Using structure of Gene Ontology for Prognosis and Inference from Microarray data

Shashank Srivastava* Snigdha Chaturvedi† Arnab Bhattacharya‡

Abstract

Recently, there has been much interest in the study and analysis of microarrays for mining valuable biological information, building predictive models for pathological conditions, and unravelling latent correlations signifying biological pathways. Several techniques have focused on identifying differentially expressed genes, and proposed representations of the microarrays through dimensionality reduction techniques to overcome the 'curse of dimensionality'. Statistical tests such as Anova and the Fisher's test, and methods such as clustering and SVD decomposition have been useful in determining differentially regulated genes, and building statistical models of medical conditions from gene-expression values while ignoring noise and normal variations. Recently, the Gene-set approach has been proposed to evaluate expression patterns of gene groups instead of individual genes. These methods, however, do not allow direct inference of relations between gene-expression values and genetic concepts. This study provides an approach to extend the statistical method using additional knowledge infusion from the structure of the Gene Ontology database. We propose a hierarchical approach towards data representation, which bypasses several limitations of existing methods, and can directly yield biological understanding and interpretability. The proposed method is tested on two standard datasets. Prognostic predictions from our method are seen to validated precisely by existing biological literature and highly specific control-studies. The proposed representation also shows predictive potential, and classification accuracies from our novel representation scheme using decision trees compare favorably with statistical methods.

Keywords: Bioinformatics, Microarrays, Dimensionality-reduction, Gene-Ontology, Inference

1 Background and Motivation

Microarray is a fast growing technology, which can measure expression levels of thousands of genes across the genome simultaneously at both transcriptional and DNA levels. A sample microarray with discrete recordings is a matrix where each row corresponds to an 'experiment' and the columns correspond to genes (that could possibly influence the experiment). The 'experiments' can measure the expression levels of genes in two or more variants of cells, such as diseased or normal cells. Hence, a single microarray can represent the expression level of tens of thousands of genes in different medical conditions (or diseases). Since microarray data presumably contains a large amount of latent information about biological processes and genetic phenomena, two relevant questions would follow naturally:

1. Can gene-wise differences in gene expressions help in distinguishing among medical conditions?

2. Can the data lead to a better understanding of involved etiology and biological processes?

The first question has received much attention for its implications and potential in medical prognosis. Classical machine learning and data mining algorithms for microarrays, however, must face the essential question of a large number of attributes (genes), and few training instances [35, 36] (experiments). This often leads to the problem of insufficient training for both parametric and non-parametric methods, or over-fitting which results in poor predictive performance. Most predictive models ameliorate the 'curse of dimensionality' by either discarding some genes altogether, or through approximative projections such as PCA.

Methods for dimensionality reduction using the K-means algorithm, hierarchical clustering, self-organizing maps (SOMs) and other models have been explored. Notable methods include the Gene Shaving clustering approach by Tibshirani et al, and singular value decomposition of expression values by Alter et al. [6].

Methods that test statistical significance of differing distributions of individual genes in different medical conditions, and which choose the most significantly differing genes have often been studied. However, this general approach suffers from several shortcomings. Firstly, choosing individual genes as attributes violates the iid (independent and identically distributed) assumptions

*Department of Computer Science, Indian Institute of Technology Kanpur

†IBM India Research Labs, New Delhi

‡Department of Computer Science, Indian Institute of Technology Kanpur

of features, since differentially regulated genes in a biological process are seldom expressed clinically in independent ways. Secondly, the setting of a threshold for the statistically significant genes is usually a non-trivial problem [8, 7]. The list of differential genes hence obtained by this general method, is then mined using biological literature for finding correlations and associations. Dragichi et al. [38] propose a global functional profiling method that, given a list of regulated genes, gives a statistical p-value to gene profiles. The magnitude of p-value is determined by comparing the number of differentially regulated genes in each gene profile with the expected number from a hypergeometric distribution. This approach is shown to be meaningful in terms of biological interpretability and discovery, but suffers from the mentioned limitations of the individual gene approach, and does not take directly into account actual expression levels of individual genes.

Another issue with hypothesis testing of individual genes is that while the method can provide lists of differentially expressed genes, discarding other genes can often result in loss of significant information. This problem is made worse in cases when sets of genes show subtle, synchronized expressive behaviour (thus making the iid assumption even worse). This has led to research using Gene sets, pioneered by Mootha et al. Their study, proposing the GSEA approach, revealed that genes involved in oxidative phosphorylation are only modestly but synchronizedly underexpressed in human diabetic muscle, and are related to significant variations in human metabolism. An individual gene approach fails in the scenario. Gene-sets using various biological sources such as the Swiss-prot database [9] and the Gene Ontology [10] have been used in several studies.

Still more importantly than predictive modeling that can aid prognosis, molecular level information in form of expression levels can lead to higher level biological interpretations, assisting better understanding of the processes involved, and identification of critical biological and etiological pathways. The statistical hypothesis testing approach of individual genes can provide lists of differentially expressed genes in two medical conditions. Currently, biological cues are arrived at through searching these genes in genetic and biochemical literature for co-occurrence and correlations. In this work, our proposed approach is aimed at assisting the analytic process and bypassing some of the limitations of the above approaches. For an understanding of latent processes and pathways that underlie gene expression values at a given biological snapshot, it would be helpful if a given disease instance could be viewed quantitatively in terms of genetic concepts. We attempt to formulate such an approach of representing a disease instance using Gene Ontology (GO) terms, and using the sets of differentiating GO terms for making biological inferences.

Our contributions in this paper are:

1. We formulate and validate a concise representation scheme for microarray experiments that infuses hierarchical biological context from the Gene Ontology using GO concepts. This representation is intuitive in terms of interpretability and biological understanding.

2. The method identifies important biological phenomenon and the gene functions that differentiate two disease conditions. The distinction can guide biologists and biochemists in understanding and exploring pathological and general physiological phenomenon.

3. In the process, we also explore an alternative scheme to overcome the high dimensionality associated with predictive models for microarray data. In this sense, the classification potential of the representation also establishes the stability of our approach.

The structure of the paper is as follows: In Section 1.1, we review the structure of the Gene Ontology and describe two standard datasets used in this work. In Section 2, we explain our representation scheme for microarray experiments; and describe our general approach in detail. In Section 3, we make biological inferences about clinical conditions for each of our datasets, and validate them with empirical evidence from biological literature. We also look at the performance of the proposed representation on a classification task over the two datasets, and compare the representation scheme with two clustering based approaches. Section 4 concludes with a summary of the proposed method, shortcomings, and suggested directions of further enquiry.

1.1 Preliminaries

1.1.1 Gene Ontology The Gene Ontology (GO) [29] is a large directed acyclic graph (DAG) that provides a controlled vocabulary (or ontology) to describe attributes of genes and gene products of organisms. The GO contains three structured ontologies that describe genes in terms of the following properties:

1. Biological processes
2. Molecular functions
3. Cellular components

The GO ontology is hierarchical in nature, and can be visualised as a DAG rooted at a node *GO* and

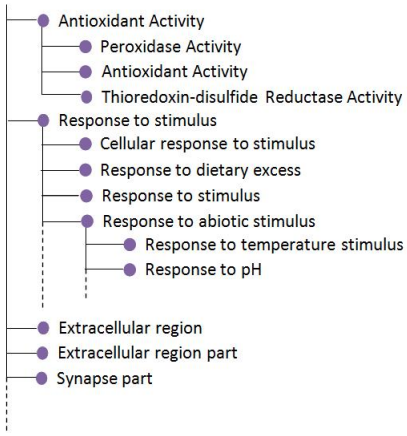

Figure 1: Hierarchy in the GO

various GO annotations forming its descendant nodes. Any two connected nodes are related by 'is a' or 'part of' relations. Genes associated with a particular GO concept appear in the GO graph as the children of the node. An illustrative segment of the GO structure is shown in Figure 1. The hierarchy of the GO is intuitively seen: "Response to temperature stimulus" is a child node of "Response to abiotic stimulus" which in turn is a child node of "Response to stimulus". A gene maybe related to two or more or more genetic functions, and hence two or more GO terms (which ocan be mutually uncorrelated, and occur in different parts of the GO graph). Hence there can be cross edges between nodes in the GO structure. However, the graph shows an acyclic structure since edge relations to genes only show directional GO term to Gene mappings.

1.1.2 Datasets Two freely available datasets have been used in the current study: the Leukemia microarray dataset [32] and the DLBCL microarray dataset [31]. The Leukemia dataset consists of 5147 gene expression profiles of 72 patients with 47 instances of Acute Lymphoblastic Leukemia (ALL) and 25 instances of Acute Myeloid Leukemia (AML). The DLBCL dataset consists of 7070 gene expression profiles of 77 patients, 58 of which were suffering from Diffused large B-cell lymphoma (DLBCL) and 19 from follicular lymphoma (FL).

2 Methods and Technical Solutions

While many statistical methods based only on expression values of genes have been useful for the purpose of classification, purely statistical method neither incorporate domain knowledge, nor provide any insights about the biochemical and physical processes associated with the concerned genes and diseases. In our approach, we represent a disease instance in terms of GO concepts, and identify differentiating GO concepts in this representation to draw biological inferences. The essential steps in our method can be briefly summarized as follows:

1. Create a functional profile for each gene expressing relative involvement of the gene in various GO concepts (gene functions) present at a chosen granularity (distance of GO concept from root node) in the Gene Ontology.

2. Express each disease instance of the microarray in terms of identified GO concepts using the functional profile and degree of expression (up/down regulation) of the genes, and the relaitve association of any gene with a GO concept.

3. Identify GO concepts with differential expression in the disease classes. Rank GO concepts by interclass distribution difference or predictive ability to draw biological inferences.

Hereon, we visualize a microarray as a $K \times M$ array, where K is the number of experiments (or disease instances) and M is the total number of genes being monitored by the microarray.

2.1 Step1: Creating Functional Profiles We describe a gene by its relative appearance within different contexts of specific GO concepts(s). Each of the M genes is represented in terms of the GO terms within any context which it occurs in the Gene Ontology. Since GO terms at the lower abstraction levels of the GO are very specific, and too many in number to be interpretable, we initially consider the granularity of GO terms at the third level of the GO tree (starting with the root, labelled *GO*, as the first level). Hence, we have a manageable, but sufficiently descriptive list of GO terms spanning the three ontologies for biological processes, molecular functions and cellular components.

Suppose the genes in the dataset are collectively found to be involved in a total of N distinct GO terms, at the third level of the GO tree. Each of the M genes is then represented as an N-dimensional vector (called the functional profile of the gene), with the i^{th} component describing the number of times the gene appears in the directed sub-graph of the GO rooted at the corresponding i^{th} GO term. This is illustrated in table 1, where functional profiles of genes are visualised as an $M \times N$ array, and numbers in a cell describe the frequency of occurrence of a gene within all contexts of a GO term at the third level of the GO tree. It is important to note that a gene can be counted multiple times at different levels of granularity of the GO. The

underlying assumption is that higher occurrence of a gene (within different contexts of a generic GO concept) implies a greater relation between the gene in particular and the GO term

Table 1: Functional Profiles of genes for a typical microarray

Gene	GOterm 1	GOterm 2	.....	GOterm N
gene 1	1	0	.....	2
gene 2	0	3	.....	1
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
gene M	1	1	.....	0

2.2 Step 2: Representation by Gene Functions

We estimate the degree of up/down regulation of each of the M genes present in the dataset, relative to normal levels of expression. Upregulation of a component at a biological instance is mirrored by a higher expression level of corresponding genes. Conversely, downregulation is coordinated with a decrease in expression of the corresponding gene. In this work, we implicitly assume that deviation in gene expression levels reflect upregulation and downreguation of corresponding gene products and corresponding proteins.

Regulation of gene expression is correlated with increase and inhibition of related gene functions associated with corresponding GO terms either directly, or indirectly through regulatory pathways. Assuming a linear model of expression, the degree of up/down regulation is ideally estimated as the difference between an expression level and the expression level for normal (average) cells. We estimate the average expression level of the population as:

$$\mu_i = \frac{\sum_{j=1}^{t} C^j \mu_i^j}{t} \tag{2.1}$$

where,
μ_i is the estimated mean expression level of gene g_i for the population,
μ_i^j is mean expression level of gene G_i for class j,
t is the number of (diseased and normal) classes.

An experiment (disease instance) k, can now be represented as a vector, $\overline{y_k}$, of GO terms in the following manner:

$$\overline{y_k} = \sum_{i=1}^{i=M} (x_{i_k} - \mu_i)\overline{f_i}, \tag{2.2}$$

where,
x_{i_k}=expression level of gene G_i in disease instance k,
$\overline{f_i}$=functional profile of gene G_i.

In this approach, the expression level of a gene that occurs more frequently in context of a GO term (and hence, in a large number of processes in context of the GO term) is given a higher weight. Hence, each of the K disease instances (or, experiments) in the dataset is represented as a numeric feature vector of N GO terms instead of M genes. Since $N << M$, the number of 'dimensions' is significantly reduced. However, not all of the GO terms are significant for statistical inference, as explained in the next subsection.

We now look at the possibility of inferring biological concepts from differences in expression of the N GO terms in the concerned disease classes. For most genes, expression levels behave similarly for different disease classes.

2.3 Step 3: Inferring Biological Concepts

In our case, populations in both datasets consist of binary classes. An observation of class distributions of the components of individual GO terms on both datasets suggests that most GO concepts are similarly represented in the two classes (see Figures 5 and 6).

In both datasets, however, a small number of GO concepts are seen to be very significantly different in the two concerned classes (Figures 3 and 4). These GO terms could be expected to characterise differences between the classes, and have major discriminatory potential. While the number of differentiating GO terms at this level is tractable (≈ 10), and analysis at this level can benefit from domain knowledge of the specialising biologist, we rank the GO concepts on the basis of class separability. The following distance measures for distributions were tried for ranking GO terms:

1. Euclidean and Manhattan distance
2. Student's two sample t-test
3. Bhattacharyya coefficient [2]
4. Histogram Match distance [1]
5. Kullback Leibler divergence [3, 4]

Some of the most differentially expressed GO terms in the two classes were commonly identified by most of theese measures. Based on classificatory potential of the best features on a cross-validation task, t-test score was marginally found to be most reliable measure. This is inspite of the inexact assumption of Gaussian distributions of GO term expressions. The t-test, designed in 1915 by Gosset for the Guinness Breweries of

Dublin [**?**], tests whether the means of two distributions (assumed to be normal) are significantly different taking into account the variance of the distributions. In all further work, t-test scores were used to identify the best features. However, while higher statistical scores can imply a greater differentiating power and significant differences in distributions within the disease classes for a GO term, it is suggested that the class distributions be observed by medical/biochemical specialists in addition to calculating score since in a few cases, even smaller variations in some critical gene functions may have greater biological significance.

Relative expression of the differentiating GO terms thus identified could potentially yield important biochemical pointers which might assist biochemists in their analysis of diseases. For example, "higher level of viral reproduction activity in DLBCL than FL" is one such non-trivial pointer for the DLBCL dataset, that gets substantiated by empirical facts in Section 3.

The approach in this work offers scope for subsequent mining of the Gene Ontology. Often, a particular GO term (say, GT1) can have high differentiating power, but might be too general to make any significant inference. For example, "Response to stimulus" is one such function which does not yield any significant biological understanding at the first level, and there exist several types of stimuli. The routine described in Sections 2.2 and 2.3 can be iteratively called to mine deeper into the GO, starting at GT1. For this purpose, functional profiles of the M genes consider only GO terms that occur directly below GT1 in the Gene Ontology.

In our experiments, we observed that on moving to finer levels of GO, the differentiating score for a few GO terms is often distinctly greater than rest of the GO terms indicating the precise factor (sometimes gene) responsible for differentiating the disease classes. For example in case of Leukemia dataset, Figure 2 shows that, the score of the GO term, "Response to temperature stimulus", is significantly greater than other children of "Response to abiotic stimuli" (the t-test scores have been mentioned in parentheses in the figure) which implies its greater contribution in differentiating among the disease classes of Leukemia dataset than other abiotic stimuli. In other cases, no single factor is individually very differentiating, illustrating the utility of the hierarchical structure of the GO in forming intuitive gene sets at different levels. For example in case of Leukemia dataset, one of the GO terms "Extracellular region part" was observed to be a differentiating GO term with a high t-test score of 4.71 (see Figure 2). However, it was observed that none of its children GO terms had a remarkably high t-test score (the highest being 3.50) implying the possibility of further investigation into its role in differentiating the two disease classes.

Discussions on inferences drawn for both datasets, along with validating evidence, and their predictive power are described in detail in Section 3.

3 Empirical Evaluation

Context: ALL and AML are both fast-growing cancers of the blood and bone marrow. Although the risk factors for the two diseases have been identified, the specific causes of AML and most ALL are still unknown.

DLBCL [28] and FL are two commonly occurring types of lymphoma of the B cell lineage with DLBCL being the more aggressive lymphoma while FL being the relatively indolent category. About 25-60% of the FL cases transform to DLBCL after which they become indistinguishable [15].

Biological Inferences:
In this section, we observe differentials in distributions of the GO terms to make biological inferences using the representation in Section 2.2, and evaluate them against existing biological literature.

3.1 Leukemia Dataset The Leukemia dataset has occurrences of 55 GO terms at the third level of the Gene Ontology. Observation of class distributions of the GO terms with high differential scores (as calculated in Section 2.3) led to the identification of primarily 4 differentiating GO terms at this level.

Figure 2 summarises the results for the Leukemia dataset. It shows some of the relevant GO terms at the 3^{rd}, 4^{th} and 5^{th} levels of GO. The GO terms mentioned in bold fonts are the ones that were found to be most differentiating using our method, the rest being the non-differentiating ones. The figures in parentheses indicate corresponding t-test scores of the GO terms.

Figures 3, 4, 5 and 6 show distributions of the two most differentiating functions ("Antioxidant Activity", "Response to Stimulus"), and two other randomly chosen GO terms ("Reproductive Process", "Rhythmic Process") exhibiting difference between differentiating GO terms, and the typical pattern of similar distributions in both classes.

We next outline the inferences from our method aalong with corroborative evidence from biological literature whenever obtainable. On the basis of differentials observed in histogram distributions for GO terms at third level of abstraction, ALL seems to be generically characterized as distinct from AML in the following terms:

1. **Inference 1:** Higher anti-oxidant activity in AML

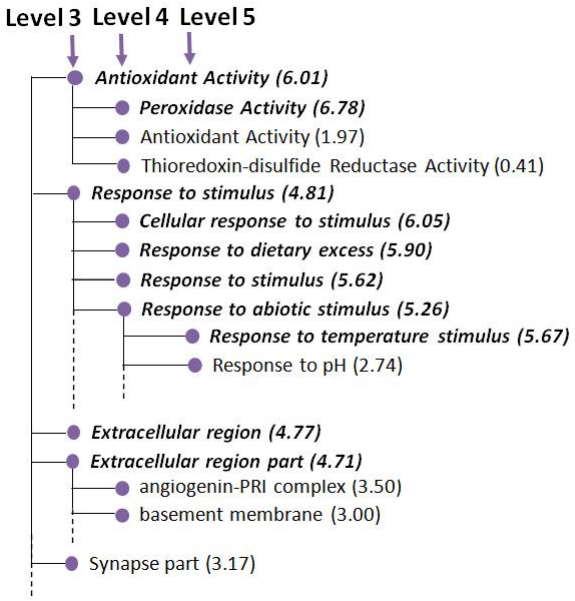

Figure 2: Significant GO terms and corresponding t-test scores (mentioned in parentheses) at various levels of the GO for the Leukemia dataset

Evidence: "Superoxide dismutase(SOD) is dimeric antioxidant enzyme .... The overall mean SOD levels of leukemic patients (AML, ALL and CML) were significantly low compared to that of the normal controls, the reduction being more prominent in ALL." The SOD levels in case of AML and ALL were measured to be 120.21 ± 16.31 and 111.68 ± 5.25 [34].

2. **Inference 2:** Greater response to stimulus in AML

3. **Inference 3:** Lower Activity related to extracellular region and extracellular region part in case of ALL

However, the GO terms at this level (such as "Response to stimulus") are too general to be medically significant. We therefore, mine the GO to deeper levels for two of the diffentiating GO terms obtained at this level: "Antioxidant Activity" (4^{th} level) and "Response to Stimulus" (5^{th} level). These results are also shown in Figure 2. As a result of this deeper mining, we could draw the following inferences:

1. **Inference 1:** Higher peroxidase activity in AML
   Evidence 1: "The peroxidase and esterase stains are particularly useful for the diagnosis and accurate morphologic classification of AML. Unequivocal peroxidase positivity is generally diagnostic of AML." [16]
   Evidence 2: It was also experimentally observed that the mean peroxidase activity value in case of AML and ALL was -12.6 and -0.6 respectively. [18]

2. **Inference 2:** Greater cellular response to stimulus in AML

3. **Inference 3:** Greater response to abiotic stimulus (specifically temperature stimulus) in AML
   Evidence 1: Work of [20] "demonstrate that a temperature-sensitive p53 induces temperature-dependent decreases in the expression of the apoptosis-suppressing gene bcl-2 in the murine leukemia cell M1". It is noteworthy that M1 is a subclass of AML.
   Evidence 2: A study conducted by [19] shows that low grade B cell ALL are CD52 positive while Tcell ALL and almost all AML are CD52 negative. And, temperature is one of the principal factors synergistically modulating epididymal CD52 expression [21, 22]

The relevance of these predictions is strengthened by the following excerpt. "Recent studies are evaluating CAMPATH-1H, which is an antibody to CD52, as a novel method of T cell depletion. In ALL, Campath-1H treatment may have the added benefit of anti-leukemia activity since malignant lymphoblasts express CD52." [17]

We observe a high t-test score for the GO term "Response to dietary excess" at the 4^{th} level of the GO, inspite of a lack of concrete evidence. However, a positive correlation of AML with dietary excess is suspected, especially for infants [11]. The proposed approach seems to support this hypothesis, and indicates a direction for empirical medical inquiry.

We also see that some GO terms identified are too general (eg. "Extra-cellular region", "Response to stimulus"), and don't directly map to specific biological functions/concepts. Mining of these pointers can be expected to be expedited with improvements in robustness of the Gene Ontology structure.

3.2 DLBCL dataset The results of mining of the DLBCL dataset have been summarised in Figure 7. For this dataset, there were 57 GO terms at the third level of the GO tree. Most GO terms are equally expressed in the two disease classes, DLBCL and FL. However, 4 of the 55 GO terms were seen to be significantly different in the two classes, as shown in Figure 7.

Distributions of two differentiating GO terms ("Viral Reproduction" and "Antioxidant Activity") are shown in Figures 8 and 9. Several biological predictions could be made at this level, distinguishing DLBCL from FL, the major ones being:

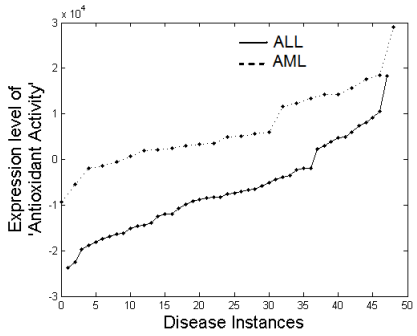

Figure 3: "Antioxidant Activity" in Leukemia dataset (t-test score = 6.01)

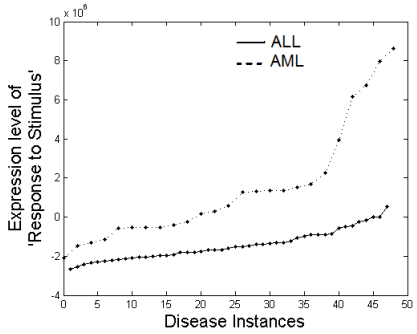

Figure 4: "Response to stimulus" in Leukemia dataset (t-test score = 4.81)

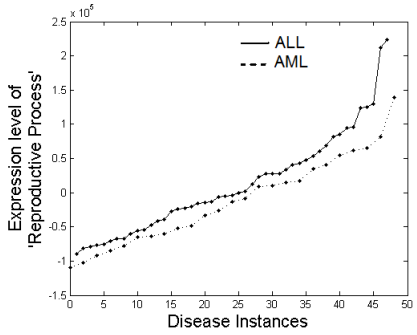

Figure 5: "Reproductive Process" in Leukemia dataset (t-test score = 1.46)

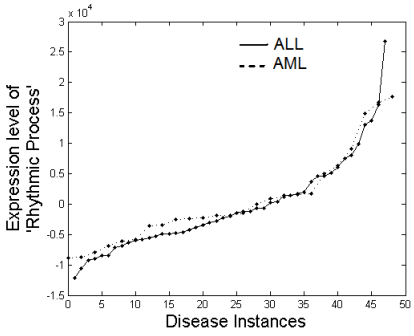

Figure 6: "Rhythmic Process" in Leukemia dataset (t-test score = 0.47)

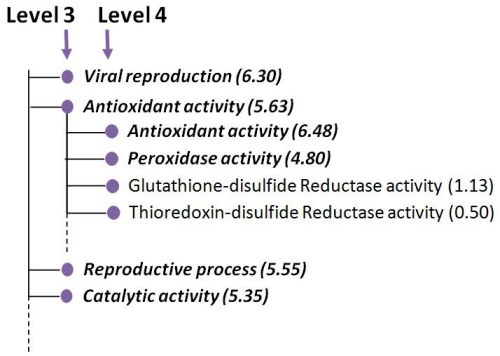

Figure 7: Significant functions and t-test scores (mentioned in parentheses) at various levels of the GO for the DLBCL dataset

1. **Inference 1:** Higher levels of viral reproduction in DLBCL
   Evidence 1: "In HIV patients, 3 histologic subtypes of NHL are more common: high grade immunoblastic, diffuse large B cell (DLBCL) and Burkitt lymphomas. Indolent B cell lymphoma represents only a small proportion of all lymphomas occurring among HIV patients. Follicular lymphoma (FL) is generally considered an indolent NHL characterized by a follicular pattern." [5].
   Evidence 2: "The possible role of a co-infection with Epstein-Barr virus (EBV) has also been investigated ..... EBV DNA was detected in 40.0% of FL, in 72.7%of DLBCL" [12].
   Evidence 3: "The HCV rates for patients with different lymphopropliferative diseases were 7.5% (27/361) for DLBCL, ... and 4.2% (4/96) for FL." [14].
   Evidence 4: "In conclusion, persistent human hepatitis virus infections, especially HCV, may play an important role in the tumorigenesis of splenic DLBCL in Japan." [13]

2. **Inference 2:** Lower anti-oxidant activity in FL

3. **Inference 3:** Higher reproductive process and catalytic activity in DLBCL

The GO terms present at finer levels were considered for a more specific analysis of "Antioxidant Activity" in subsequent mining (illustrated in Figure 7). At this level, it can be reasonably inferred that in addition to antioxidants in general, peroxidase plays an important role in differentiation of DLBCL from FL. Following are the inferences predicted at the finer level.

1. **Inference 1:** Lower anti-oxidant activity in FL
   Evidence 1: Calculated expression levels of Acidic Phospholipase A2 (PLA_2), an anti-oxidant protein, was plotted for cases of DLBCL and FL (Figure 10 reproduced from [27]) and it was observed that the mean expression level in case of DLBCL was much larger than in FL [27].
   Evidence 2: Secondary DLBCLs (which transform from FLs) have lower levels of anti-oxidant inhibitors [26].

2. **Inference 2:** Lower peroxidase activity in FL
   Evidence: Evidence suggests that DLBCL and FL have different Bcl-2 protein expression levels and that there exist a relationship between Bcl-2 and Glutathione-peroxidase (GSH-PO). The following excerpts illustrate the facts.
   Evidence 1: "Bcl-2 gene rearrangement was seen in the vast majority of Jordanian FL cases and approximately one third of all DLBCL cases.....The presence of bcl-2 gene rearrangement in DLBCL may define a subset of lymphomas that may be biologically and clinically unique and different from the rest of DLBCL." [23]
   Evidence 2: "Expression of bcl-2 protein was shown in 51% DLBCLs. Bcl-2 protein was expressed in 89% of FLs with t(14;18), in contrast to 25% of FLs without t(14;18)." [24]
   Evidence 3: "These findings strongly suggest that the expression of Glutathione-Peroxidase (GSH-PO) and Bcl-2 in the glandular epithelial cells of the rat ventral prostate is testosterone-dependent. Furthermore, co-expression of GSH-PO and Bcl-2 in the prostatic cells are considered to be normal or adaptive aspects of the cells." [25]

Strong evidences were found vindicating predictions of the proposed approach regarding antioxidants, peroxidase and viral activity. Independent studies suggest a positive correlation of people infected with HIV, EBV and HCV, with DLBCL rather than FL.

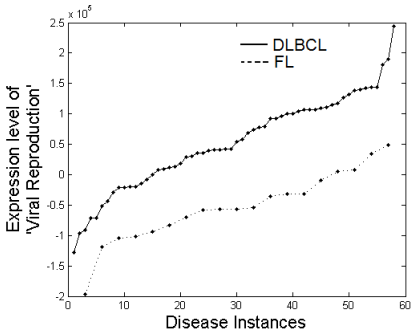

Figure 8: "Viral Reproduction" in DLBCL dataset (t-test score = 6.30)

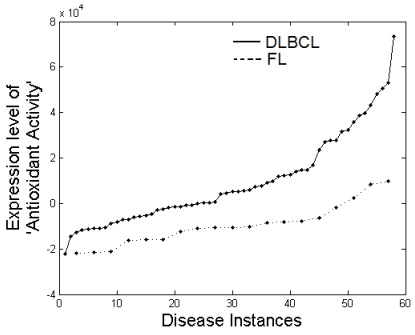

Figure 9: "Antioxidant Activity" in DLBCL dataset (t-test score = 5.63)

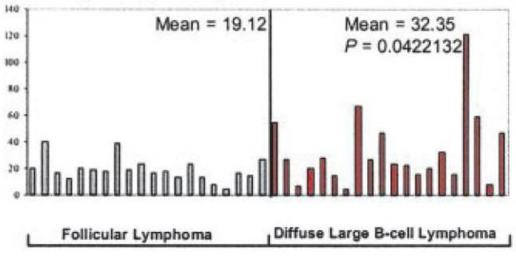

Figure 10: Expression levels of PLA_2 (antioxidant protein) in cases of FL and DLBCL (reproduced from [27])

3.3 Classification The approach described in this work offers a concise representation of the microarray data in terms of GO concepts. Predictive models in terms of this representation can be more informative in terms of interpretability.

In this section, we test the proposed representation for claissifcatory potential. It is understood that the representation scheme is not formulated with an intention to building predictive models, and due to small sizes of datasets at hand, the comparision is not entirely rigorous. However, any meaningful representation would be expected to have reasonable classificatory potential. Here, we compare the proposed representation against the standard K-means methods based on clustering, and show that the proposed method yields a decent classification accuracy. Much work has been done, involving standard machine learning algorithms, especially involving clustering of genes by expression values, and principal component analysis. In particular, a variety of clustering methods have been used for mining microarray data [38, 39, 37].

3.3.1 Expression-based clustering Recently, clustering methods have been employed to cluster genes in a logical order, with similar genes grouped together. The basic principle behind this method [33] is that among a comprehensive set of genes in a microarray, several genes may have partially redundant information in terms of expression values profile. Hence, it is proposed that 'similar genes', i.e. genes having redundant information should be clubbed together into one cluster and, hence, each cluster can now be represented by a *prototype gene*.

The K-means unsupervised learning algorithm is used to cluster genes on the basis of expression levels in the recorded experiments. The centroid of the cluster now represents genes belonging to a single cluster. The expression level of the estimated *prototype gene* is, hence, the arithmetic mean of genes belonging to the corresponding cluster. *Prototype genes* thus obtained are used as the feature set for building a classifier. The number of clusters (say, C) is an important consideration in this approach. Very small (or large) value may result in an insufficiently small feature space (or overfitting). Classifiers with different values of C were built, and the best performing results reported.

3.3.2 Clustering by functional profile As a more biologically meaningful alternative, we cluster genes on the basis of their positions of occurrence in the Gene Ontology, instead of the expression values, using the K-means clustering algorithm. Since the clustering is based on the functional profiles, genes belonging to same cluster are expected to have same relative involvement in gene activities and hence are expected to form a meaningful cluster. For each cluster, we define the expression level of the *prototype gene* as the arithmetic mean of the expression levels of all the genes in it. The number of clusters is again empirically determined. The value corresponding to the best performing classifier is reported.

3.3.3 Our method The method described in Sections 2.1 and 2.2 leads to a representation where each disease instance is represented in terms of a few GO terms. This representation can be used for classification. Since, generally the number of GO terms considered, N, is comparable to the number of instances, K, in the microarray, using all attributes may lead to overfitting. Hence, the ten most differentiating GO terms (ranked according to highest t-test scores) are chosen as composing the feature vector.

3.4 Validation Feature sets with attributes as described above were used to build classifiers. WEKA's[30] Functional Decision Tree was used for classification, and predictions compared against the annotated ground truth. All models were validated using Leave-One-Out Cross Validation (LOOCV).

Classifiers built from the three approaches described were evaluated comparing predictions of the classificatory models with annotated ground truths for both data. Although both datasets had a considerable skew in class distribution, no resampling was done, or metaheuristics used; so that performances reflect the base classifiers. A detailed summary of the classification results is given in Table 2.

Clustering by gene functional profiles is seen to perform worse than either of the other two classifiers on the (easier) Leukemia dataset, but does better than both on the harder DLBCL dataset. This approach needs further exploration with respect to cluster composition, and comparison with clustering by expression values.

The results indicate that the GO term based representation approach performs as well as the K-means clustering on both datasets. While the primary aim of our study is not to build classifiers with high classification accuracy, the performance of the classifier suggests classificatory potential in such a representation, and establishes the stability of the approach. The potential of such a representation for statistical modeling can be further explored.

4 Discussion

Our approach suggests a concise representation scheme for microarray experiments that infuses biological context from the Gene Ontology in terms of GO terms,

Table 2: Performance of classifiers for the two datasets

Datasets	Expression based method		GO based method		Our approach	
	No. of features	LOOCV Accuracy	No. of features	LOOCV Accuracy	No. of features	LOOCV Accuracy
Leukemia	15	95.8%	10	90.3%	10	94.4%
DLBCL	25	88.3%	11	90.6%	10	85.2%

accounting for the relative importance of every gene in context of the GO term. This representation is intuitive in terms of interpretability and biological understanding, and can identify and predict important biological phenomena and gene functions that differentiate two disease conditions. Subsequent mining of Gene Ontology within the differentiating GO terms is seen to work well, and can give an in-depth analysis of a medical condition.

Predictions of the method (with respect to both datasets) can serve as pointers for biological research for both datasets. The approach can be especially useful in case of diseases like leukemia, where underlying biological processes and critical pathways are still unknown. It can also assist in understanding the etiology of various diseases. Mining of non-specific pointers that do not directly map to biological concepts needs to be explored.

The proposed representation also offers a go-around for the curse of dimensionality, and shows potential in terms of classification accuracy. Further exploration in this regard can lead to better predictive models.

References

[1] M Werman, S Peleg, and A Rosenfeldf,*A Distance Metric for Multidimensional Histograms*, Computer Vision, Graphics and Image Processing, (32), Elsevier Science, pp. 328-336.

[2] A Bhattacharyya, *On a measure of divergence between two statistical populations defined by their probability distributions*,Bulletin of the Calcutta Mathematical Society,(35) (1943), pp. 99-109.

[3] S Kullback, *The Kullback-Leibler distance*, The American Statistician, (41), 1987, pp. 340-341.

[4] S. Kullback, and R.A. Leibler, *On Information and Sufficiency*, The Annals of Mathematical Statistics, 22(1), Institute of Mathematical Statistics, 1951, pp. 79-86.

[5] Elisabete Moreira, Carmen Lisboa, Sergio Alves, Ilidia Moreira, Elsa Fonseca, and Filomena Azevedo, *Cutaneous lesions as the first manifestation of systemic follicular lymphoma in an HIV patient*, Dermatology Online Journal, 14 (7), UCD Department of Dermatology, 2008

[6] Orly Alter, Patrick Brown, and David Botstein, *Singular value decomposition for genome-wide expression data processing and modeling*, Proc Natl Acad Sci USA, 97(18), National Academy of Sciences, 2000.

[7] KH Pan, CJ Lih, and SN Cohen, *Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays*, Proc Natl Acad Sci USA, 102(25), National Academy of Sciences, 2005.

[8] T Breslin, P Edén, and M Krogh, *Comparing functional annotation analyses with Catmap*, BMC Bioinformatics, 15, BioMed Central, 2004.

[9] K Virtaneva, F A Wright, S M Tanner,B Yuan, W J Lemon, A Michael, Caligiuri, C D Bloomfield, A Chapelle, and R Krahe, *Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics, Proc Natl Acad Sci USA, 98(3), National Academy of Sciences, 2001, pp. 1124-1129.*

[10] D Nam, S B Kim, S K Kim, S Yang, S Y Kim, and I S Chu, ADGO: analysis of differentially expressed gene sets using composite GO annotation*, Bioinformatics, 22 (18), Oxford University Press, 2006, pp. 2249-2253.*

[11] Julie Ross, Dietary flavonoids and the MLL gene: A pathway to infant leukemia?*, Proc Natl Acad Sci USA, 97 (9), National Academy of Sciences, 2000, pp. 4411-4413.*

[12] Garbuglia, Iezzi, Capobianchi, Pignoloni, Pulsoni, Sourdis, Pescarmona, Vitolo, and Mandelli, Detection of TT virus in lymph node biopsies of B-cell lymphoma and Hodgkin's disease, and its association with EBV infection*, International Journal of Immunopathology Pharmacology (IJIPP), 16(2), Biolife, 2003, pp. 109-118.*

[13] M. Takeshita, H. Sakai, S. Okamura, Y. Oshiro, K. Higaki, O. Nakashima, N. Uike, I. Yamamoto, M. Kinjo, and F Matsubara, Splenic large B-cell lymphoma in patients with hepatitis C virus infection*, Human Pathology, W B Saunders CO-Elsevier Inc., 2005, 36(8), pp. 878-885.*

[14] Dino Veneri, Massimo Franchini, Roberta Zanotti, Francesco Frattinia, Federica Randon, Marianna Rinaldi, and Giovanni Pizzolo, Prevalence of hepatitis C virus infection among patients with lymphoproliferative disorders: A single center survey*, American Journal of Hematology*, 82(11), John Wiley and Sons, 2007, pp. 1031.

[15] Elaine Sarkin Jaffe,*Pathology and genetics of tumours of haematopoietic and lymphoid tissues*, in Intl Agency for Research on Cancer, 2003, pp. 171-172.

[16] Joseph Mazza, *Manual of Clinical Hematology*, Lippincott Williams & Wilkins, 2001, pp. 217.

[17] M J Laughlin, and H M Lazarus, *Allogeneic stem cell transplantation*, Humana Press, 2002, pp. 42.

[18] C P Tsakona, S E Kinsey,and A H Galdstone, *Use of Flow Cytochemistry via the H*1 in FAB Identification of Acute Leukaemias*, Acta Haematologica, 88(2-3, KARGER, 1992, pp. 72-77.

[19] S J Rodig, J S Abramson, G S Pinkus, S P Treon, D M Dorfman, H Y Dong, M A Shipp, and J L Kutok, *Heterogeneous CD52 Expression among Hematologic Neoplasms: Implications for the Use of Alemtuzumab (CAMPATH-1H)*, Clinical Cancer Research, 12(23), American Association for Cancer Research, 2006, pp. 7174-7179.

[20] T Miyashita, S Krajewski, M Krajewska, H G Wang, H K Lin, D A Liebermann, B Hoffman, and J C Reed, *Tumor suppressor p53 is a regulator of Bcl-2 and bax gene expression in vitro and in vivo*, Oncogene, 9(6), Nature Publishing Group, 1994, pp. 1799-1805.

[21] C Kirchoff, C Osterhoff, I Pera, and S Schroter, *Function of human epididymal proteins in sperm maturation*, Andrologia, 38(4-5), Allemagne, 1998, pp. 225-232.

[22] C Kirchhoff, R Carballada, B Harms, and I Kascheike, *CD52 mRNA is modulated by androgens and temperature in epididymal cell cultures*, Mol Reprod Dev, 56(1), Wiley-Liss, 2000, pp. 26-33.

[23] N M Almasri, J Al-Alami, and M Faza, *Bcl-2 gene rearrangement in Jordanian follicular and diffuse large B-cell lymphomas*, Saudi Med J, 26(2), Medical Services Department, 2005, pp. 251-255.

[24] B F Skinnider, D E Horsman, B Dupuisab, and R D Gascoyne, *Bcl-6 and Bcl-2 protein expression in diffuse large B-cell lymphoma and follicular lymphoma: Correlation with 3q27 and 18q21 chromosomal abnormalities*, Human Pathology, 30(7), Elsevier Science, 1999, pp. 803-808.

[25] M Murakoshi, R Y Osamura, and K Watanabe, *Immunocytochemical Localization of Glutathione-Peroxidase (GSH-PO) and Bcl-2 in the Rat Ventral Prostate*, Acta Histochem Cytochem, 36(4) (2003), pp. 335-343.

[26] Margaret, David, Lisa, Robin Roberts, Thomas Grogan, Thomas Miller, Larry Oberley, and Margaret Briehl, *A redox signature score identifies diffuse large B-cell lymphoma patients with a poor prognosis*, Blood, 106 (10), The American Society of Hematology, 2005, pp. 3594-3601.

[27] Elenitoba-Johnson, Jenson, Abbott, Palais, Bohling, Lin, Tripp, Shami, Wang, Coupland, Buckstein, Perez-Ordonez, Perkins, Dube, and Lim, *Involvement of multiple signaling pathways in follicular lymphoma transformation: p38-mitogen-activated protein kinase as a target for therapy*, Proc Natl Acad Sci USA, 100(12), National Academy of Sciences, 2003, pp. 7259-7264.

[28] *http://lymphoma.about.com/od/nonhodgkinlymphoma/p/dlbcelllymphoma.htm*

[29] *http://www.geneontology.org/*

[30] *http://www.cs.waikato.ac.nz/ml/weka/*

[31] *http://www.ailab.si/orange/datasets/DLBCL.htm*

[32] *http://www.ailab.si/orange/datasets/leukemia.htm*

[33] Blaise Hanczar, Mélanie Courtine, Arriel Benis, Corneliu Hennegar, Karine Clément, and Jean-Daniel Zucker, *Improving classification of microarray data using prototype-based feature selection*, SIGKDD Explor. Newsl., 5(2), ACM, New York, 2003, pp. 23-30.

[34] A R Poongothai, S Vishnupriya, and D Ragunadharao, *Quantitative variation of superoxide dismutase levels in leukaemias*, Indian Journal of Human Genetics, 10(1), Medknow Publications, 2004, pp. 9-12.

[35] Eric Bair, and Robert Tibshirani, *Machine learning methods applied to DNA microarray data can improve the diagnosis of cancer*, SIGKDD Explor. Newsl., 5(2), ACM, New York, 2003, pp. 48-55.

[36] Gregory Piatetsky-Shapiro, and Pablo Tamayo, *Microarray data mining: facing the challenges*, SIGKDD Explor. Newsl., 5(2), ACM, New York, 2003, pp. 1-5.

[37] R Tibshirani, T Hastie, M Eisen, D Ross, D Botstein, and P Brown, *Clustering methods for the analysis of DNA microarray data*, Technical Report, (1999), Department of Statistics, Stanford U.

[38] S Draghici, P Khatri, R P Martins, G C Ostermeier, S A Krawetz, *Global functional profiling of gene expression*, Genomics, 81(2), (2003), pp. 98-104.

[39] Hastie, Tibshirani, Eisen, Alizadeh, Levy, Staudt, Chan, Botstein, and Brown, *'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns.*, Genome Biol., 1(2), BioMed Central Ltd., 2000, pp. RESEARCH0003.

[40] V K Mootha, C M Lindgren, and K F Eriksson, *PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.*, Nat Genet., 34(3), BioMed Central Ltd., 2003, pp. 267-273.