Open Access

# Yeast Gene Function Prediction from Different Data Sources: An Empirical Comparison

Ying Liu*

*Department of Mathematics and Information Sciences, University of North Texas at Dallas, 7300 University Hills Blvd Dallas, TX 75241, USA*

**Abstract:** Different data sources have been used to learn gene function. Whereas combining heterogeneous data sets to infer gene function has been widely studied, there is no empirical comparison to determine the relative effectiveness or usefulness of different types of data in terms of gene function prediction. In this paper, we report a comparative study of yeast gene function prediction using different data sources, namely microarray data, phylogenetic data, literature text data, and a combination of these three data sources. Our results showed that text data outperformed microarray data and phylogenetic data in gene function prediction ($p<0.01$) as measured by sensitivity, accuracy, and correlation coefficient. There was no significant difference between the results derived from microarray data and phylogenetic data ($p>0.05$). The combined data led to decreased prediction performance relative to text data. In addition, we showed that feature selection did not improve the prediction performance of support vector machines.

**Keywords:** Gene Prediction, Microarray data, Phylogenetic Data, Literature Text Data, Comparative Study.

## 1. INTRODUCTION

Functional genomics studies gene function on a large scale by conducting parallel analysis of gene expression for a large number of genes [1, 2]. This research is a natural successor to the genome sequencing efforts such as, for example, the Human Genome Project, and is made possible by the DNA microarrays. Such arrays, which allow researchers to simultaneously measure the expression levels of thousands of different genes, produce overwhelming amounts of data. In response, much recent research has been concerned with automating the analysis of microarray data [3]. Current approaches mainly concentrate on applying clustering techniques to the expression data, in order to find clusters of genes demonstrating similar expression patterns. The assumption motivating such search for co-expressed genes is that simultaneously expressed genes often share a common function. However, there are several reasons that cluster analysis alone cannot fully address this core issue [3].

High-throughput gene and protein assays give a view into the organization of molecular cellular life through quantitative measurements of gene expression levels [1]. Increasing quantities of high-throughput biological data have become available to assess functional relationships between gene products on a large scale. Different data sources can be used to predict gene function.

First, gene function can be inferred from DNA microarray expression data. DNA microarray is based on the assumption that genes with similar functions have similar expression profiles in cells. This is utilized by inductive learning methods that predict the function of genes that have an unknown function (unknown genes), from their expression-similarity with genes with a known function (known genes) [3]. Currently, techniques pursued for microarray data analysis concentrate on applying clustering methods directly on the expression data. However, cluster analysis alone cannot fully address the issue of gene function prediction [3]. Furthermore, many high-throughput methods sacrifice specificity for scale. Whereas gene coexpression data are an excellent tool for hypothesis generation, microarray data alone often lack the degree of specificity needed for accurate gene function prediction [4].

Secondly, gene function can be inferred from phylogenetic profiles. The complete genomic sequences of human and other species provide a tremendous opportunity for understanding the functions of biological macromolecules [5]. Phylogenetic profiles are derived from a comparison between a given gene and a collection of complete genomes. Each profile characterizes the evolutionary history of a given gene. There is evidence that two genes with similar phylogenetic profiles may have similar functions, the idea being that their similar pattern of inheritance across species is the result of a functional link [6].

Finally, one more data source that can be used to infer the gene function is the scientific literature. The function of many genes is described in the literature. By relating documents talking about well understood genes to documents discussing other genes, we can predict, detect, and explain the functional relationships between the genes that are involved in large-scale experiments. A number of groups are developing to organize genes. The web tool, PubGene, finds links between pairs of genes based on their co-occurrence in MEDLINE abstracts [7]. Liu *et al*. [8, 9] developed a tool to retrieve functional keywords automatically from biomedical literature for each gene, and then cluster the genes by shared functional keywords. Using a similarity-based search in document space, Shatkay *et al*. [3] developed an approach

*Address correspondence to this author at the Department of Mathematics and Information Sciences, University of North Texas at Dallas, 7300 University Hills Blvd Dallas, TX 75241, USA; Tel: 972-338-1573; Fax: 972-338-1911; E-mail: ying.liu@unt.edu

1875-0362/11 2011 Bentham Open

for utilizing literature to establish functional relationships among genes on a genome-wide scale.

Different data sources have been used to infer gene functions [3-9]. Furthermore, heterogeneous data sources have been combined to predict gene functions [10-12]. However, there is no empirical comparison to determine the relative effectiveness or usefulness of different types of data in terms of gene function prediction. In this paper, we performed a comparative study for functional prediction of *Saccharomyces cerevisiae* genes from different data sources. Data from three different types of sources were compared: microarray data, phylogenetic profile data, biomedical literature data, and a combination of the three heterogeneous data sets. The goal was to determine the relative effectiveness or usefulness of this data in terms of gene function prediction.

## 2. METHODS

### 2.1. Data Sources

1. The first data set derives from a collection of DNA microarray hybridization experiments [13]. Each data point represents the logarithm of the ratio of expression levels of a particular gene under two different experimental conditions. The data consists of a set of 79-element gene expression vectors for 2,465 yeast genes [4]. These genes were selected by Eisen *et al.* [13] based on the availability of accurate functional annotations. The data were generated from spotted arrays using samples collected at various time points during the diauxic shift [14], the mitotic cell division cycle [15], sporulation [16], and temperature and reducing shocks [4].

2. In addition to the microarray expression data, each of the 2,465 yeast genes is characterized by a phylogenetic profile [5]. In its simplest form, a phylogenetic profile is a bit string, in which the Boolean value of each bit indicates whether the gene of interest has a close homolog in the corresponding genome. The profiles employed in this paper contain, at each position, the negative logarithm of the lowest E-value reported by BLAST version 2.0 [17] in a search against a complete genome, with negative values (corresponding to E-values greater than 1) truncated to 0. Two genes in an organism can have similar phylogenetic profiles for one of two reasons [4]. First, genes with a high level of sequence similarity will have, by definition, similar phylogenetic profiles. Second, for two genes which lack sequence similarity, the similarity in phylogenetic profiles reflects a similar pattern of occurrence of their homologs across species [4]. This coupled inheritance may indicate a functional link between the genes, on the hypothesis that the genes are always present together or always both absent because they cannot function independently of one another [4].

3. In this paper, the experiments are carried out using gene functional categories from the MIPS Comprehensive Yeast Genome Database (CYGD) (*http://mips.helmholtz-muenchen.de/genre/proj/yeast/index.jsp*). The database contains several hundred functional classes, whose definitions come from biochemical and genetic studies of gene function [4]. For each of the genes, the abstracts used to curate the CYGD were extracted and formed a document. Abstracts may occur in more than one document if they refer to multiple genes. All the documents form a document database. Since each document represents one gene, we use the words *document* and *gene* interchangeably.

The abstracts in each document were tokenized, stemmed by Porter's stemming algorithm, and filtered by a stop list [9]. The standard term frequency-inverse document frequency (TFIDF) function was used [18] to assign the weight to each word in the document. TFIDF combines term frequency (TF), which measures the number of times a word occurs in the gene's set of abstracts (reflecting the importance of the word to the gene), and inverse document frequency (IDF), which measures the information content of a word – its rarity across all the abstracts in the background set. The inverse document frequency (IDF) is calculated as:

$$idf^{a} = \log \frac{N}{df^{a}} \tag{1}$$

where $idf^{a}$ denotes the inverse document frequency of word *a* in all the documents; $df^{a}$ denotes the number of abstracts in which word *a* occurs; and N is the total number of abstracts in all the documents.

TFIDF is defined as:

$$tfidf_{g}^{a} = tf_{g}^{a} \times idf^{a} \tag{2}$$

$tfidf_{g}^{a}$ denotes the weight of the word *a* to the gene *g*; $tf_{g}^{a}$ the number of times word *a* occurs in gene *g*.

To distribute the word weights over the [0, 1] interval, the weights resulting from TFIDF were often normalized by *cosine normalization*, given by

$$Weight_{g}^{a} = \frac{tfidf_{g}^{a}}{\sqrt{\sum_{s=1}^{|W|} \left(tfidf_{g}^{s}\right)^{2}}} \tag{3}$$

where |W| denotes the number of words in the abstracts of gene *g*.

Each document, which corresponded to one gene, was modeled as an *M*-dimensional TFIDF vector, where *M* is the number of distinct words in the document. Formally, a document was a vector ($tfidf_1$, $tfidf_2$, … , $tfidf_M$), where $tfidf_i$ is the *tfidf* value of word *i*.

4. These three types of data mentioned above are combined by concatenating the three types of vectors to form a single set of vectors. This is also called early integration or feature integration [4]. We used feature integration because feature integration considers the various types of data at once, making a single prediction for each gene with respect to each functional category [4].

Prior to learning, the gene expression, phylogenetic profile, text TFIDF vectors, and the combined data are adjusted

to have a mean of 0 and a variance of 1. The gene expression and phylogenetic profile data were from [4].

### 2.2. Classifier

In this study, Support Vector Machine (SVM) was used for gene function prediction. SVMLight v.3.5 was used [19]. SVM has been widely used in gene and protein function prediction [12, 20]. Linear kernel and polynomial kernel were applied.

### 2.3. Cross-Validation of the Models

The normal method to evaluate the prediction results is to perform cross-validation on the prediction algorithms [21]. Tenfold cross-validation has been proved to be statistically good enough in evaluating the prediction performance [22, 23]. In this paper, each of the data sets (microarray, phylogenetic, text mining, and the combined data sets) was partitioned into ten subsets with both positive and negative genes spread as equally as possible between the sets. Each of these sets in turn was set aside while a model was built using the other nine sets. This model was then used to classify the genes in the tenth set, and the accuracy computed by comparing these predictions with the actual category. This process was repeated ten times and the results averaged [24].

### 2.4. Feature SELECTION

The feature selection method we used in this study is MIT correlation score, which is also known as the signal-to-noise score [25] that helps to eliminate the "noisy" features. For a given feature *i*, we compute the mean and standard deviation of that feature across the positive examples ($\mu_i^+$ and $\sigma_i^+$, respectively) and across the negative examples ($\mu_i^-$ and $\sigma_i^-$, respectively). The MIT correlation score is defined as MIT ($i$) = $\frac{|\mu_i^+ - \mu_i^-|}{\sigma_i^+ + \sigma_i^-}$. When making selection, we simply take those features with the highest scores as the most discriminatory features. For the text data, the features are the terms or the distinct words.

### 2.5. Performance Measures

Several statistics were used as performance measures:

(1). Accuracy: the proportion of correctly classified instances:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

where true positives (TP) denote the correct predictions of positive examples; true negatives (TN) are the correct predictions of negative examples; false positives (FP) represent the incorrect prediction of negative examples into the positive class; and false negatives (FN) are the positive examples incorrectly classified into the negative class.

(2). Sensitivity: the percent of positive examples which were correctly classified;

$$Sensitivity = \frac{TP}{TP + FN} \quad (5)$$

(3). Specificity: the percent of negative examples which were correctly classified;

$$Specificity = \frac{TN}{TN + FP} \quad (6)$$

(4). Positive Predictive Value (PPV): the percentage of the examples predicted to be positive that were correct:

$$PPV = \frac{TP}{TP + FP} \quad (7)$$

(5). Negative Predictive Value (NPV): the percentage of the examples predicted to be negative that were in fact negative.

$$NPV = \frac{TN}{TN + FN} \quad (8)$$

(6). Correlation Coefficient (CC): It is also known as Simple Matching Coefficient (SMC). CC depends not only on sensitivity and specificity, but also on PPV and NPV.

$$CC = \frac{(TP * TN) - (FN * FP)}{\sqrt{(TP + FN) * (TN + FP) * (TP + FP) * (TN + FN)}} \quad (9)$$

Paired t-tests were performed to evaluate whether the results obtained from the four data sets were significantly different from each other.

## 3. RESULTS

The database contains different functional classes, whose definitions come from biochemical and genetic studies of gene function. The experiments reported here used 8 CYGD functional categories which have the most genes available in the CYGD data set as of July 30th, 2009, (Table **1**).

**Table 1. The Gene Function Categories Studied in this Paper**

| Function Category | Function |
|---|---|
| 1 | Metabolism |
| 10 | Cell cycle and DNA processing |
| 11 | Transcription |
| 12 | Protein synthesis |
| 14 | Protein fate (folding, modification, destination) |
| 20 | Cellular transport, transport facilitation and transport routes |
| 34 | Interaction with the cellular environment |
| 42 | Biogenesis of cellular components |

### 3.1. Gene Function Prediction from Microarray, Phylogenetic, and Text Data

The results of gene function prediction from different data sources were shown in Fig. (**1**), Fig. (**2**), Table **2**, and Table **3**.

When microarry data was used and linear kernel was applied for gene function prediction, all the genes in each category, except category #12, were mis-classified (true positive = 0), which can be observed, in Fig. (**1**), that the sensitivity values were 0's. Similar results can be observed when phylogenetic data was used and linear kernel was applied to classify gene function except category #1 (Fig. **1**). When linear kernel was applied and text data was used, the results derived from text data significantly outperformed those derived from microarray data and phylogenetic data ($p<0.01$). SVM can correctly classify the function of the genes in category #12 with an accuracy of 0.963 and a sensitivity of 0.7.

When polynomial kernel was applied, the results derived from text data outperformed those derived from microarray data and phylogenetic data ($p< 0.05$) except category #1 (Fig. **2**). No significant difference was observed between the gene function prediction results derived from microarray data and phylogenetic data ($p >0.05$) (Fig. **2**).

For text data, linear kernel outperformed polynomial kernel ($p<0.01$) as measured by sensitivity, PPV, accuracy, and CC. Polynomial kernel worked significantly better than linear kernel ($p<0.01$) for microarray data, and phylogenetic data (Fig. **1**, Fig. **2**, Table **2**, and Table **3**). A linear-SVM can outperform a polynomial-SVM because the noise contained in the dataset can be amplified by the high-order polynomial kernel into the feature-space, which may weaken the classifier's discriminative power.

### 3.2. Gene Function Prediction from Combined Data

From the Fig. (**1**), Fig. (**2**), Table **2**, and Table **3**, we can see that using combined data to classify yeast gene function did not improve the SVM performance. When linear kernel was applied, the results derived from text data significantly outperformed those derived from the combined data, as measured by sensitivity ($p<0.01$) (Fig. **1A**), accuracy ($p<0.05$) (Fig. **1C**), and CC ($p<0.05$) (Table **2**). There was no significant difference between the combined data results, microarray data results, and phylogenetic data results ($p>0.05$).

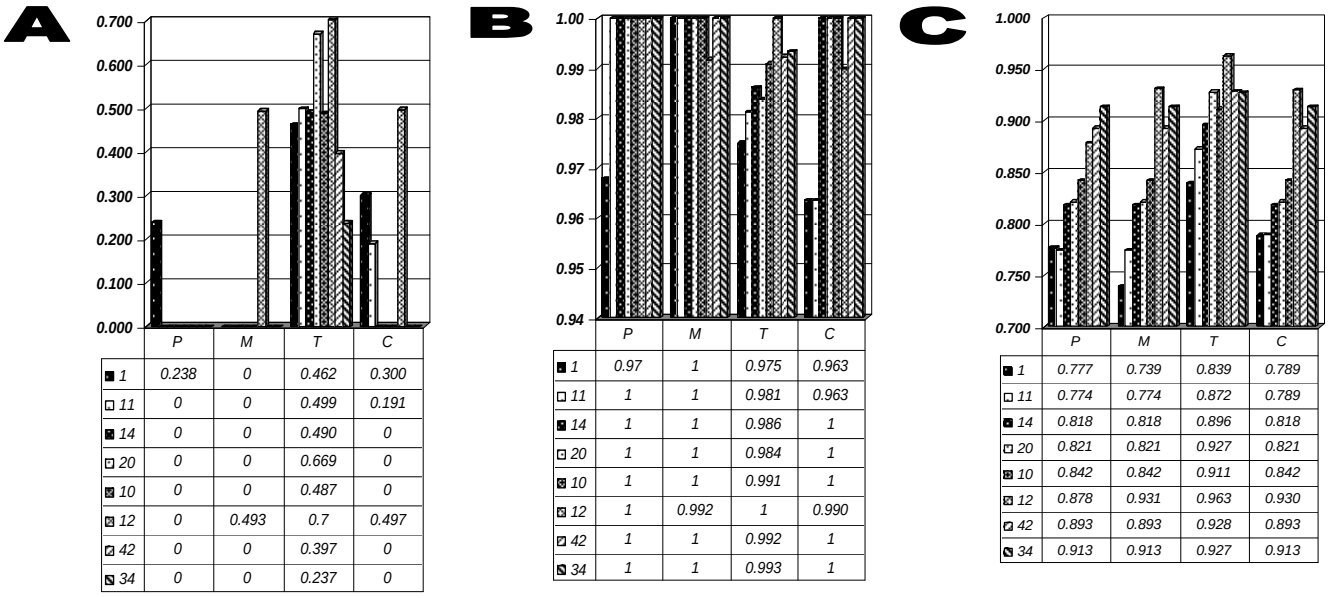
Similar to microarray data and phylogenetic data, polynomial kernel worked significantly better than linear kernel ($p<0.01$) for combined data (Fig. **1**, Fig. **2**, Table **2**, and Table **3**). However, the results derived from text data with linear kernel still outperformed those derived from combined data with polynomial kernel ($p<0.01$).

### 3.3. Feature Selection

In this study, the MIT was used as the feature selection method to test if feature selection can improve SVM performance on gene function prediction using text data. Linear kernel was applied. The experiments demonstrated that, MIT, a naïve feature selection algorithms, which does not take into account the heterogeneity of the data, did not yield improved prediction performance (Fig. **3**). Highest sensitivity, accuracy, PPV, NPV, and CC were obtained when all the features were used.

## 4. DISCUSSION

A primary goal in biology is to understand the molecular machinery of the cell. The sequencing projects provide us one view of this machinery. A complementary view is provided by data from microarray hybridization experiments. High-throughput techniques, such as DNA microarray and sequencing, accompanied by an increase in the number of publications discussing gene-related discovery, provide the researchers great resources to understand the gene function better. In this paper, we classified yeast gene functions from different data sources. CYGD database categorizes the yeast genes into different categories, of which we analyzed eight (category numbers 1, 11, 14, 20, 12, 10, 42, and 34).




| | P | M | T | C |
|---|---|---|---|---|
| 1 | 0.238 | 0 | 0.462 | 0.300 |
| 11 | 0 | 0 | 0.499 | 0.191 |
| 14 | 0 | 0 | 0.490 | 0 |
| 20 | 0 | 0 | 0.669 | 0 |
| 10 | 0 | 0 | 0.487 | 0 |
| 12 | 0 | 0.493 | 0.7 | 0.497 |
| 42 | 0 | 0 | 0.397 | 0 |
| 34 | 0 | 0 | 0.237 | 0 |

| | P | M | T | C |
|---|---|---|---|---|
| 1 | 0.97 | 1 | 0.975 | 0.963 |
| 11 | 1 | 1 | 0.981 | 0.963 |
| 14 | 1 | 1 | 0.986 | 1 |
| 20 | 1 | 1 | 0.984 | 1 |
| 10 | 1 | 1 | 0.991 | 1 |
| 12 | 1 | 0.992 | 1 | 0.990 |
| 42 | 1 | 1 | 0.992 | 1 |
| 34 | 1 | 1 | 0.993 | 1 |

| | P | M | T | C |
|---|---|---|---|---|
| 1 | 0.777 | 0.739 | 0.839 | 0.789 |
| 11 | 0.774 | 0.774 | 0.872 | 0.789 |
| 14 | 0.818 | 0.818 | 0.896 | 0.818 |
| 20 | 0.821 | 0.821 | 0.927 | 0.821 |
| 10 | 0.842 | 0.842 | 0.911 | 0.842 |
| 12 | 0.878 | 0.931 | 0.963 | 0.930 |
| 42 | 0.893 | 0.893 | 0.928 | 0.893 |
| 34 | 0.913 | 0.913 | 0.927 | 0.913 |

**Fig. (1).** The prediction performance, sensitivity (**A**), specificity (**B**), and accuracy (**C**), of the linear-kernel support vector machines with different data sets as inputs. P: phylogenetic data; M: microarray data; T: text data; C: combined data. The series 1, 11, 14, 20, 10, 12, 42, and 34 are the functional categories tested in this study.

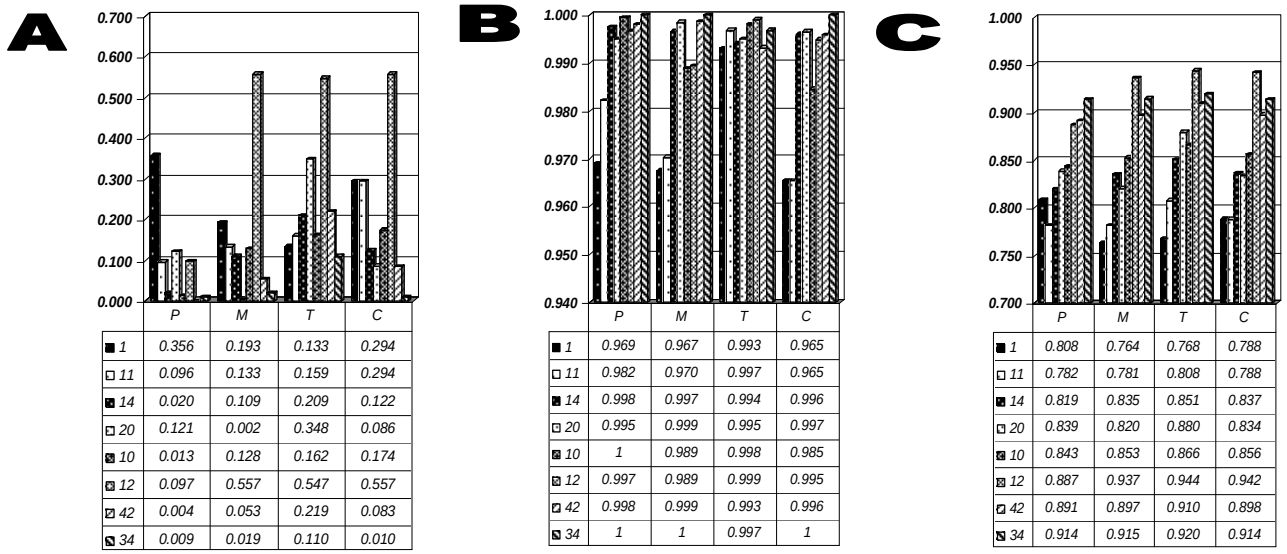



**Fig. (2).** The prediction performance, sensitivity (**A**), specificity (**B**), and accuracy (**C**), of the polynomial-kernel support vector machines with different data sets as inputs. P: phylogenetic data; M: microarray data; T: text data; C: combined data. The series 1, 11, 14, 20, 10, 12, 42, and 34 are the functional categories tested in this study.

Although the idea of combining heterogeneous data sets to infer gene function is not new [4], there is no empirical comparison to determine the relative effectiveness or usefulness of difference types of data in terms of gene function prediction. In this paper, we report a comparative study of yeast gene function prediction using different data sources, namely microarray data, phylogenetic data, literature text data, and a combination of these three data sources.

### 4.1. Effect of Different Data Sources on Gene Function Prediction

The results showed that, using SVM as the classifier, text data can provide better prediction results than microarray data and phylogenetic data, particularly when linear kernel was applied (Fig. **1**, and Table **2**) as measured by sensitivity, PPV, NPV, accuracy, and CC. These results confirmed that the CYGD predictions we tested are not learnable from

**Table 2. The Prediction Performance (PPV, NPV and CC) of support Vector Machines (Linear Kernel) with Different Data Sets as Inputs**

| Kernel Type | Functional Category | Phylogenetic Data | | | Microarray Data | | | Text Data | | | Combined Data | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **PPV** | **NPV** | **CC** | **PPV** | **NPV** | **CC** | **PPV** | **NPV** | **CC** | **PPV** | **NPV** | **CC** |
| Linear | 1 | 0.718 (0.08) | 0.782 (0.04) | 0.320 (0.07) | - | 0.739 (0.03) | * | **0.876** (0.06) | 0.837 (0.03) | **0.553** (0.04) | 0.739 (0.07) | 0.795 (0.04) | 0.374 (0.07) |
| | 11 | - | 0.774 (0.03) | * | - | 0.774 (0.00) | * | **0.884** (0.03) | 0.871 (0.01) | **0.601** (0.07) | 0.567 (0.02) | 0.804 (0.01) | 0.253 (0.03) |
| | 14 | - | 0.818 (0.00) | * | - | 0.818 (0.00) | * | **0.890** (0.05) | 0.897 (0.01) | **0.610** (0.07) | - | 0.818 (0.00) | * |
| | 20 | - | 0.821 (0.01) | * | - | 0.821 (0.01) | * | **0.906** (0.04) | 0.932 (0.01) | **0.739** (0.04) | - | 0.821 (0.00) | * |
| | 10 | - | 0.842 (0.00) | * | - | 0.842 (0.00) | * | **0.914** (0.03) | 0.912 (0.01) | **0.625** (0.06) | - | 0.842 (0.00) | * |
| | 12 | - | 0.878 (0.00) | * | 0.894 (0.08) | 0.934 (0.01) | 0.632 (0.07) | **0.997** (0.01) | 0.960 (0.01) | **0.817** (0.05) | 0.883 (0.05) | 0.934 (0.01) | 0.629 (0.05) |
| | 42 | - | 0.893 (0.01) | * | - | 0.893 (0.01) | * | **0.866** (0.10) | 0.932 (0.01) | **0.554** (0.04) | - | 0.893 (0.04) | * |
| | 34 | - | 0.913 (0.00) | * | - | 0.913 (0.00) | * | **0.765** (0.03) | 0.932 (0.01) | **0.387** (0.04) | - | 0.913 (0.00) | * |

-: means no positive was predicted; *: means no value was calculated because of being divided by zero.
Each value is an average value over ten-fold cross-validation. Values in the brackets are the standard errors. Bold values mean significant difference.

**Table 3. The Prediction Performance (PPV, NPV and CC) of Support Vector Machines (Polynormial Kernel) with Different Data Sets as Inputs**

| Kernel Type | Functional Category | Phylogenetic Data | | | Microarray Data | | | Text Data | | | Combined Data | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PPV | NPV | CC | PPV | NPV | CC | PPV | NPV | CC | PPV | NPV | CC |
| Polynomial | 1 | 0.807 (0.06) | 0.809 (0.03) | 0.446 (0.05) | 0.651 (0.11) | 0.772 (0.05) | 0.258 (0.11) | **0.901** (0.11) | 0.764 (0.03) | 0.278 (0.06) | 0.744 (0.05) | 0.794 (0.04) | 0.372 (0.06) |
| | 11 | 0.640 (0.08) | 0.789 (0.01) | 0.180 (0.04) | 0.580 (0.07) | 0.794 (0.01) | 0.192 (0.09) | **0.950** (0.05) | 0.803 (0.01) | 0.329 (0.07) | 0.744 (0.05) | 0.794 (0.04) | 0.372 (0.06) |
| | 14 | 0.533 (0.07) | 0.820 (0.00) | 0.08 (0.06) | 0.874 (0.09) | 0.834 (0.01) | 0.262 (0.10) | 0.903 (0.08) | 0.850 (0.01) | 0.382 (0.10) | 0.881 (0.11) | 0.836 (0.01) | 0.286 (0.06) |
| | 20 | 0.855 (0.12) | 0.839 (0.01) | 0.277 (0.07) | 0.100 (0.33) | 0.821 (0.01) | 0.03 (0.04) | **0.950** (0.04) | 0.876 (0.02) | 0.528 (0.06) | 0.857 (0.15) | 0.834 (0.01) | 0.234 (0.06) |
| | 10 | 0.300 (0.44) | 0.843 (0.00) | 0.162 (0.09) | 0.687 (0.14) | 0.858 (0.01) | 0.251 (0.07) | 0.859 (0.32) | 0.864 (0.01) | 0.329 (0.15) | 0.684 (0.14) | 0.864 (0.01) | 0.294 (0.08) |
| | 12 | 0.873 (0.20) | 0.888 (0.01) | 0.255 (0.08) | 0.889 (0.08) | 0.942 (0.01) | 0.671 (0.07) | 0.991 (0.02) | 0.941 (0.01) | 0.711 (0.05) | 0.944 (0.05) | 0.942 (0.01) | 0.697 (0.07) |
| | 42 | 0.050 (0.17) | 0.893 (0.01) | 0.04 (0.02) | 0.717 (0.37) | 0.898 (0.01) | 0.178 (0.10) | 0.821 (0.16) | 0.914 (0.01) | 0.391 (0.07) | 0.743 (0.24) | 0.900 (0.01) | 0.220 (0.03) |
| | 34 | 0.200 (0.42) | 0.914 (0.00) | 0.105 (0.09) | 0.300 (0.51) | 0.915 (0.01) | 0.224 (0.10) | 0.585 (0.43) | 0.922 (0.01) | 0.224 (0.16) | 0.200 (0.02) | 0.914 (0.00) | 0.209 (0.04) |

Each value is an average value over ten-fold cross-validation. Values in the brackets are the standard errors.

either microarray data or phylogenetic data [4]. Pavlidis *et al.* [4] pointed out that the failure to classify the gene functions from microarray data or phylogenetic data was not a failure of SVM model. Rather, for many functional categories, the data are simply not informative. The microarray data is only informative if the genes in the category are coordinately regulated at the level of transcription under the condition tested. Simultaneous expressed genes may not always share a function. On the other hand, genes that are functionally related may demonstrate strong anti-correlation in their expression levels, (a gene may be strongly suppressed to allow another to be expressed) [3]. Similarly, phylogenetic data are limited in resolution in part because relatively few genomes are available. In particular, among the genomes from which phylogenetic profiles were derived, all but one is bacterial. Thus it is difficult to generate useful phylogenetic profiles for genes that are specific to eukaryotes [4].

One complement data source we can use to classify gene functions is literature data. With the advancement of genome sequencing techniques comes as an overwhelming increase in the amount of literature discussing the discovered genes [26]. As an illustrative example, the number of PubMed documents containing the word *gene* published between the years 1970-1980 is a little over 35,000, while the number of such documents published between the years 1990-2000 is 402,700 – over a ten fold increase [3]. The gene functions have been described in the literature. Therefore, we believe that the gene functions can be classified by revealing coherent themes within the literature. Content-based relationships among abstracts are then translated into functional connections among genes. Liu *et al.* [8, 9, 25] developed a system to retrieve functional keywords automatically from biomedical literature for each gene, and then cluster the genes by shared functional keywords. The keywords extracted by the system revealed a wealth of potential functional concepts, which were not represented in existing public databases [27]. The system also clustered the genes into appropriate functional groups based on the functional keyword association [8, 9].

Our gene function prediction by text data strategy is similar to the document categorization in information retrieval. In our case, each document is the collection of abstracts which are related to a specific gene. Document categorization, defined as classifying documents into categories according to their topics or main contents in a supervised manner, organizes large amounts of information into a small number of meaningful categories and improves the information retrieval performance either *via* term-weighting, or query expansion.

### 4.2. Combining Heterogeneous Data Sets for Gene Function Prediction

The problem of learning from multiple information sources has been extensively studied in machine learning where it is called as multi-modal learning. Generally there are two types of multi-modal learning: feature level integration and semantic integration [28]. The feature integration combines the information at the feature level and performs learning in the joint feature space. The correlation structure

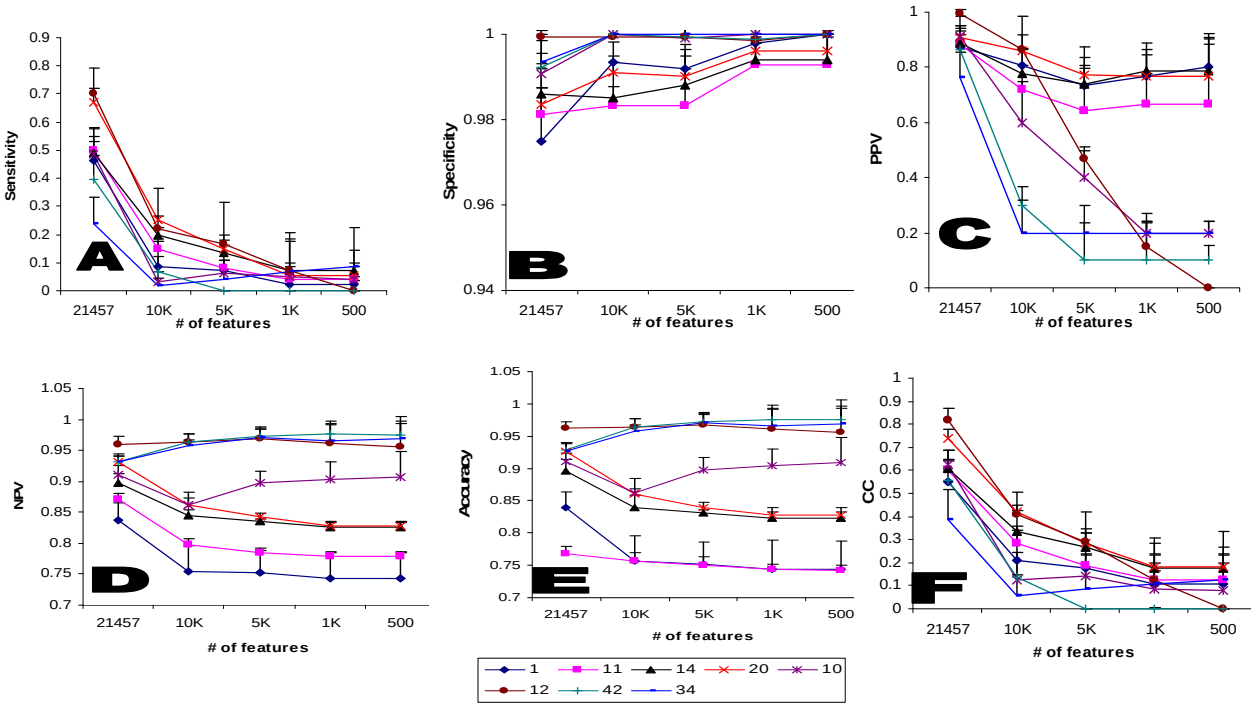



**Fig. (3).** Effect of feature selection in combination of SVM classifier on sensitivity (**A**), specificity (**B**), PPV (**C**), NPV (**D**), accuracy (**E**), and correlation coefficient (**F**). Note the different scales on the vertical axes. The horizontal axes refer to the number of features used by SVM to predict the gene function. Error bars indicated the standard errors. The series 1, 11, 14, 20, 10, 12, 42, and 34 are the functional categories tested in this study.

between different sources can be discovered *via* learning. The semantic integration, on the other hand, first builds individual models based on separate information sources and then combines these models *via* some processes say, mutual information maximization [29]. Li *et al.* [28] listed four reasons why semantic integration was preferred over feature integration. However, Pavlidis *et al.* [4] argued that feature integration considers the various types of data at once, making a single prediction for each gene with respect to each functional category. They also argued that the performance of SVMs when data types are combined and a single hypothesis is formed is superior to combining different independent hypotheses [4].

In this study, we used feature integration. The results showed that the combined data did not improve the prediction results, especially compared with text data. Our results confirmed the conclusion drawn by Pavldis *et al*. [5] when they studied gene function learning from microarray data and phylogenetic data. Learning from different data types is not always a good idea. The combined data led to decreased prediction performance relative to an SVM trained on a single type of data (e.g. text data). In this case, the decrease occurs when one data type provides much more information than the others, indicating that the inferior data types (e.g. microarray and phylogenetic data) contribute noise that disrupts learning [5].

### 4.3. Effect of Feature Selection on Gene Function Prediction

In this study, MIT correlation score was used as the feature selection method. By treating each feature independently, MIT correlation score does not take into account possible correlations between features. But MIT has the advantages of simplicity and efficiency. Prediction performance declined as features are removed. MIT has been successfully applied to gene expression data analysis for cancer prediction [25].

The results of the experiments indicated that SVM did not benefit from feature selection (Fig. **2**), which had been reported in text prediction [30-32]. Taira and Haruno [33] compared SVM and decision tree in text categorization, and the best average performance was achieved when all the features were given to SVM, which was a distinct characteristic of SVM compared with the decision tree learning algorithm. Joachims [19] argued that, in text prediction, feature selection was often not needed for SVM, as SVM tends to be fairly robust to overfitting and can scale up to considerable dimensionalities. SVM avoids overfitting by choosing the maximum margin separating hyperplane from among the many that can separate the positive from negative examples in the feature space. Also, the decision function for classifying points with respect to the hyperplane only involves dot products between points in the feature space. Because the algorithm that finds a separating hyperplane in the feature space can be stated entirely in terms of vectors in the input space and dot products in the feature space, a support vector machine can locate the hyperplane without ever representing the space explicitly, simply by defining a function, called a kernel function, that plays the role of the dot product in the feature space. This technique avoids the computational burden of explicitly representing the feature vectors [34].

## 5. CONCLUSION

The results in this paper showed a rather counter-intuitive result that the literature text data can provide more accurate prediction results over microarray and phylogenetic data in case of the CYGD database containing all the genes of yeast whose function is already known. Combining different data types did not provide better performance than using only a single data type, text data.

## REFERENCES


[1] T. R. Hvidsten and J. Komorowski, "Predicting gene function from gene expressions and ontologies," *Pacific Symp. Biocomput.*, vol. 6, pp. 299-310, 2001.

[2] C. Lippert, G. Zoubin and K. M. Borgwardt, "Gene function prediction from synthetic lethality networks via ranking on demand," *Bioinformatics*, vol. 26, pp. 912-918, 2010.

[3] H. Shatkay, S. Edwards, W. J. Wilbur and M. Boguski, "Genes, themes, and microarrays: using information retrieval for large-scale gene analysis," *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, vol. 8, pp. 317-28, 2000.

[4] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman and D. Botstein, "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*)". *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 8348-8353, 2003.

[5] P. Pavlidis, J. Weston, J. Cai and W. S. Noble, "Learning gene functional predictions from multiple data types", *J. Comput. Biol.*, vol. 9, pp. 401-411, 2002.

[6] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles", *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 4285-4288, 1999.

[7] T. K. Jenssen, A. Aegreid, J. Komorowski and E. Hovig, "A literature network of human genes for high-throughtput analysis of gene expression", *Nat. Genet.*, vol. 178, pp. 139-143, 2001.

[8] Y. Liu, B. J. Ciliax, K. Borges, V. Dasigi, A. Ram, S. B. Navathe and R. Dingledine, "Comparison of Two Schemes for Automatic Keyword Extraction from MEDLINE for Functional Gene Clustering", *Proceedings of 2004 IEEE Computational Systems Bioinformatics Conference (CSB2004)*, Stanford University, August vol. 16-19, 2004, pp. 394-404, 2004.

[9] Y. Liu, "Text mining biomedical literature for discovering gene-to-gene relationships: a comparative study of algorithms", *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 2, pp. 62-76, 2005.

[10] Y. Chen and D. Xu, "Genome-scale protein function prediction in yeast *Saccharomyces cerevisiae* through integrating multiple sources of high-throughput Data", *Pacific Symp. Biocomput.*, vol. 10, pp. 471-482, 2005.

[11] W. S. Noble and A. Ben-Hur, "Integrating Information for Protein Function Prediction Bioinformatics -- From Genomes to Therapies". T. Lengauer, Ed. Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, 2007, vol. 3.

[12] M. Re and G. Valentini, "Prediction of Gene Function Using Ensembles of SVMs and Heterogeneous Data Sources", *Collection of Applications of Supervised and Unsupervised Ensemble Methods*, pp. 79-91, 2009.

[13] M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, "Cluster analysis and display of genome-wide expression patterns", *Proc. Natl. Acad. Sci USA*, vol. 95, pp. 14863-14868, 1998.

[14] J. DeRisi, V. R. Iyer and P. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale", *Science*, vol. 278, pp. 680-686, 1997.

[15] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization", *Mol. Biol. Cell*, vol. 9, pp. 3273-3297, 1998.

[16] S. Chu, J. DeRisi, M. B. Eisen, J. Mulholland, D. Botstein, P. Brown and I. Herskowitz, "The transcriptional program of sporulation in budding yeast", *Science*, vol. 282, pp. 699-705, 1998.

[17] S. F. Alschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucl. Acids Res.*, vol. 25, pp. 3389-3402, 1997.

[18] G. Salton and C. Buckley, "Text-weighting approaches in automatic text retrieval", *Inform. Process. Manage.*, vol. 24, pp. 513-523, 1998.

[19] T. Joachims, "Text categorization with support vector machines: learning with many relevant features", *Proceedings of ECML-98*, pp. 137-142, 1998.

[20] Y. Guan, C. L. Myers, D. C. Hess, Z. Barutcuoglu, A. A. Caudy and O. G. Troyanskaya, "Predicting gene function in a hierarchical context with an ensemble of classifiers," *Genome Biol.*, vol. 9, (Suppl 1), p. S3, 2008.

[21] A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer prediction," *Appl. Bioinform.*, vol. 3, pp. S75-S83, 2003.

[22] I. H. Witten and E. Frank, "*Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*", Morgan Kaufmann; San Francisco, 1999.

[23] Y. Liu, "A Comparative study on feature selection methods for drug discovery", *J. Chem. Inf. Comp. Sci.*, vol. 44, pp. 1823-8, 2004.

[24] D. Bahler, B. Stone, C. Wellington and D. W. Bristol, "Symbolic, neural, and Bayesian machine learning models for predicting carcinogenicity of chemical compounds", *J. Chem. Inf. Comp. Sci.*, vol. 40, pp. 906-914, 2000.

[25] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, "Molecular prediction of cancer: class discovery and class prediction by gene expression monitoring", *Science*, vol. 286, pp. 531-537, 1999.

[26] D. Rebholz-Schuhmann, H. Kirsch and F. Couto, "Facts from text—Is text mining ready to deliver?", *PLoS Biol.*, vol. 3, no. 2, p. e65, 2005.

[27] Y. Liu, M. Brandon, B. Ciliax, S. Navathe and R. Dingledine, "Text Mining Functional Keywords Associated with Genes", *Medinfo 2004*, San Francisco, CA, Sept. 7th-Sept. 12th, pp. 292-296, 2004.

[28] T. Li, S. Zhu, Q. Li and M. Ogihara, "Gene functional prediction by semi-supervised learning from heterogeneous data", *Proceedings of the 2003 ACM symposium on Applied computing*, pp. 78-82, 2003.

[29] S. Becker, "Mutual information maximization: models of cortical self-organization", *Netw. Comput. Neural Syst.*, vol. 7, pp. 7-13, 1996.

[30] Y. Yang and J. O. Pederson, "A comparative study on feature selection in text categorization", *International Conference on machine Leaning (ICML'97)*, pp. 412-420, 1997.

[31] M. Rogati and Y. Yang, "High-performing feature selection for text prediction", *CIKM'02*, pp. 659-661, 2002.

[32] J. Brank, "Interaction of feature selection methods and linear prediction models", *Workshop on Text Learning* (TextML-2002).

[33] H. Taira and M. Haruno, "Feature selection in SVM text categorization", *AAAI-99*, pp. 480-486, 1999.

[34] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. Sugnet, M. Ares and D. Haussle, "Knowledge-based analysis of microarray gene expression data by using support vector machines", *Proc. Natl. Acad. Sci. USA*, vol. 97, pp. 262-267, 2000.


Received: April 07, 2011 Revised: April 26, 2011 Accepted: May 02, 2011

© Ying Liu; Licensee *Bentham Open.*

This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted, non-commercial use, distribution and reproduction in any medium, provided the work is properly cited.