Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

A Knowledge Driven Regression Model for Gene Expression and Microarray Analysis

Rong Jin, Luo Si, Shireesh Srivastava, Zheng Li, and Christina Chan

Abstract— **The linear regression model has been widely used in the analysis of gene expression and microarray data to identify a subset of genes that are important to a given metabolic function. One of the key challenges in applying the linear regression model to gene expression data analysis arises from the sparse data problem, in which the number of genes is significantly larger than the number of conditions. To resolve this problem, we present a knowledge driven regression model that incorporates the knowledge of genes from the Gene Ontology (GO) database into the linear regression model. It is based on the assumption that two genes are likely to be assigned similar weights when they share similar sets of GO codes. Empirical studies show that the proposed knowledge driven regression model is effective in reducing the regression errors, and furthermore effective in identifying genes that are relevant to a given metabolite.**

I. INTRODUCTION

One of the important research questions in the analysis of gene expression data is to identify the subset of genes that are the most relevant to a given biological process. To identify the genes that are significant to a metabolic function, given the expression levels of genes and the measurement of a metabolic flux under a number of conditions, a linear regression model can be built between the gene expression levels and the metabolic measurements. Then, the weights assigned to the genes by the linear regression model are used to determine the importance of the genes, i.e., the larger the magnitude of the regression weight, the more important the gene. However, the challenge in analyzing micro-array data arises from the fact that even the simplest biological system (e.g., yeast) consists of thousands of genes while the number of conditions that are used to obtain the metabolic measurement is usually no more than a hundred. Hence, there will be numerous ways to weigh thousands of genes that fit equally well with the metabolic measurement under a small number of conditions, and it is difficult to determine which one is the best. This is also called the sparse data problem in statistics [7].

In the past, a number of studies have been devoted to the sparse data problem in linear regression. Many approaches are based on the technique of dimension reduction. The key idea of dimension reduction is to reduce the number of input variables by combining them, either linearly or nonlinearly, to form a small number of latent variables. These latent variables then serve as the inputs to the regression model. Well known dimension reduction approaches include Principle Component Analysis (PCA) [1], Independent Component Analysis (ICA) [3], Linear Discriminative Analysis (LDA) [4], and Partial Least Squre (PLS) [6]. In addition to dimension reduction, regularization is another commonly used approach for the sparse data problem [7]. It introduces a penalty term into the objective function of the linear regression model that favors the sparse solutions. The optimal weights are obtained by minimizing both the penalty term and the regression errors simultaneously.

In this paper, we address the sparse data problem by exploring the prior knowledge of genes from the Gene Ontology (GO) database. We assume that regression weights assigned to genes are determined not only by the regression errors but also by the prior knowledge of genes. In particular, we will minimize the difference between the regression weights of two genes whenever they share the similar set of GO codes. Based on this assumption, we present a knowledge driven regression model, which chooses the weights of genes not only by minimizing the regression errors but also by its consistency with the prior knowledge from the GO database.

II. PRELIMINARIES

Let $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ denote the gene expression data and the metabolic measurements acquired from n different conditions. Each $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d}) \in \mathbf{R}^d$ represents the expression levels of d different genes under the i-th condition, and every $y_i \in \mathbf{R}$ is the metabolic measurement for the i-th condition. The linear regression problem is to find a set of weights $\mathbf{w} = (w_1, w_2, \ldots, w_d) \in \mathbf{R}^d$ for the d genes that minimizes the regression error. Based on the assumption that important genes tend to be assigned large weights than irrelevant genes, we will choose the genes with the largest magnitude of regression weights as the ones that are significant to the given metabolic function.

Following the traditional statistical approaches [8], we can cast the linear regression problem as the following optimization problem:

$$\min_{\mathbf{w}} \quad l_e = \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \tag{1}$$

R. Jin is with Faculty of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA rongjin@cse.msu.edu

L. Si is with Faculty of Computer Science, Purdue University, West Lafayette, IN 47907, USA lsi@cs.cmu.edu

S. Srivastava is with the Dept. of Chemical Engineering, Michigan State University, East Lansing, MI 48824, USA srivas14@egr.msu.edu

Z. Li is with the Dept. of Chemical Engineering Michigan State University, East Lansing, MI 48824, USA lizheng1@gmail.com

C. Christina is with Faculty of Chemical Engineering Michigan State University, East Lansing, MI 48824, USA kris@egr.msu.edu

1-4244-0033-3/06/$20.00 ©2006 IEEE.

The solution to the above optimization problem is

$$\mathbf{w} = (XX^\top)^\dagger X\mathbf{y} \tag{2}$$

where $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. Operator $\dagger$ is the pseudo inverse. To deal with the sparse data problem, the ridge regression [5] introduces the penalty term $\|\mathbf{w}\|_2^2$ into the objective function of Eqn. (1), which leads to the following optimization problem:

$$\min_{\mathbf{w}} \quad l_r = \tau_e \|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \tag{3}$$

where parameter τ_e weights the importance of the penalty term against the regression error. The solution to the ridge linear regression model is

$$\mathbf{w} = (XX^\top + \tau_e I_d)^{-1} X\mathbf{y} \tag{4}$$

where I_d is the identity matrix of size $d \times d$.

III. Knowledge Driven Regression Model

The goal of the knowledge driven regression model is to incorporate the qualitative information of genes from the GO dataset into the numerical regression model. In the section, we will first describe the procedure of converting the GO information of genes into a gene similarity matrix. We will then describe the procedure of incorporating the gene similarity matrix into the regression model.

A. Computing the Gene Similarity Matrix

The key idea of converting the qualitative information of genes from the GO database into the quantitative measurement is through the similarity measurement. To this end, we first represent each gene by the set of assigned GO codes. For the i-th gene g_i, we denote its GO codes by the vector $\mathbf{t}_i = (t_{i,1}, t_{i,2}, \ldots, t_{i,n_i})$ where n_i is the number of terms. We also refer to $\mathbf{t}_i$ as the GO profile for the gene g_i. We then compute the similarity between a pair of genes based on the overlap between the GO profiles of the two genes. The underlying hypothesis is that two genes tend to play similar roles in a biological process when their share similar GO profiles.

A simple way to compute the gene similarity is to treat the GO profile of each gene as a document. The similarity between two genes is then computed based on the overlap between the two documents. We refer to this approach as the "**Document Similarity**" (DS) method.

The problem with the document similarity approach is that it treats any two GO codes as two completely independent identities that do not have any correlation. However, one of the key advantages of using gene ontology is that the GO codes are organized into a hierarchical structure, and the relationship between any two genes can be expressed by their relative positions in the hierarchy. More specifically, two different GO codes can be strongly related if their position in the ontology are close to each other. In order to exploit the correlation among different GO codes, we first compute the pairwise similarity between two GO codes, and then apply the similarity of GO codes to estimate the similarity between two genes. Let us denote the similarity between two GO codes t_i and t_j by $f_t(t_i, t_j)$. We assume that two GO codes are correlated when their paths in gene ontology heavily overlap. In other words, we assume that the similarity between two GO codes is proportional to the length of their overlapped paths. This assumption leads to the following similarity measurement:

$$f_t(t_i, t_j) \quad = \quad \frac{|\{e | e \in \mathcal{Z}_i \wedge e \in \mathcal{Z}_j\}|}{|\mathcal{Z}_i|} \tag{5}$$

where $\mathcal{Z}_i = \{z_{i,1}, z_{i,2}, \ldots, z_{i,m_i}\}$ represents the path for the GO code t_i. However, since the gene ontology is a forest instead of a tree, each GO code could have multiple parents and there could be multiple paths related to each code. Hence, we consider two different ways of calculating the similarity between GO codes. The first approach represents each term by its shortest path in the gene ontology, and directly use the above formulism to compute the similarity between GO codes. We refer to this method as the "**Single Path**" (SP) method. The second approach includes multiple pathes in the computation. In particular, the similarity between any two GO codes is computed as the maximum similarity among all the possible paths of the two GO codes, which leads to the following definition of similarity of GO codes:

$$f'_t(t_i, t_j) \quad = \quad \max_{\mathcal{Z}_i, \mathcal{Z}_j} \frac{|\{e | e \in \mathcal{Z}_i \wedge e \in \mathcal{Z}_j\}|}{|\mathcal{Z}_i|} \tag{6}$$

We refer to this method as the "**Multiple Path**" (MP) method. Given the similarity of GO codes, the similarity between two genes g_i and g_j is then computed as:

$$\text{sim}(g_i, g_j) \quad = \quad \frac{1}{n_i} \sum_{k=1}^{n_i} \max_{1 \le l \le n_j} f_t(t_{i,l}, t_{j,k}) \tag{7}$$

The three similarity measurements defined above, namely the document similarity, the single path based similarity, and the multiple path based similarity, reflect different levels of prior knowledge. The document similarity convey the most shallow knowledge of gene similarity since it ignores the fact that different GO codes could be strongly related. The single path based similarity convey significantly richer knowledge than the document similarity by utilizing the structure of the gene ontology in the computation of code similarity. The multiple path based similarity conveys more information about the genes than the single path based similarity since it exploits the fact that each GO code could have multiple parents in the gene ontology. In this study, we will examine how different levels of prior knowledge encoded in the similarity matrix impacts on the regression accuracy.

B. The Knowledge Driven Regression Model

The above procedure encodes the GO information of genes into the pairwise similarity. We denote the pairwise similarity between any pair of genes by the similarity matrix matrix $S \in \mathcal{R}^{d \times d}$ where each element $S_{i,j} = \text{sim}(g_i, g_j)$ represents the similarity between the genes g_i and g_j. We will describe, in this subsection, how to utilize the pairwise similarity of

genes to reduce the regression error and therefore improve the accuracy of gene selection.

As stated before, the key assumption behind the proposed regression model is that two genes are likely to be assigned similar weights when they share a high similarity in their GO profiles. Based on this assumption, we introduce the following quantity:

$$l_g = \sum_{i,j=1}^{d} S_{i,j}(w_i - w_j)^2 \quad (8)$$

Notice that, if our assumption is true, we would expect that the regression weights $\mathbf{w}$ will result in a small value for the function l_g. This is because, following our assumption, any two genes g_i and g_j will be assigned with similar weights w_i and w_j when their share a high similarity $S_{i,j}$. As a result, we expect the quantity $S_{i,j}(w_i - w_j)^2$ to be small for most pairs of genes, and consequently a small value for the function l_g. We can also write the expression in (8) in the matrix form, i.e.,

$$l_g = \mathbf{w}^\top L \mathbf{w} \quad (9)$$

where matrix L is often called the "**Graph Laplacian**" in spectral graph theory. In particular, matrix L is defined as $L = D - S$, where matrix $D = \text{diag}(D_1, D_2, \ldots, D_d)$ is a diagonal matrix and each diagonal element is calculated as $D_i = \sum_{j=1}^{d} S_{i,j}$.

In order to force the regression weights to be consistent with the gene information from the GO database, similar to the regularized approach, we introduce l_g as the penalty term to the objective function of the ridge regression model in (3). This leads to the following optimization problem:

$$\min_{\mathbf{w}} \quad \mathbf{w}^\top(\tau_l L + \tau_e I_d)\mathbf{w} + \sum_{i=1}^{n}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2 \quad (10)$$

where parameter τ_l and τ_e balances the tradeoff among l_e, l_g, and penalty term $\|\mathbf{w}\|_2^2$. The solution to the above problem is

$$\mathbf{w} = (XX^\top + \tau_e I_d + \tau_l L)^{-1} X\mathbf{y} \quad (11)$$

As indicated in the above expression, we have two parameters, i.e., τ_e and τ_l, to be determined. The value of these two parameters will be chosen through the paradigm of leave one out cross validation, which is commonly used in statistics for determining unknown parameters. More specifically, we will vary the value for both parameters, and for every setup of τ_e and τ_l, we will compute its regression error using the leave one out cross validation as follows: given the gene expression levels and the metabolic measurements under n different conditions, we will leave the first condition out and train the knowledge driven regression model based on the data of the remaining $n-1$ conditions using the fixed value of τ_e and τ_l. We will then evaluate the regression error for the first condition using the regression weights estimated from the remaining $n-1$ conditions. We will compute the same leave one out regression error for each condition, and average the error over n different condition. Finally, we will choose the parameters τ_e and τ_l that provides the minimum regression error averaged over n conditions.

TABLE I
EXPERIMENTAL CONDITIONS

Media	TNF-α	TNF-α	TNF-α
Control	0	20	100
BSA	0	20	100
Palmitate(0.7mM)	0	20	100
Oleate(0.7mM)	0	20	100
Linoleate(0.7mM)	0	20	100

IV. EXPERIMENT

The goal of this experiment is twofold:

- Will the proposed knowledge driven regression model be able to improve the regression accuracy, and furthermore identify the genes that are relevant to a given metabolic function?
- How different ways of computing gene similarity will affect the performance of the proposed knowledge driven regression model?

A. Experimental Data

Microarray gene expression and metabolic data were obtained for HepG2 cells exposed to free fatty acids (FFAs) and tumor necrosis factor (TNF)-α. The experimental design applied in the study is shown in Table I. FFAs and TNF-α were chosen because they have been shown to be associated with many hepatic disorders and to alter hepatic function. Furthermore, in obese people both these factors are elevated. Therefore, it is important to identify how these factors interact to cause changes in hepatic metabolism. For each condition, we obtained cDNA microarray gene expression levels. The data consisted initially of 19458 genes. We then applied the ANOVA variance analysis, Genetic Algorithm/Partial Least Square (GA/PLS) and and Constrained Independent Component Analysis (CICA) to identify a subset of genes whose expression levels were strongly correlated with the measurement of the metabolic process(es) of interest. This selection process identified the 100 most relevant genes. Among these 100 genes, Gene Ontology based classification identified 40 genes in 5 different functional groups that were significantly enriched ($p < 0.01$). The functional groups of lipid metabolism, phosphorus metabolism (signaling), cell death, cell proliferation and electron transport were significantly enriched. Measurements of 49 metabolites were obtained with biochemical assays and HPLC.

Assumptions: In this regression problem, we assumed that the value of each metabolic measurement is a linear combination of the expression levels of the selected genes. We further assumed that there is a different linear regression model for each metabolic process and the regression model is independent of the condition used for obtaining the gene expression levels and the metabolite measurements. Thus, for each regression model, there were 48 regression weights to be determined with only 14 data points (corresponding to the number of conditions).

TABLE II

NORMALIZED REGRESSION ERRORS FOR LINEAR REGRESSION (LR), RIDGE LINEAR REGRESSION (RLR), AND KNOWLEDGE DRIVEN REGRESSION (KLR) USING DIFFERENT SIMILARITY MEASUREMENTS.

LR	KLR			RLR
	DS	SP	MP	
12.9%	14.7%	7.9%	4.4%	14.5%

B. Baseline and Evaluation

We evaluate the quality of the regression models using the normalized regression error, which is defined as follows:

$$err \quad = \quad (y - \hat{y})^2 / \sigma_y^2$$

where y and $\hat{y}$ is the true and the estimated output value, respectively. σ_y^2 is the variance of y, which is estimated from the measurement of the metabolite that are obtained under different conditions. We use the leave one out cross validation to evaluate the proposed regression models: for each left condition, a separate regression model is trained on the metabolite measurement that are obtained for the remaining seven conditions, and the metabolite measurement of the left condition is predicted using the trained regression model. We measure the normalized regression error for each condition and the error averaged over eight conditions and 49 metabolite measurements is used to indicate the quality of regression.

Two baseline models are used in this study. The first baseline model is the straightforward regression that is already described in Section II. The second baseline model is the ridge regression model that is also described in Section II. To determine the optimal value of τ_e in the regularized regression model, we further apply the leave one out cross validation to the training sets that only consist of seven conditions.

C. Experimental Results

In this experiment, we address the question whether prior knowledge of the genes improves the regression accuracy, and if so how does the different levels of knowledge impact the regression results. Table II summarizes the normalized regression errors for the straightforward linear regression, the ridge regression, and the proposed regression model that uses different similarity matrices.

First, we observe that using the simple document similarity does not reduce the normalized regression error. In fact, the normalized regression error is increased slightly from 12.9% to 14.7% when using the similarity matrix based on the document similarity. It is also surprising to observe that the regression error by the ridge regression model is in fact slightly worse than that of the straightforward linear regression model, increasing the normalized regression error from 12.9% to 14.5%.

Second, we observe that the normalized regression error is reduced from 12.9% to 7.9% and 4.4% when using the two similarity measurements that are based on the path overlapping. These two facts indicate that although the gene ontology does provide useful information about genes, the different levels of prior knowledge can have rather different impact on the regression accuracy. In particular, a shallow encoding of the prior knowledge like the document similarity is not as effective in reducing the regression errors as the two path based similarity measurement.

Finally, we analyze the genes that are assigned the weights of the largest magnitude by different regression models. Our analysis indicate that the knowledge driven regression model is more effective than the linear regression model in identifying the genes that are relevant to a given metabolite. For example, we evaluated the genes that are assigned the highest absolute weights by the two regression models for predicting oxygen uptake. The knowledge driven model identified three protein tyrosine phosphatase genes to be most negatively related, and the cytochrome P450 subfamily was identified to be most positively related. Protein tyrosine phosphatase has been known to be a target of reactive oxygen species (ROS) [2] and cytochrome P450 is involved in the oxidation of fatty acids. In contrast, neither of these two sets of genes are assigned with hight weights by the linear regression model.

V. CONCLUSION

In this paper, we present a knowledge driven regression model that incorporates the prior knowledge of genes from the Gene Ontology database. It is based on the assumption that two genes are likely to be assigned similar weights if they share similar GO profiles. Following this assumption, the proposed approach first measures the similarity between any two genes based on the overlap in their GO profiles. This gene similarity information is then used as the regularization term in the linear regression model to help identify appropriate weights. Empirical studies with the metabolic data show the promise in the performance of the proposed approach.

VI. ACKNOWLEDGEMENTS

This work is supported in part by the National Science Foundation (BES 0222747, BES 0331297, and 0425821), the Environmental Protection Agency, National Institute of Health and the Whitaker Foundation.

REFERENCES

[1] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[2] P. Chiarugi and P. Cirri. Redox regulation of protein tyrosine phosphatases during receptor tyrosine kinase signal transduction. *Trends Biochem Sci*, 28(9):509–514, 2003.

[3] P. Common. Independent component analysis, a new concept? *Singal Processing*, 36:287–314, 1994.

[4] J.H. Friedman. Regularized discriminative analysis. *J. Amer. Stat. Assoc.*, 84:165–175, 1989.

[5] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimatin for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[6] A. Hskuldsson. Pls regression methods. *J. Chemometrics.*, 2(3):211–228, 1988.

[7] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal . Statist. Soc. B.*, 58:267–288, 1996.

[8] S. Weisberg. *Applied Linear Regression*. Wisley, New Yor, 1980.