Selecting Informative Genes from Microarray Dataset by Incorporating Gene Ontology

Xian Xu Aidong Zhang
State University of New York at Buffalo
Department of Computer Science and Engineering
Buffalo, NY 14224, USA
xianxu,azhang@cse.buffalo.edu

Abstract

Selecting informative genes from microarray experiments is one of the most important data analysis steps for deciphering biological information imbedded in such experiments. However, due to the characteristics of microarray technology and the underlying biology, namely large number of genes and limited number of samples, the statistical soundness of gene selection algorithm becomes questionable. One major problem is the high false discover rate. Microarray experiment is only one facet of current knowledge of the biological system under study. In this paper, we propose to alleviate this high false discover rate problem by integrating domain knowledge into the gene selection process. Gene Ontology represents a controlled biological vocabulary and a repository of computable biological knowledge. It is shown in the literature that gene ontology-based similarities between genes carry significant information of the functional relationships [3]. Integration of such domain knowledge into gene selection algorithms enables us to remove noisy genes intelligently. We propose an add-on algorithm applied to any single gene-based discriminative scores integrating domain knowledge from gene ontology annotation. Preliminary experiments are performed on publicly available colon cancer dataset [2] to demonstrate the utility of the integration of domain knowledge for the purpose of gene selection. Our experiments show interesting results.

1. Introduction

Cancer classification traditionally relies upon morphological appearance of tumor. However, there has been increasing interest in classifying tumor molecularly with the advent of DNA microarray technologies for large-scale transcriptional profiling [13]. A typical DNA microarray study utilizes several DNA microarray chips on different tissue samples and generates a numerical array with thousands of rows (genes) and tens of columns (experiments/DNA chips). Sample classification can be performed by treating gene expression levels as features describing molecular states of the tissue in question. Various canned classification algorithms in the machine learning literature have been used for such classification task.

Compared to other disciplines DNA microarray datasets bare their unique characteristics. The foremost trait of such datasets is the over abundance of features and lacking of samples, which renders results of any classification algorithms less statistically significant. Using excessive features will normally degrade performance of a machine learner [11] while increasing computational cost at the same time. It is aslo much more difficult for biologists to understand classification results when expression levels of thousands of genes are used for sample class prediction. As a result, gene selection is generally performed before microarray dataset is fed to classification algorithms [2, 4, 5, 12, 15, 16]. However, the statistical soundness of using feature selection algorithms on microarray dataset becomes questionable. On one hand, the threshold of selected discriminative scores needs to be adjusted for multiple selections which has already attracted a lot of research interest [9]. On the other hand, selected genes are not always biologically relevant due to the nature of microarray experiments and the small sample size. When the sample size is limited, even random noise could result in some significant discriminative scores. We will give an example of this in the next section.

In this paper, we propose to use gene ontology [8] annotation to alleviate the problems of gene selection on small sized samples. GO annotation represents a large repository of biological knowledge that is accessible to computational algorithms. Intuition behind our work is simple: while it is likely that even random gene expression can archive relatively high discriminative scores when the number of samples is limited, it is less likely that several random genes annotated with the same GO term all have relatively high discriminative scores. Gene ontology has been incorporated into several microarray data analysis and visualization algorithms/tools for various purposes, in the context of cluster validation [6], visualization of distribution of some scores over GO terms [7]. Authors in [3, 14] concluded that the GO-driven similarity and expression correlation are significantly interrelated. Another avenue of research focuses on detecting over-represented GO terms in a set of co-expressed genes [1].

Our algorithm first examines, for each GO term, if genes annotated with it have statistically higher discriminative scores. This is an indication of correlation between the corresponding GO term and sample class labels. We call these GO terms "informative". Informative genes are then selected from genes that are annotated with such informative GO terms and yield high discriminative scores at the same time. Our algorithm is a generic add-on algorithm that can be attached to many if not all gene selection algorithms.

This paper is organized as follows. In Section 2, we demonstrate the inherent problem of gene selection using single gene discriminative scores. In Section 3, our new GO based gene selection algorithm is detailed. In Section 4, experimental results on publicly available colon cancer dataset are reported. We conclude this paper and give direction of future work in Section 5.

2 The Problem of Gene Selection on Small-Sized Samples

In this section, we give an example of the problems facing gene selection from limited samples. Alon [2] published their experimental results on colon cancer. The dataset consists of 62 microarray experiments on normal and colon cancer tissues. Expression levels of 2000 genes are monitored. After log-transformation, expression levels in this dataset range from 0.76 to 4.32 with mean of 2.30 and standard deviation of 0.49. This is a commonly used benchmark dataset for data analysis algorithms on microarray datasets.

Let us examine the statistical significance of gene selection on such datasets. We use four approaches to assess the false discovery rate of gene selection algorithms based on t-scores by calculating t-scores for randomly generated gene expression arrays. In the first two approaches, we generate random expression arrays of the same size (2000×62) from uniform/normal distribution with same parameters as original log transformed dataset. We also experimented generating from empirical distribution of original log transformed data by assuming the histogram as an approximate probability density function. Later we generate random expression array by randomly reassigning sample class labels. The cutoff t-scores for choosing top 50, 100 and 200 genes from the original log transformed dataset are 3.82, 3.23 and 2.7, respectively. Experimental results summarized in Table 1 show the number of random genes having larger-than-cutoff t-scores in each case. Those random genes would have been selected by t-score based algorithms. All the random experiments are repeated 1000 times.

Table 1. Gene Selection On Random Dataset. Each entry shows the number of genes that score greater than cutoff t-scores in each dataset (original and random generated)

# of Genes *t-scores*	*Top 50* *3.82*	*Top 100* *3.23*	*Top 200* *2.7*
Orig	50	100	200
Rand. uni	1.06 ± 1.04	5.32 ± 2.28	21 ± 4.41
Rand. nor	0.81 ± 0.89	4.69 ± 2.19	19.55 ± 4.43
Rand. emp	0.71 ± 0.83	4.475 ± 1.97	19.5 ± 4.04
Rand. rel	0.86 ± 6.39	4.50 ± 22.35	20.98 ± 62.32

From these experiments we observe that even randomly generated expression levels may result in high discriminative scores. Suppose we merge the original dataset with a random dataset, resulting in a 4000 genes by 62 sample expression matrix. If we were to choose top 200 genes using t-scores from such merged dataset, more than 10% of selected genes would come from the random generated portion of data. The situation becomes more hopeless if we add more random genes or the number of samples is even smaller. When 8000 random genes are added to the merged dataset, the probability of selecting such random genes in top 300 gene list is more than 30%.

3 Integrating Biological Knowledge into Gene Selection Process

One way to overcome this apparent drawback in the feature selection process is the integration of domain knowledge. Biologists have long been doing this. Gene selection is only the starting point in many biological studies. Genes may be addded/removed to/from selected gene set at a later stage pending on other biological evidences. However, to our surprise, there are not many feature selection algorithms proposed in the literature to actually utilize domain knowledge. Although there are already a plethora of online computable biological knowledge bases.

In this section, we propose an add-on algorithm to existing single gene-based gene selection algorithms. Given a single gene-based discriminative scores (of the sample class labels), our algorithm processes the score using biological information contained within GO annotation. This results in a new class of discriminative scores prefixed with the name

”GO adjusted”.

Definition 1 ***Informative Genes*** *are those genes having discriminative scores larger than* θ*, or* $\mathcal{F}(g) > \theta$*. Assume* F *be a single gene based discriminative score.*

Definition 2 ***Discriminative Power*** *of a GO term is defined as the percentage of informative genes among all genes that are annotated with such GO term. Here* $g \in go$ *denotes that a gene* g *is annotated by GO term* go *and* $|go|$ *denotes the number of genes that are annotated by GO term go.*

$$DP(go) = \frac{|\{g|g \in go \wedge \mathcal{F}(g) > \theta\}|}{|go|}$$

Definition 3 ***Informative GO Term*** *is defined as those GO terms go whose discriminative power is larger than* γ *and the number of informative genes annotated with go is larger than* β*.*

$$DP(go) > \gamma \quad \textbf{and} \quad |\{g|g \in go \wedge \mathcal{F}(g) > \theta\}| > \beta$$

Definition 4 ***GO adjusted discriminative score*** *is defined using one single gene discriminative score* $\mathcal{F}(g)$ *and a GO term go, where* $g \in go$*.*

$$\mathcal{F}_a(g,go) = \begin{cases} 0 & \textit{if go is not informative;} \\ \mathcal{F}(g) & \textit{if go is informative.} \end{cases}$$

Definition 5 ***Best GO adjusted discriminative score*** *is defined as the best GO adjusted discriminative score out of all possible GO annotations of a single gene.*

$$\mathcal{F}_b(g) = \max_{\forall go, g \in go} \mathcal{F}_a(g,go)$$

Given a single gene-based discriminative score $\mathcal{F}$ and three parameters θ, β, γ, our algorithm calculates a modified single gene-based discriminative score named “best GO adjusted score.” The basic idea behind our algorithm is straightforward. While the expression levels of random genes may correlate with sample class labels by chance, it is far less likely that majority of these random genes will also have common GO annotation. In other words, it is far less likely that those random genes will have valid biological connections, either participating in the same biological processes, or manifesting the same molecular functions, or being found in the same cellular components.

Informative genes are defined by Def. 1 to be those genes whose single gene discriminative scores $\mathcal{F}$ pass threshold θ. Discriminative power of a GO term is defined by Def. 2 as the percentage of informative genes among all genes that are annotated with the GO term in question. Discriminative power of a GO term with respect to sample class labels measures the collective discriminative power of genes annotated with that GO term. This in turn measures how different biological processes, cell components and molecular functions are affected under different experimental conditions. The higher discriminative power of a GO term, the stronger a GO term is correlated with sample class labels. The value of θ has same range as the corresponding single gene discriminative score, which is the user’s estimate of what a significant score is for $\mathcal{F}$. We further call a GO term informative GO term if such a GO term satisfies the two conditions in Def. 3. First, more than β informative genes needs to be annotated by a GO term in order for that GO term to be called informative. Secondly, the percentage of informative genes among all genes annotated by the GO term needs to surpass threshold γ. These two criteria are set to fend off the effect of random genes. We will discuss how to choose these parameters later.

Single gene-based discriminative score $\mathcal{F}$ is then modified according to the discriminative power of GO terms. To get rid of noisy genes, informative genes are only selected from those informative GO terms. We essentially strengthen single gene-based scores if significant amount of other genes that share common known biological annotation with the given gene are also discriminative of sample class labels. We define “GO adjusted discriminative score” $\mathcal{F}_a(g,go)$ according to Def. 4. The score is 0 if the annotating GO term is non-informative, otherwise it is the same as the single gene discriminative score, or $\mathcal{F}_a(g,go) = \mathcal{F}(g)$. Here we assume single gene-based discriminative score is positive and the larger the score is, the more discriminative the corresponding gene is. Each gene product is annotated with potentially multiple GO terms. We define “best GO adjusted score” $\mathcal{F}_b(g)$ to be the best “GO adjusted score” for a gene among all its annotation. We assume the transitivity of gene annotation in this work. If the direct annotating GO term of a gene is not informative, the gene may still be considered if any parent GO terms of the direct annotating GO term are informative.

Details of our algorithm and complexity analysis are omitted in this paper due to space concern. For details please refer to our technical report.

4 Experiment

Experiment Setup Our primary concern is the false discovery rate of any given single gene-based algorithm. We measure this false discovery rate by repeating experiments on randomized dataset. Our experimental dataset consists of the original dataset from public available data sources and the random portion that are generated as noise. We measure for each single gene-based algorithm the percentage of genes that are selected coming from the randomized portion of data and use this number as an estimate of the false discovery rate.

The original data set is appended with several repetition of blocks of random data. Each random block has the same number of genes and same number of experiments as the original dataset. The expression levels in each random block is generated using normal distribution with same parameters as the original log transformed dataset. Each gene in a random block corresponds to a real gene in the original dataset and share the same GO annotations of it. We call the number of random blocks the random size of a given experiment. For each experimental setup, we generate 200 sets of random data and report the average number of false positive.

Data Preparation The data used in this work are all publicly available. We choose the widely used benchmark microarray dataset: colon cancer [2]. For gene ontology, we downloaded a copy from GO web site at 10/15/2004. We collected GO annotation for genes used in alon colon cancer microarray experiment from SOURCE [10] online database on 11/1/2004. We do notice that SOURCE annotation may have been updated. However, the dataset suffices to test our algorithm.

Not all genes in the original dataset have GO annotation in SOURCE database. However, a decent majority of the original genes have been annotated. For colon cancer dataset, out of original 2000 genes, we found 1495 of them have been annotated with at least one of the 9137 GO annotation. That is, on average, a little more than 6 GO annotation per gene. Since our algorithm relies on gene ontology to provide necessary biological insight, we choose only those genes that currently have GO annotation for further analysis. We expect more and more GO annotation will be available in the future.

Result Table 2 and 3 summarizes our experimental result on colon cancer dataset. It shows the false positive rate measured as the number of random genes being chosen by four single gene-based gene selection algorithms: S2N (signal to noise ratio) tscore and their GO adjusted counterpart. For each experiment we also show the number of overlapping genes, genes that are selected both both original single gene based discriminative scores and their GO adjusted counterparts. From this table, we observe that GO adjusted scores perform consistently better as less random genes (10%) are chosen by GO adjusted scores as informative genes. The trends shown in the tables are clear: the more random genes included the more false positive; the more number genes selected the more false positive. We generally have no knowledge and control of how many genes included in microarray dataset are random with respect to sample class labels. However, we do know now that selecting excessive informative genes from microarray dataset becomes troublesome. We also report the number of genes selected by both original discriminative score and the best GO adjusted score. The percentage of overlapping genes range from 60% to 80%, indicating majority of genes selected by original single gene based algorithms also cluster in GO annotation.

Table 2. Performance of GO adjusted scores as measured in false positive using S2N score

Rand. Size	# of Genes	S2N score	GO S2N score	Diff S2N score	Gene Overlap
1	50	1.40	1.15	0.25	42.45
2	50	2.40	2.15	0.25	39.05
3	50	4.15	3.75	0.40	39.85
4	50	4.05	3.55	0.50	28.64
1	100	5.30	4.09	1.21	68.63
2	100	8.78	7.57	1.21	68.06
3	100	17.24	14.84	2.40	56.38
4	100	15.02	12.55	2.47	59.63
1	200	19.10	17.45	1.65	167.85
2	200	33.15	30.30	2.85	153.90
3	200	43.10	39.45	3.65	144.10
4	200	49.60	46.90	2.70	132.65

Now we describe how we choose the three parameters required by GO adjusted scores: θ, β, γ in these experiments. For θ, we set it to the cutoff score of selected genes. This essentially states that if user is to choose top 100 genes, we deem top 100 ranked genes as informative genes. Using the transitivity assumption of GO annotation, the virtual top GO node in our algorithm annotates every genes in question. Let the ratio between the number of informative genes and the number of genes in the virtual top GO node be γ_0, experiments show good results when γ is set to 1.4-1.6 times γ_0. When γ is set to γ_0, GO adjusted scores revert to original single gene-based discriminative score since the virtual top GO node becomes informative. With the increase in γ, we are taking information in GO annotation more assertively. In our experiments β is set to 1. We tested β values from 0 to 2, results are similar.

5 Conclusion and Future Work

Gene selection plays an important role in the analysis of microarray dataset. Genes that express differently in different sample conditions are selected for further biological investigation. However, due to the limited sample size and complex underlying biology, gene selection algorithms are haunted by excessive false positive rate. In this work, we proposed to integrate biological domain knowledge imbedded in gene ontology and its annotation into the process of gene selection. This is the first attempt of such integration, to our best knowledge. Our experimental result shows this

Table 3. Performance of GO adjusted scores as measured in false positive using t-score

Rand. Size	# of Genes	t score	GO t score	Diff t score	Gene Overlap
1	50	1.40	1.20	0.20	41.85
2	50	2.90	2.25	0.65	38.60
3	50	3.95	3.3	0.65	38.10
4	50	4.90	4.25	0.65	38.55
1	100	5.15	4.26	0.89	66.28
2	100	9.06	7.38	1.68	74.06
3	100	18.11	16.49	1.62	60.14
4	100	15.58	13.20	2.38	67.76
1	200	19.40	17.60	1.80	166.35
2	200	31.95	30.15	1.80	155.30
3	200	44.70	42.30	2.40	136.75
4	200	53.35	49.65	3.70	129.45

is a promising direction. Using gene ontology and its annotation, the probability of selecting random genes in our experiments is reduced more than 10% on average. Our algorithm is a wrapper algorithm upon any if not all single gene based discriminative scores. Our algorithm behaves differently with different choice of one parameter γ, taking GO annotation into consideration within the feature selection process at various degrees. This provides an interesting way to integrate "old" knowledge in GO ontology with new information from microarray experiments.

We ignored genes without GO annotation in this work for simplicity. Although majority of genes (75%) in our study are annotated by at least one GO term, genes that are not currently annotated may be of interest nonetheless. For these genes, traditional discriminative scores can be used instead. Best GO adjusted score defined by Def 5 is comparable to original discriminative scores since they are essentially the same if the annotating GO term is informative. A straightforward way to handle gene selection for genes that are not currently annotated would be to compute original discriminative scores. Such discriminative scores can then be combined with best GO adjusted scores for the purpose of gene selection.

The discriminative power of a GO term is defined to describe the correlation between GO terms and sample class labels. This is different from the previous research in the correlation between gene expression similarity and GO annotation similarity [3, 14], which has some interesting implication by itself. It provides a bridge between gene ontology and disease ontology. The ability of coupling GO terms and disease symptoms may prove useful to provide biologist new insight into disease pathology.

References

[1] F. Al-Shahrour, R. Daz-Uriarte, and J. Dopazo. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, 2004.

[2] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissue probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A.*, 96(12):6745–50, 1999.

[3] F. Azuaje and O. Bodenreider. Incorporating ontology-driven similarity knowledge into functional genomics: An exploratory study. In *In Proc. of IEEE BIBE 2004*, 2004.

[4] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. volume 7, pages 559–83, 2000.

[5] T. Bø and I. Jonassen. New feature subset selection procedures for classification of expression profiles. *Genome Biology*, 3(4):research0017.1–0017.11, 2002.

[6] N. Bolshakova, F. Azuaje, and P. Cunningham. A knowledge-driven approach to cluster validiy assessment. *Bioinformatics*, 21(10):2546–2547, 2005.

[7] J. Cheng, S. Sun, A. Tracy, E. Hubbell, J. Morris, V. Valmeekam, A. Kimbrough, M. Cline, G. Liu, R. Shigeta, D. Kulp, and M. Siani-Rose. Netaffx gene ontology mining tool: a visual approach for microarray data analysis. *Bioinformatics*, 20(9):1462–1463, 2004.

[8] G. Consortium. The gene ontology (go) database and informatics resource. *Nucleic Acids Research*, 32:D258–D261, 2004.

[9] X. Cui and G. A. Churchill. Statistical tests for differential expression in cdna microarray experiments. *Genome Biol.*, 4(4):210, 2003.

[10] M. Diehn, G. Sherlock, et al. Source: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Research,*, 31(1):219–223, 2003.

[11] A. K. Jain, R. P. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, 22(1):4–37, 2000.

[12] P. J. Park, M. Pagano, and M. Bonetti. A nonparametric scoring algorithm for identifying informative genes from microarray data. In *Proc. PSB*, 2001.

[13] D. K. Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics*, 32:502–8, 2002.

[14] H. Wang and F. Azuaje. Gene expression correlation and gene ontology-based similarity: An assessment of quantitative relationships. In *In Proc. of IEEE CIBCB 2004*, 2004.

[15] L. Yu and H. Liu. Redundancy based feature selection for microarray data. In *Proc. of SIGKDD*, 2004.

[16] J. H. Zar. *Biostatistical Analysis*. Prentice Hall, 1998.