Clustering Labeled Data and Cross-Validation for Classification with Few Positives in Yeast

Miles Trochesset and Anthony Bonner
University of Toronto
Department of Computer Science
10 King's College Rd.
Toronto, ON
M5S-3G4, Canada

{mtroches,bonner}@cs.toronto.edu

ABSTRACT

This paper presents two standard machine learning algorithms, one used in a non-standard way, for predicting the biological functions of essential genes in a systematic and comprehensive manner. We used gene expression and phenotype data from *Saccharomyces cerevisiae*. Determining gene function is simplified to a series of binary classification problems and one of the challenges of this learning task lies in the extremely small number of positives, compared with large amounts of negatives samples. We develop a method based on unsupervised hierarchical clustering used with labeled data to search for regions of high concentrations of positives and make predictions for the unlabeled genes. We also investigate the supervised logistic regression classifier as a baseline for comparing to our technique. Both of these methods are based on different views of the data and we found that depending on the biological processes being predicted, one or the other of these approaches performs better, although our method makes more confident predictions for more biological processes. The outcomes of the research are twofold: first we build a new biological data mining method based on existing machine learning tools that are readily accepted in the biological community. Second we make biological predictions of gene functions, each associated with a level of confidence and all above 50% precision.

1. INTRODUCTION

This paper investigates data mining and machine learning techniques for predicting, in a systematic and comprehensive manner, the possible functions of all putative and known genes (a gene may have several biological functions) in a yeast organism called *Saccharomyces cerevisiae*. We focused more intensely on making predictions for unlabeled genes, and decided to analyze the predictions of labeled genes in the future. Unlabeled genes are genes for which no function has yet been determined, whereas labeled genes are known to have at least one function. Systematic approaches for identifying the biological functions of genes, especially the unlabeled, are needed to ensure rapid progress from genome sequence to directed experimentation and applications (such as drug discovery).

The functions we learned are biological processes. Since relatively few genes are involved in a typical biological process, there are far more negatives than positives (as little as 0.01% of positives in the genome for certain biological processes), although some biological processes involve up to 60% of the genes in the genome. The learning task is made even harder by the fact that the samples we have comprise only about 10% of the genes in the genome (but required tremendous amounts of biological work to obtain nonetheless), 15% of which are unlabeled. So the number of positives available in our samples can be extremely small for some biological processes.

We examined two different methods based on two views of the data. The first view is that the positives and negatives can be separated by a hyperplane, which we fit using logistic regression. In the second view, the data constitutes as a sea of negatives with some small islands of positives of unknown size and number. We identify these concentrations of positives using hierarchical clustering on labeled data, which is not the standard unsupervised way of using this algorithm. We found that for some biological processes, one or the other method performs better, although our hierarchical method produces more confident predictions for more biological processes. Also, the method we develop allows the analysis of biological processes for which we have as little as 5 positive samples, unlike logistic regression which was unable to make predictions when the number of positives was below 20.

In this application, the cost associated with experimentally testing predictions lead us to performing leave-one-out cross-validation, not only to control how well the classifiers are behaving and draw ROC curves, but really to build decision rules for classifying samples. This is a main point in our methodology and we will explain it's details later.

Our analysis uses two types of data, gene expression from cDNA microarrays and growth phenotype data. Whole-genome expression profiling, facilitated by the development of DNA microarrays [12; 21], represents a major advance in genome-wide functional analysis. A single assay can measure the transcriptional response of thousands of genes, and often a full genome, to a change in cellular state such as disease, cell-cycle, cell division, response to stress and chemical compounds, or genetic perturbation and mutations. The scientific community agrees that gene expression alone cannot give a full picture of the cell state, because transcripts such as mRNAs need to be translated into proteins which sometimes need to be activated and each of these steps can be

regulated. Therefore more data types are needed to analyze regulation of the cell at a finer level of granularity. This is another reason why we chose to include sources of phenotype data in this study.
A lot of the classification work using machine learning has been done in cancer classification [1; 2; 9; 14; 15; 17; 19; 24] rather than predicting ontologies. This task is investigated in [5] but only for 6 classes (which were not defined by the Gene Ontology). Our approach is designed for making prediction for any of the classes in the Gene Ontology (on the order of a thousand different classes).

2. OVERVIEW OF THE DATA AND PREPROCESSING

The data used in this paper was gathered at Hughes Lab at the Banting and Best Department of Medical Research in the University of Toronto. In order to investigate the function of essential genes, which are required for survival and therefore cannot be knocked out, Hughes lab constructed a particular type of mutant yeast strains for two thirds of all the essential genes [16]. Construction was suspended because of project deadlines and financial reasons. These 602 mutants allow direct experimentation on the essential genes. There is a one-to-one correspondence between an essential gene and a mutant strain. The following datasets were collected and used for predicting gene function :

- *gene expression* from DNA microarrays measuring the abundance of gene transcripts of the mutant cells relative to the wild-type strain for the entire genome. The 291 samples, corresponding to 218 essential genes with replicates (out of the 602 constructed mutants), are publicly available on NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) through accession number GPL1229. After quantification, hybridized samples were normalized using background subtraction, followed by a LOWESS smoother to correct for dye discrepancies and by a high pass filter to remove any sorts of spatial artifacts (scratches, dust, gradient across the array or red corners ...) After investigating several techniques for imputing missing values, we used BPCAfill [18], which performed the best (using normalized root mean squared error as the measure of goodness of fit) on simulated datasets with the same proportion of missing data (approximately 13%). In the end the dimensionality of the data was reduced from 6307 genes on the arrays to 20 using principal components analysis (PCA) [11; 20] by selecting the eigenvectors associated with the 20 largest eigenvalues of the covariance matrix.

- *size distribution* measures the distribution of cell sizes for 591 of the 602 mutant strains. Normalization procedure: strains were grown by batches and this dataset was normalized so as to make the median of the median distribution of strains grown on the same batch to coincide for all batches. Validation of this normalization was done by verifying that the distribution of control wild-type strains grown in all batches coincided. The distributions were measured at 256 points, and the dimensionality was reduced to 8 by PCA. All growth phenotype datasets are available at http://hugheslab.med.utoronto.ca/Mnaimneh.

- *Drug Response* looks at the sensitivity of the mutant strains to different chemical compounds in 27 experimental conditions. 685 mutant strains, corresponding to 585 mutant strains with replicates, were grown on plates with one drug and the size of the colonies were compared to wild-type grown with the same drug. The value reported in the dataset was the log P-value that a difference existed between the two groups.

- *Morphology* represents the morphological features of the mutant cells which were visually inspected for 17 different characteristics such as elongated, budded or pointed cells. This data is the only type which is categorical. A 1 indicates that the feature was slightly observed for all mutant cells, a 0 indicates it was not. On rare occasion other types appear, 0.5 means the feature was slightly observed but the phenotype was not penetrant, 2 moderately observed for all cells, 2.5 moderately observed but the phenotype was not penetrant, 3 severely observed for all cells.

Each dataset covers a different set of the 602 constructed mutants , although these sets intersect, and the number of positive samples for a particular biological process depends on the dataset being used. A simple solution was to use these datasets independently.
Finally the gene labels we learned, which are organized in a hierarchical manner according to the Gene Ontology (GO) [23], were downloaded from the (SGD) Saccharomyces Genome Database [6; 7]. We used the *biological process* type of the GO database as our labels for gene function because the biologists we work with were interested in these rather than *molecular function* or *cellular component*. Almost 40% of the genes in the genome have no label for all of the the biological processes, we call these genes unlabeled. 15% of the 602 constructed mutants were uncategorized. Some GO biological processes are so broad and general that they involve thousands of genes, such as *protein metabolism [GO:0019538]* or *cell organization and biogenesis [GO:0016043]*. In fact, large top-level (high in the GO hierarchy) categories involving hundreds of genes are often not specific enough to verify experimentally. Therefore we have restricted this study by not showing biological processes that clearly involved too many genes to be interesting.

3. CLASSIFICATION BY HYPERPLANE

In this section we examine the case where the two classes are separable by a hyperplane. This is a strict assumption about the data, but it leads to predictions with high level of confidence for some biological processes nonetheless and represents a baseline for comparing the results obtained with the second view which we describe in the next section. We choose to fit the hyperplane using logistic regression [11] because of it's simplicity and also because it is well understood, and accepted in the biology community [3]. In 3.1 we investigate a method by which we can easily build decision rules customized to a particular biological process for classifying samples, precision being the only user-defined parameter. We apply these decision rules to the unlabeled samples in 3.2.
Each gene can be involved in several biological processes and therefore this is not the classical machine learning approach in which samples can belong to one class only, and of course

several genes can be involved in a biological process. We learned biological processes independently, which simplified the problem to discriminating between two classes for each biological process: either a gene is involved or it is not.

3.1 Cross Validation For Customized Decision Rules

We trained logistic regression classifiers by leave-one-out cross-validation on the labeled samples of each of the biological processes we chose to learn. Each time we computed the posteriors $P(Y = 1|X = x)$ where x was the sample set aside, Y denotes the class label (which takes the value 1 if a gene is involved in the biological process, and 0 otherwise). We had little choice but to use leave-one-out cross validation because, having so few positives in our samples (as little as 5 positives), we could not afford to waste labeled data by separating it into training and test sets.

In order to classify a sample we need to build a decision rule. One very simple rule could be to classify as positive any sample for which the posterior probability is above 0.5. Here we are faced with a decision making problem which needs a little more attention because of the cost associated with making false predictions. In molecular biology, running experiments is very expensive and we want to be very confident about the prediction being true before testing it in wet lab. All the cost of decisions is biased toward false predicted positives in this application and false negatives aren't given as much importance. As a result to increase our confidence on the predicted positives, we computed conservative thresholds for discriminating between classes, each depending on the particular biological process. A sample will be classified as positive if it's posterior is above that threshold $P(Y = 1|X = x) > t$. In the logistic regression setting, the classes are separated by the hyperplane defined by the equation $\theta^T x = 0$. When the input x is on the hyperplane,

$$P(Y = 1|X = x) = P(Y = 0|X = x) = 0.5 \qquad (1)$$

Raising the threshold, corresponds to translating that hyperplane in the direction of θ (or $-\theta$). Our procedure consists of translating the hyperplane toward the positive samples until the ratio of true positives to false positives is sufficiently high. Therefore we use cross validation, not only to control how well the classifiers are performing, but really to build decision rules for classifying the unlabeled samples.

The measure of satisfaction we used for translating the hyperplane is precision, which is the ratio of true positives to predicted positives, i.e.

$$\text{precision} = \frac{\text{true positives}}{\text{predicted positives}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (2)$$

Predicted negatives cannot be confirmed experimentally (at least at Hughes Lab which is providing us with the data), so knowledge is gained only when predicted positives are confirmed and it is indeed precision biologists are interested in and not overall classification performance.

For a particular biological process, one approach could be to choose the threshold that leads to the maximum precision computed using all labeled samples, but we prefer to take a more conservative approach by setting a user-defined precision. That way predictions will only be made for biological processes for which the classifier reaches that precision at some threshold. For biological processes for which logistic regression performed poorly, no predictions will be made. Because precision is not a monotonic function in t, we chose the lowest threshold leading to the desired precision since this solution maximizes the recall (also known as sensitivity in the signal processing and biological worlds), which is the percentage of positives which are predicted positives:

$$\text{recall} = \frac{\text{true positives}}{\text{all positives}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (3)$$

We computed five thresholds for each GO biological process, corresponding to precision levels of 100%, 85%, 75%, 60% and 50% based on the labeled data. The precision level used to classify a sample, along with the distance of that sample to the translated hyperplane leads to different of confidence levels.

3.2 Predicting functions for the unlabeled genes

For classifying the unlabeled samples, we trained a logistic regression classifier per biological process using all the available labeled samples and then computed the posterior probabilities $P(Y = 1|X = x)$ where x were the unlabeled samples. Unlabeled samples were classified as positive whenever their posterior was greater than the threshold, and predicted positives were reported.

Predictions were grouped by the precision level used and by biological process and are separated into batches depending on which dataset was used. Each prediction has four fields: a GO biological process, a systematic gene name, the precision level used for computing the threshold and finally the difference between the gene's posterior probability and the threshold which characterizes the distance from the translated hyperplane. All this data was assembled in tab delimited files available as supplementary data.

For increasing the significance of the precisions computed, we forced them to be based on a minimum of 10 predicted positives. We call *confident prediction* one that satisfies that constraint. We only reported confident predictions based on thresholds corresponding to 50% precision and above, this means that we can never make predictions for biological processes involving fewer than 5 genes.

It is worth underlying the fact that precision levels reported are minimums. A sample being predicted positive at a precision level could also have been predicted positive at a higher precision level. Summaries of these predictions are shown in Table 1-3. In these tables we report the number of unlabeled genes predicted grouped by biological process and by precision level. We indicate the number of known genes involved in each biological process as well as the number of positive samples available in the dataset used.

We observed that the procedure of fitting a hyperplane using logistic regression converged only for biological processes having more than 10 positive in our samples. In fact we observed that no confident predictions were made for biological processes involving fewer than 20 positives in our samples. The method we develop in the next section does not have this limitation.

4. HIERARCHICAL CLUSTERING ON LABELED DATA

In this section we investigate a method based on a different view of the data. We consider here that positive samples represent small islands among a sea of negatives, but we don't know how many islands there are nor their size. One

Number of predictions for GO-BP:	known in GOBP	# pos in samples	precision .6	precision .5
transcription [GO:0006350]	534	39		6
transcription, DNA-dependent [GO:0006351]	505	39		6
cell proliferation [GO:0008283]	571	37	5	8
RNA metabolism [GO:0016070]	336	34		10
cell cycle [GO:0007049]	494	33	4	6
RNA processing [GO:0006396]	297	33		4
biosynthesis [GO:0009058]	803	30		5
mitotic cell cycle [GO:0000278]	288	30		5
ribosome biogenesis and assembly [GO:0042254]	186	26		18
ribosome biogenesis [GO:0007046]	151	24		17
macromolecule biosynthesis [GO:0009059]	449	21	1	1
protein biosynthesis [GO:0006412]	442	21	1	1
DNA replication and chromosome cycle [GO:0000067]	219	20		1
transcription from Pol I promoter [GO:0006360]	149	20	7	8

Table 1: Summary of confident predictions made by logistic regression on the gene expression data

Number of predictions for GO-BP:	known in GOBP	# pos in samples	precision .6	precision .5
transcription [GO:0006350]	534	142		4
transcription, DNA-dependent [GO:0006351]	505	141		4
RNA metabolism [GO:0016070]	336	128	4	15
RNA processing [GO:0006396]	297	127		17
cell proliferation [GO:0008283]	571	116		5
ribosome biogenesis and assembly [GO:0042254]	186	85	3	15
ribosome biogenesis [GO:0007046]	151	79	3	15
transcription from Pol I promoter [GO:0006360]	149	76		14
rRNA processing [GO:0006364]	121	65		12
organelle organization and biogenesis [GO:0006996]	550	61		2

Table 2: Summary of confident predictions made by logistic regression on the cell size distributions

possibility would be to use k-nearest neighbors (kNN), but unfortunately we have no idea what to expect for k, and a simple majority vote would not work because of the high number of negatives almost everywhere (including regions of relatively high concentrations of positives). We develop an algorithm based on hierarchical clustering that circumvents these problems.

Clustering has been used extensively in functional genomics to analyze gene expression data [2; 4; 8; 13; 22] and is probably what biologists use and trust most. Biologists often use hierarchical clustering on gene expression data. For example, they usually display the resulting dendrogram immediately beside the gene expression data from which it was derived, and label the leaves of the dendrogram with gene names and/or biological processes. The method we develop here is based on this methodology, but extends it to an automated process. It also has the advantage of using all of the known functions of the genes in the hierarchical tree and not just their main function.

Our method looks for regions in the data space of high concentrations of positives. All that is required is some notion of "distance" between all pairs of elements. In contrast, logistic regression does not work for the morphology dataset because, although the data is technically real valued, it is still too categorical for the fit to converge.

4.1 Details of the Procedure

We first build a hierarchical tree on all available labeled and unlabeled samples using hierarchical agglomerative clustering [10] with average linkage. In constructing the tree, we ignore the labels on the data. In this way, we can include both labeled and unlabeled data in the tree, and more importantly, we can use the same tree for each biological process, thus saving on computing time, since the tree need only be built once. Thus, the construction of the tree can be viewed as a preprocessing step whose cost is amortized over all the biological processes. However, after the tree is constructed, it is not possible to add new unlabeled samples to the data.

We used the correlation coefficient between two samples as a measure of the distance between them rather than Euclidean distance. This is because the actual level of expression of two genes is less important than their profiles being correlated among a set of experiments. For example, the measured expression of a gene might be twice that of another gene in the same pathway because of experimental factors such as oligonucleotide probe quality (folding into a stable secondary structure, melting temperature etc).

Following the construction of the tree, we use it to build a classifier for each biological process. Recall that each such process provides a different set of labels for genes. Since the leaves of our tree represent genes, each leaf is assigned the label of the gene it represents. Leaves for unlabeled genes are labeled as negative, since it is likely that an unlabeled gene is not involved in any particular biological process. (We also flag such leaves, so as to remember that they are unlabeled). We can now look in the tree for regions of high concentrations of positive leaves, after which we assign labels to all the unlabeled genes that fall in such regions. These assignments represent our classifiers predictions.

To make these assignments, the algorithm computes a score σ for each internal node in the tree, reflecting the concentration of positives at the leaves under the node.

$$\sigma = \frac{\#\text{ of positives at leaves}}{\#\text{ of leaves}} \times \left(1 - \alpha e^{-\#\text{ of positives}}\right) \quad (4)$$

The first factor in this formula is the proportion of positive leaves under the node, it reflects the concentration of positives in the region of the data space in which the leaves are. The second factor tends to one when the number of positives raises, and tends to zero as the number of positives

Number of predictions for GO-BP:	known in GOBP	# pos in samples	precision .6	precision .5
RNA metabolism [GO:0016070]	336	140		1
RNA processing [GO:0006396]	297	139		1
cell proliferation [GO:0008283]	571	127	3	5
ribosome biogenesis and assembly [GO:0042254]	186	93		5
ribosome biogenesis [GO:0007046]	151	86		6
rRNA processing [GO:0006364]	121	71		3

Table 3: Summary of confident predictions made by logistic regression on the drugs dataset

```
Build hierarchical tree on all labeled and
unlabeled samples.
For each GO biological process GO-BPi do {
  Label the leaves according to GO-BPi.
  Label unlabeled samples as negatives.
  For each sample Sj do {
    Relabel Sj as negative.
    Compute the score of all internal nodes.
    Compute the score of Sj as maximum score of
    all it's ancestors.
  }
  Find lowest threshold that achieves
  user-specified precision.
  Classify unlabeled samples using this threshold.
  Report predicted positives.
}
```

Figure 1: Algorithm Pseudo-code

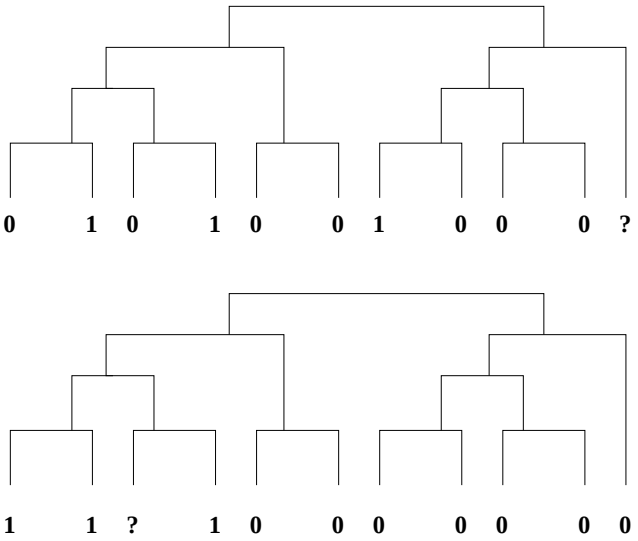

Figure 2: Toy hierarchical tree reused with labels from two biological processes

decreases. It gives more importance (higher score) to nodes with more positive leaves, *i.e.*, to larger regions of positive concentration, since we regard such regions to be more statistically significant. We have used $\alpha = 0.5$ and haven't investigated tweaking this parameter nor using other functions for the second factor of this equation. We then define the score of a leaf to be the maximum score of all it's ancestors (internal nodes). Since unlabeled samples are leaves in the tree, they automatically receive a score, which we use to classify them.

Before building decision rules, we use a technique similar to the cross validation of the previous section. At each iteration, we effectively remove a labeled sample by treating it as unlabeled. The scoring process described above is repeated each time. This provides a score for the labeled sample being treated as unlabeled. Each labeled leaf is scored in this way.

It is now easy to build a decision rule. We simply set a threshold, and a leaf is classified as positive if its score is above the threshold. To evaluate the rule, we apply it to labeled leaves, and compare each leaf's true label to its predicted label. A threshold that achieves a user-specified precision is then chosen. Finally, using this threshold, we use the decision rule to classify all the unlabeled data.

The pseudo-code for this algorithm is given in Figure 1. A toy example of how the tree is reused for each biological process is given in Figure 2. Our method is very fast, the whole process from building the tree to reporting predicted positives in all biological processes took a few seconds for each dataset on a Pentium IV 2GHz. This should be contrasted with the logistic regression methodology which required approximately a half hour for each dataset.

4.2 Results

Predicted positives were reported for all four datasets and assembled in tab delimited files. A prediction has four fields: a GO biological process, the gene systematic name, the difference between the score of the unlabeled leaf and the threshold used, and the precision corresponding to that threshold. The precision and the difference between the score and the threshold represent the confidence we have in the prediction. Summaries of these predictions (except for the morphology dataset) are shown in Table 4-6. We did not show the summaries for the morphology dataset, the number of confident predictions made were approximately the same as in Tables 5 and 6.

Comparing the two methods for identical datasets (Table 1 vs. 4, Table 2 vs. 5 and Table 3 vs. 6), we observe that our method produces many more confident predictions, at precision levels 50% and 60% (even 75% with the drugs dataset), and for more biological processes. In particular, our hierarchical method made prediction for 18 biological processes involving fewer than 20 positive in the samples whereas logistic regression produced none.

In Figure 3 we show the ROC curves of a couple of the classifiers used for making predictions, obtained by the method we developed. We clearly see that our method performs better than guessing the majority class, *i.e.* classify as negative every time, and achieves very high true positive rates at thresholds for which the false positive rates are still very low. For example, the classifier used for predicting genes involved in *glycerophospholipid biosynthesis* reaches a true

Number of predictions for GO-BP:	known in GOBP	# pos in samples	precision .6	precision .5
transcription [GO:0006350]	534	39		18
transcription, DNA-dependent [GO:0006351]	505	39		18
RNA metabolism [GO:0016070]	336	34	9	11
RNA processing [GO:0006396]	297	33	11	11
ribosome biogenesis and assembly [GO:0042254]	186	26	19	21
ribosome biogenesis [GO:0007046]	151	24	12	19
protein modification [GO:0006464]	361	23		3
organelle organization and biogenesis [GO:0006996]	550	22		1
macromolecule biosynthesis [GO:0009059]	449	21		9
protein biosynthesis [GO:0006412]	442	21		9
transcription from Pol I promoter [GO:0006360]	149	20	11	11
rRNA processing [GO:0006364]	121	18	8	8
catabolism [GO:0009056]	276	16		2
cytoskeleton organization and biogenesis [GO:0007010]	255	14		2
mRNA processing [GO:0006397]	124	14		4
macromolecule catabolism [GO:0009057]	176	12		1
lipid metabolism [GO:0006629]	190	11		1
lipid biosynthesis [GO:0008610]	111	11		1
RNA splicing [GO:0008380]	112	10		4
mRNA splicing [GO:0006371]	92	10		4
microtubule-based process [GO:0007017]	94	8		1
microtubule cytoskeleton organization and biogenesis [GO:0000226]	86	8		1
M-phase specific microtubule process [GO:0000072]	62	8		1
membrane lipid metabolism [GO:0006643]	85	6		1
membrane lipid biosynthesis [GO:0046467]	62	6		1
phospholipid metabolism [GO:0006644]	64	5		1
phospholipid biosynthesis [GO:0008654]	48	5		1
glycerophospholipid metabolism [GO:0006650]	34	5		1
glycerophospholipid biosynthesis [GO:0046474]	30	5		1

Table 4: Summary of confident predictions made by our clustering method on the gene expression data

positive rate of 100% for less than 2% false positive rate.

5. CONCLUSION & FUTURE WORK

We developed a method based on hierarchical clustering for labeled data to find regions in the data space of relatively high concentration of positives. This technique allows the analysis of biological processes involving very few genes. With this method, we were able to make confident predictions at precisions of 50% and above for biological processes for which our samples contained as few as 5 positives. The methodology developed here is not restricted to learning essential genes, but could be applied to any set of genes.

We used correlation as a measure of similarity between pairs of elements and average linkage to build the hierarchical tree. It would be interesting to investigate different distance metrics and especially other linkage strategies such as single linkage, which produces clusters that aren't necessarily compact.

We focused on making predictions for unlabeled genes. However, it would be biologically interesting to report cases in which a gene's true label is negative but whose predicted label is a confident positive. This is because negative labels in our dataset are sometimes wrong. A more challenging task would be to use datasets concurrently for the intersecting samples and independently for disjoint sets of samples. Also finding methods for learning biological processes concurrently rather than independently is one of our future goals. We are thinking of using the Gene Ontology hierarchy to propagate up the hierarchy predictions made lower down, because if a gene is involved in a biological process, it is also involved processes above it in the hierarchy. This isn't completely trivial because the hierarchy is not a tree and a process can have several parents. More interestingly, if a prediction is made in a biological process having children, we would like to find methods for making the prediction more specific by propagating it down the hierarchy as far as possible.

Acknowledgments

We wish to thank Hughes Lab for providing us with their data.

6. REFERENCES

[1] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, February 2000.

[2] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Science*, 96(12):6745–6750, June 1999.

[3] J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant. Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, 22(1), January 2004.

[4] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal Of Computational Biology*, 6:281–297, 1999.

[5] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceeding*

Number of predictions for GO-BP:	known in GOBP	# pos in samples	precision .5
transcription [GO:0006350]	534	142	5
transcription, DNA-dependent [GO:0006351]	505	141	5
cell proliferation [GO:0008283]	571	116	1
cell cycle [GO:0007049]	494	99	1
ribosome biogenesis and assembly [GO:0042254]	186	85	5
ribosome biogenesis [GO:0007046]	151	79	2
transcription from Pol I promoter [GO:0006360]	149	76	3

Table 5: Summary of confident predictions made by our clustering method on the cell size distributions

Number of predictions for GO-BP:	known in GOBP	# pos in samples	precision .75	precision .6	precision .5
transcription [GO:0006350]	534	159	2	5	5
transcription, DNA-dependent [GO:0006351]	505	158	2	5	5
RNA metabolism [GO:0016070]	336	140			27
RNA processing [GO:0006396]	297	139			17
DNA metabolism [GO:0006259]	379	68			6
mRNA processing [GO:0006397]	124	60			4
nuclear organization and biogenesis [GO:0006997]	213	42		3	3
chromosome organization and biogenesis (sensu Eukarya) [GO:0007001]	178	35		3	3
establishment and/or maintenance of chromatin architecture [GO:0006325]	155	32		3	3
DNA packaging [GO:0006323]	155	32		3	3

Table 6: Summary of confident predictions made by our clustering method on the drugs dataset

of the National Academy of Science, 97(1):262–7, January 2000.

[6] K. R. Christie, S. Weng, R. Balakrishnan, M. C. Costanzo, K. Dolinski, s S Dwight, S. R. Engel, B. Feierbach, D. G. Fisk, J. E. Hirschman, E. L. Hong, L. Issel-Tarver, R. Nash, A. Sethuraman, B. Starr, C. L. Theesfel, R. Andrada, G. Binkley, Q. Dong, C. Lane, M. Schroeder, D. Botstein, and J. M. Cherry. Saccharomyces genome database (sgd) provides tools to identify and analyze sequences from saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Research*, 32:D311–D314, 2004.

[7] S. S. Dwight, M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, G. Sherlock, A. Sethuraman, S. Weng, D. Botstein, and J. M. Cherry. Saccharomyces genome database (sgd) provides secondary gene annotation using the gene ontology (go). *Nucleic Acids Research*, 30(1):69–72, 2002.

[8] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science*, 95(25):14863–14868, December 1998.

[9] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–14, October 2000.

[10] D. Hand, H. Mannila, and P. Smyth. *Principles of Datamining*. MIT Press, 2001.

[11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[12] T. R. Hughes, M. Mao, A. R. Jones, J. Burchard, M. J. Marton, K. W. Shannon, S. M. Lefkowitz, M. Ziman, J. M. Schelter, M. R. Meyer, S. Kobayashi, C. Davis, H. Dai, Y. D. He, S. B. Stephaniants, G. Cavet, W. L. Walker, A. Westand, E. Coffey, D. D. Shoemaker, R. Stoughton, A. P. Blanchard, S. H. Friend, and P. S. Linsley. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology*, 19(4):342–347, April 2001.

[13] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, July 2000.

[14] J. Khan, J. S. Wei, M. Ringner, et. al, and P. S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6):673–679, 2001.

[15] Y. Lee and C. K. Lee. Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, 19(9):1132–9, June 2003.

[16] S. Mnaimneh, A. P. Davierwala, J. Haynes, J. Moffat, W.-T. Peng, W. Zhang, X. Yang1, J. Pootoolal, G. Chua, A. Lopez, M. Trochesset, D. Morse, N. J. Krogan, S. L. Hiley, Z. Li, Q. Morris, J. Grigul, N. Mitsakakis, C. J. Roberts, J. F. Greenblatt, C. Boone, C. A. Kaiser, B. J. Andrews, and T. R. Hughes. Exploration of essential gene functions via titratable promoter alleles. *Cell*, July 2004.

[17] D. V. Nguyen and D. M. Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50, Jan 2002.

[18] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubra, and S. Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 2003.

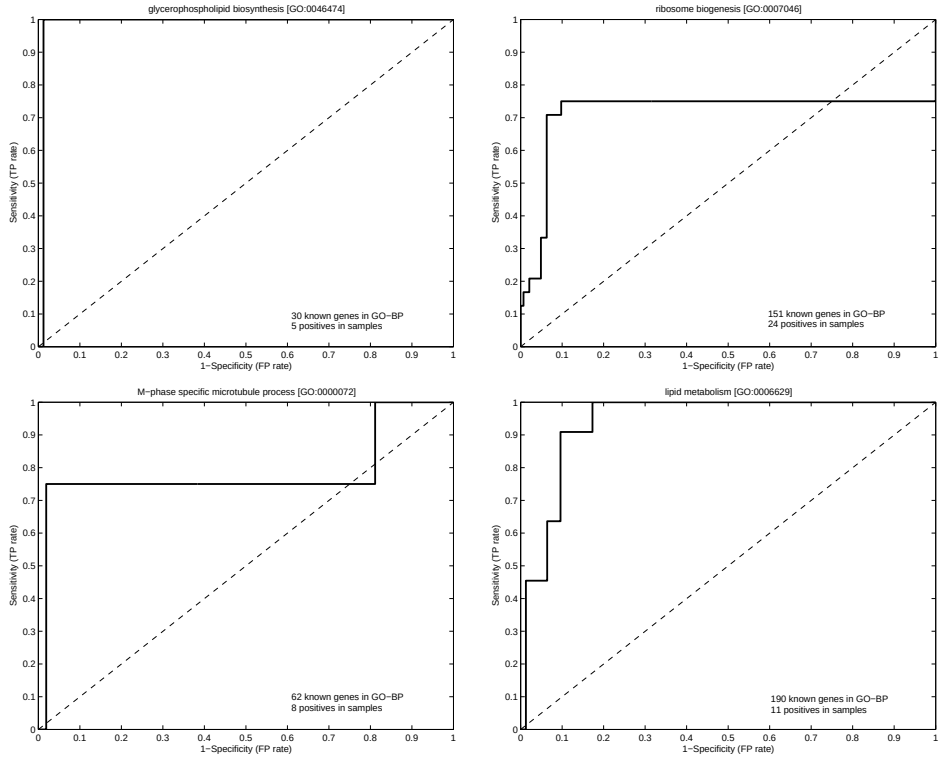

Figure 3: ROC curves for some of the classifiers we used for making predictions

[19] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of National Academy of Science*, 98(26):15149–54, December 2001.

[20] S. Raychaudhuri, J. M. Stuart, and R. B. Altman. Principal components analysis to summarize microarray experiments: Application to sporulation time series. *Proceedings of the Pacific Symposium on Biocomputing*, 5:465–466, 2000.

[21] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, October 1995.

[22] R. Sharan and R. Shamir. Click: A clustering algorithm for gene expression analysis. In *ISMB*, 2000.

[23] Gene Ontology Consortium. The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32:D258–D261, 2004.

[24] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Science*, 98(20):11462–11467, September 2001.