shown to be fairly robust against the Borges effect in terms of precision, incurring a mere mean decrease of -0.06 in precision. It will be interesting future work to investigate ways to improve the recall rates against the Borges effect.

6 Conclusions

The recent advances in microarray technology has certainly revolutionized the way molecular biologists study the functional relationships among genes. While we are now able to monitor gene expression at the genomic scale using microarray technology, there are still gaps toward whole-genome functional annotation of genes using the gene expression data. Recent work by Brown *et al.* (2000) and Mateos *et al.* (2002) have shown that while it is possible to use machine learning algorithms to systematically learn the gene

functional classes for some number of the genes in the genome, the number of genes that can be annotated this way is not yet at the genomic scale. For example, Brown *et al.* focused on learning only 5 functional classes (using sophisticated support vector machines), while Mateos *et al.* concluded that only $\sim 10\%$ of the functional classes—i.e. 8 out of 96 MIPS functional classes—are learnable by their neural networks.

In this paper, we have used an alternative semi-supervised approach based on latent semantic analysis (LSA) on the problem of whole-genome gene functional classification. Our approach is a 4-step procedure: singular value decomposition, dimension reduction by coverage, clustering by similarity, followed by assignment by majority. Using unsupervised pre-classification clustering, our approach is shown to be less susceptible to the many difficulties in whole-genome gene functional classification. For example, the inherent imbalance of training examples is handled by considering the groupings of all classes globally without the *a priori* partitioning by the often limited positive examples. The use of dimension reduction by SVD in our LSA approach effectively denoises the data to allow the subsequent clustering algorithm to focus only on significant functional components present in the expression data. Heterogeneity in the function classes is handled by allowing multiple subgroups to be called as the same functional class. With these, we have shown that the LSA approach can be useful for systematic whole-genome functional classification of genes, as indicated by the promising classification of more than 1,700 genes in the partially-annotated yeast genome into 40 distinct MIPS functional classes.

Acknowledgments

We would like to thank the reviewers for their useful suggestions. We would also like to acknowledge the earlier contribution to this work by Sanjay Padmakar Khadayate.

References

Alter, O., Brown, P. O. & Botstein, D. (2000), 'Singular value decomposition for genome-wide expression data processing and modeling', *Proc Natl Acad Sci U S A* **97**(18), 10101–6.

Alter, O., Brown, P. O. & Botstein, D. (2001), Processing and modeling genome-wide expression data using singular value decomposition, *in* M. L. Bittner, Y. Chen, A. N. Dorsel & E. R. Dougherty, eds, 'Microarrays: Optical Technologies and Informatics', Vol. 4266, pp. 171–186.

Ashburner, M., Ball, C. A., Blake, J. A. et al. (2000), 'Gene ontology: tool for the unification of biology. the gene ontology consortium', *Nat Genet* **25**(1), 25–9.

Brown, M. P., Grundy, W. N., Lin, D. et al. (2000), 'Knowledge-based analysis of microarray gene expression data by using support vector machines', *Proc Natl Acad Sci U S A* **97**(1), 262–7.

Chu, S., DeRisi, J., Eisen, M. et al. (1998), 'The transcriptional program of sporulation in budding yeast', *Science* **282**(5389), 699–705.

Deerwester, S., Dumais, S., Furnas, G. et al. (1990), 'Indexing by latent semantic analysis', *Journal of the American Society for Information Science* **41**(6), 391–407.

DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997), 'Exploring the metabolic and genetic control of gene expression on a genomic scale', *Science* **278**(5338), 680–6.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998), 'Cluster analysis and display of genome-wide expression patterns', *Proc Natl Acad Sci U S A* **95**(25), 14863–8.

Holter, N. S., Mitra, M., Maritan, A. et al. (2000), 'Fundamental patterns underlying gene expression profiles: simplicity from complexity', *Proc Natl Acad Sci U S A* **97**(15), 8409–14.

Horn, D. & Axel, I. (2003), 'Novel clustering algorithm for microarray expression data in a truncated svd space', *Bioinformatics* **19**(9), 1110–5.

Landauer, T. & Dumais, S. (1997), 'A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge', *Psychological Review* **104**, 211–240.

Mateos, A., Dopazo, J., Jansen, R. et al. (2002), 'Systematic learning of gene functional classes from dna array expression data by using multilayer perceptrons', *Genome Res* **12**(11), 1703–15.

Mewes, H. W., Frishman, D., Guldener, U. et al. (2002), 'Mips: a database for genomes and protein sequences', *Nucleic Acids Res* **30**(1), 31–4.

Raychaudhuri, S., Stuart, J. M. & Altman, R. B. (2000), 'Principal components analysis to summarize microarray experiments: application to sporulation time series', *Pac Symp Biocomput* pp. 455–66.

Spellman, P. T., Sherlock, G., Zhang, M. Q. et al. (1998), 'Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization', *Mol Biol Cell* **9**(12), 3273–97.

Tamayo, P., Slonim, D., Mesirov, J. et al. (1999), 'Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation', *Proc Natl Acad Sci U S A* **96**(6), 2907–12.

Toronen, P., Kolehmainen, M., Wong, G. & Castren, E. (1999), 'Analysis of gene expression data using self-organizing maps', *FEBS Lett* **451**(2), 142–6.

Walker, M. G., Volkmuth, W., Sprinzak, E. et al. (1999), 'Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes', *Genome Res* **9**(12), 1198–203.

Wall, M. E., Dyck, P. A. & Brettin, T. S. (2001), 'Svdman–singular value decomposition analysis of microarray data', *Bioinformatics* **17**(6), 566–8.