A Combinatorial Approach to the Analysis of Differential Gene Expression Data: The Use of Graph Algorithms for Disease Prediction and Screening *

(Extended Abstract)

Langston, M. A.[1], Lin, L.[1], Peng, X.[2], Baldwin, N. E.[1], Symons, C. T.[1] and Zhang, B.[3]

ABSTRACT

Combinatorial methods are studied in an effort to gauge their potential utility in the analysis of differential gene expression data. Patient and gene relationships are modeled using edge-weighted graphs. Two somewhat orthogonal algorithms are devised and implemented. One is based on finding optimal cliques within general graphs, the other on isolating near-optimal dominating sets within bipartite graphs. A main goal is to develop methodologies for training algorithms such as these on patient populations with known disease profiles, so that they can then be employed to classify and predict the likelihood of disease in patient populations whose profiles are not known in advance. These novel strategies are in marked contrast with Bayesian and other well-known techniques. Encouraging results are reported.

Categories and Subject Descriptors

F.2.2 [Nonnumerical Algorithms and Problems]: *systems and software*; G.2.2 [Graph Theory]: *graph algorithms*; J.3 [Life and Medical Sciences]: *biology and genetics*

General Terms

Algorithms, Performance, Theory

Keywords

Combinatorial Methods, Discrete Mathematics, Disease Prediction and Screening, Graph Algorithms, Microarray Analysis

1. INTRODUCTION

A fundamental problem in cancer treatment is early and reliable detection. The identification of a set of genes whose expression levels serve as an accurate discriminator between normal and cancerous patients would not only represent significant progress towards developing more reliable cancer diagnosis protocols, but might also identify novel therapeutic targets. With this motivation in mind, we investigate the hypothesis that only a modest number of genes may suffice for this task. We seek to develop algorithms and software for this purpose, and introduce a graph theoretical method of differential gene expression analysis. The goals of this method are to identify a set of genes useful in discriminating normal from cancerous tissues, and to use these genes in disease prediction and screening. One of the features of our algorithm is the computation of a discrimination score for each gene represented in an input microarray. This score estimates a gene's relative ability to distinguish between normal and cancerous patient tissues. We then select the highest-scoring genes, and use them to calculate a pairwise similarity metric between patients' expression profiles. With this information, we construct a complete weighted graph, in which the vertices represent patients and the edges are weighted by the similarity metric between patient vertices. A high-pass filter and a user-defined threshold are next used to transform the complete weighted graph into an incomplete unweighted graph. After an iterative application of clique enumeration, we incorporate a dominating set algorithm to highlight those genes that seem to discriminate among the greatest number of patients. The combination of these tools appears to produce some very encouraging predictive results.

In the sequel, we describe the datasets we have chosen to study, the algorithms we have devised, and the results we have obtained. We also draw some conclusions for this effort.

2. DATA EMPLOYED

The Harvard Lung Cancer Dataset [2] and the Michigan Lung Cancer Dataset [3] were selected for the study. Cel files were downloaded from the CAMDA website. The Michigan dataset

---

[1] Department of Computer Science, University of Tennessee, Knoxville, TN 37996-3450.

[2] Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN 37996-0845.

[3] Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6124.

was used for algorithmic training. This includes 86 cancer patient samples and 10 samples from individuals categorized as normal. The Harvard dataset was used for testing. This includes 190 patient samples associated with adeno carcinoma, 21 with squamos cell carcinoma, 20 with pulmonary carcinoids, 6 with small cell carcinoma and 17 samples from individuals categorized as normal. In a second Probe level, data in the cel files from each dataset were converted to background, adjusted, normalized, and log transformed. Expression measures for each probe are set using robust multi-array average (RMA) [11] in bioconductor.

3. A CLIQUE-BASED STRATEGY

Clique is a well-known *NP*-complete problem, and is typically formulated as in [8]:

Input: A graph $G=(V,E)$ and a positive integer $k \leq |V|$.

Question: Is there a subset $V' \subseteq V$ for which $|V'| \geq k$ and such

that every pair of vertices in V' is joined by an edge in E.

Clique is rapidly becoming recognized for its relevance in a variety of bioinformatics applications. In [1], for example, fast parallel algorithms have been devised and applied to extremely large microarray datasets in an effort to help identify putatively co-regulated genes in murine neural regulatory networks.

Our approach relies on having at hand at least two datasets, one for training and one for testing. For this we employ the Harvard and Michigan datasets, mainly because they are both the results of experiments using Affymetrix microarrays. For training, we use the Michigan dataset, because it represents only one type of cancer (adeno carcinoma).

Our goal in training is to develop graph-theoretic tools to help distinguish normal from cancerous patient samples. Ideally, we hope to be able to construct an unweighted graph in which edges connect only members of the same group. At that point, clique analysis should be an attractive approach for testing our methods against additional data (in this particular case, the Harvard dataset).

Our first step is to determine which genes appear to discriminate best between sample types. To accomplish this, each gene is assigned a score, and only the best genes are retained for use in subsequent steps. The distributions of the expression values of these genes should typically be expected to be bimodal. Thus the difference of the group medians give us a general measure of the difference of expression between the two groups. When we subtract the sum of the standard deviations of a gene within each group, we intend to eliminate or at least diminish the importance of any genes whose expression levels vary excessively.

The data is obtained as in Section 2 as an *n x m* matrix, *A*, of expression values, in which a row represents a test subject and a column denotes a gene. Our algorithm can be described in pidgin ALGOL, as follows.

procedure gene-score-and-select

for *i=1* to *n*

   normalize expression values in row *i* to the range [0, 1]

for *j=1* to *m*

   compute median expression value (*m*) and standard deviation (σ) on normal subject data for gene *j*

   repeat computation on cancer patient data for gene *j*

   set score(gene *j*) = |*m*(normal) – *m*(cancer)| - |σ(normal) + σ(cancer)|

delete genes with scores not exceeding zero

return remaining genes and their scores

This procedure delivers a collection of 118 genes for further evaluation. They can be classified by biological process as shown in the Figure 1 below.
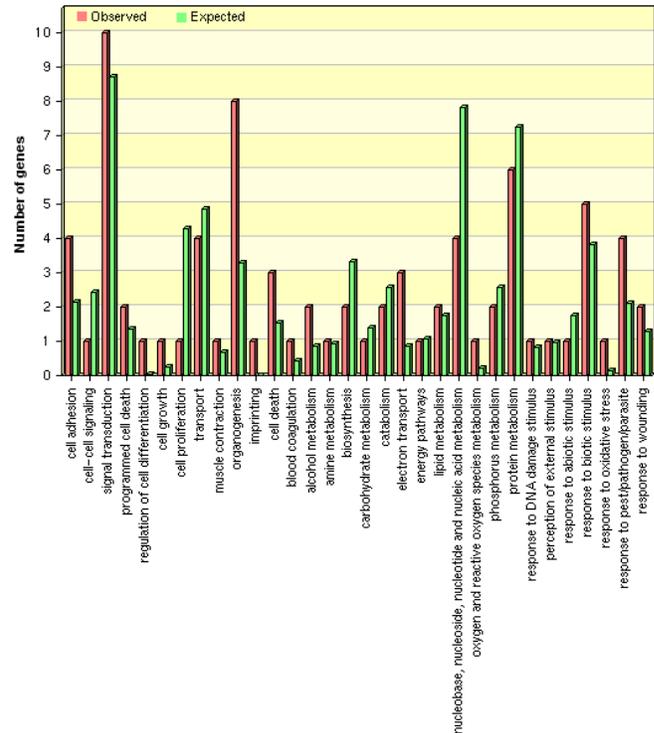

Figure 1. Bar chart of level 4 categories under biological-process

An assignment of inter-patient weights can help demonstrate the degree to which these genes and their respective scores help delineate the normal from the cancerous patients. Here the weight between patients *i* and *j* is intended to represent the degree of similarity in their respective expression levels. We compute this weight as a sum over all genes selected in the previous step, because it is these genes that seem to have the greatest potential to serve as good discriminators. Accordingly, we set *weight(i,j)* to:

$$\sum score(gene_k) \bullet (1 - \left| expression\_value_{ik} - expression\_value_{jk} \right|)$$

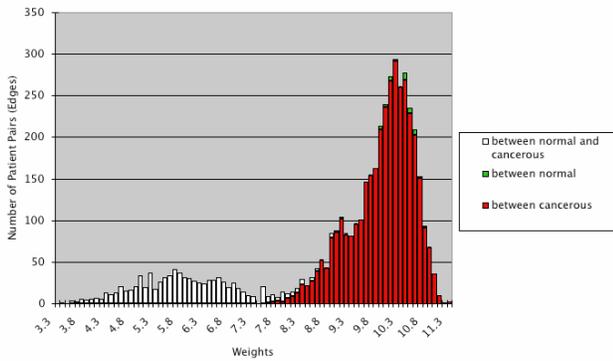The following frequency histogram is illustrative.

Figure 2. Weights between patient pairs using 118 genes (Michigan data)

As is shown, patients in the higher-weighted pairs tend to be either both normal (depicted in blue) or both cancerous (shown in red). Conversely, the lower-weighted pairs tend to be non-discriminatory, reflecting pairs in which one patient is normal and the other is cancerous. This seems to confirm our gene scoring and selecting procedure. It also suggests an initial threshold weight at which to delete edges in our next step (to be described shortly). Call this threshold *T*. We choose as a somewhat informed but still rather arbitrary starting value *T*=8.3. We use our restricted set of genes to build an edge-weighted graph. In this graph, patients are represented by vertices and the weight of an edge between a pair of patients is set using the simple summation formula already described. Any edge whose weight is less than *T* is removed. The resulting unweighted graph is then searched for all maximal cliques. Our aim is to train our codes so that we can find appropriately-sized cliques to cover both normal and cancerous groups, while minimizing cliques that overlap these groups. This requires iteration, as detailed below.

```
procedure clique-analysis
initialize edge-weighted graph of order n
  for i=1 to n
    for j = 1 to n
      set the weight of each edge
for a user-specified number of iterations do
  use T to delete edges with low weight
  find in resulting undirected graph all maximal cliques, C
  analyze C to refine the choice of T
return T
```

4. PRELIMINARY ANALYSIS

Because we know which samples are normal and which are cancerous in the Michigan dataset, we are able to iterate our method until we have a reasonable set of covering cliques. The optimal threshold seems to be centered at around *T*=9.0. We are not completely satisfied, however, with the lingering presence of overlapping cliques. Additional experimentation with gene cutoff

scores seems to indicate that the presence of genes with low scores are problematic. But neither raising the cutoff score nor additional modification of the threshold are of much use. What seems missing in our estimates of gene discrimination is a way to determine which genes impact the greatest number of patients. For this, we turn to another graph metric, dominating set.

5. REFINEMENT VIA DOMINATING SET

Dominating Set, another well-known *NP*-complete problem, can be stated as follows.

Input: A graph $G=(V,E)$ and a positive integer $k \leq |V|$.

Question: Is there a subset $V' \subseteq V$ for which $|V'| \leq k$ and every

vertex $v \in V - V'$ is joined to a vertex in V' by an edge in E.

In some sense dominating set is more difficult than clique. This is because clique can be solved by using graph complementation and vertex cover, which is fixed-parameter tractable [6]. The same cannot be said for dominating set. Thus we will only approximate solutions to dominating set.

We first assume a normal distribution of the expression values of each gene, and estimate for it the mean and standard deviation. We do this separately for each of the two subject groups (normal and cancerous) using MATLAB. Then, based on the estimated normal distribution, we calculate the p-values for the original individual expression values. From this point on, it is perhaps easiest to envision our approach with the construction of a bipartite graph. In this graph, one set of vertices represents the genes, and the opposing set represents the samples of the test subjects. We place an edge between a gene and a sample if and only if the p-value of the expression value corresponding to that gene-subject combination is greater than 0.05. Following statistical convention, we consider a p-value below this cutoff to indicate an outlier.

In this setting, we want to identify the genes that dominate (or nearly dominate) all the samples. Therefore, we winnow out from consideration any vertex (gene) whose degree with respect to either the normal or cancer samples is less than 90%. Thus, in the training dataset, a gene is eliminated if it is connected to fewer than 78 of the cancer patient samples or fewer than 9 of the normal group samples. The choice of 90% is somewhat arbitrary, but was selected only after extensive testing. Next, in an effort to to remove any remaining genes with a low possibility of discriminating between the two groups, we calculate the p-values for tests of equal means using both the Wilcoxon and t-test methods. We use both since the t-test assumes a normal distribution, while the Wilcoxon test does not. Only genes for which both p-values are less than 0.05 are retained.

Finally, for those genes that still remain, we generate scores according to the following strategy. Using a Wilcox Rank-Sum test of the two group means, we calculate the p-value of the hypothesis of equal mean. Since a smaller p-value indicates a greater probability that the two groups' expression values are different for a given gene, we use 1/p-value for the gene score.

procedure dominating-set-winnow

```
initialize edge-weighted bipartite graph of order n+m
  for i=1 to m
    for j = 1 to n
      determine the p-value (weight) of each edge(i,j)
set threshold to 0.05 and eliminate edges of low weight
delete genes that dominate < 90% of cancer samples
delete genes that dominate < 90% of normal samples
n = n - |deleted genes|
for i=1 to n
generate p-value of equal mean using Wilcoxon and t-test
delete genes with p-value greater than 0.05 for either test
n = n - |deleted genes|
for i=1 to n
  set gene score to  1/p-value
return remaining genes and their scores
```

Finally, and most importantly, we sort the genes by score and compute the intersection of the genes identified by the clique-based approach described in the last section with a like number (118) of the highest-scored genes chosen by the dominating set method.   We are left with a set containing 75 genes that have passed both the clique and the dominating set tests.

6. RESULTS

Having completed the training phase, we proceed to testing on a new dataset under the assumption that we will not know in advance which samples correspond to cancer subjects and which to patients categorized as normal. The frequency histogram that follows shows the distribution of the edge-weight scores generated on a representative section of the Harvard dataset. If our method is to be predictive, we expect to see something of a bimodal distribution. This is because weights between cancer patients will be high, weights between normal subjects will be high, and weights between members of the two groups will be low. And a bimodal distribution is in fact what we observe in Figure 3.
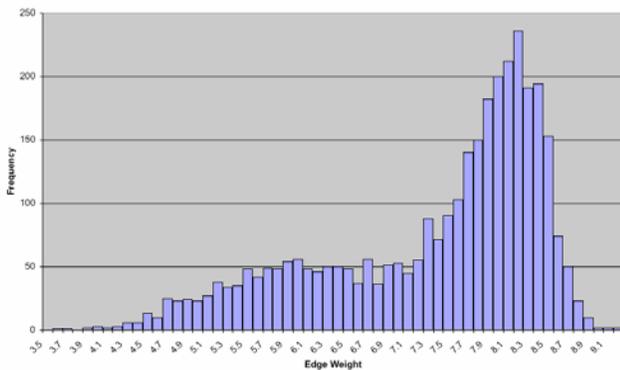

Figure 3. Weights between patient pairs using 75 genes (Harvard data)

We use this property when carrying out threshold selection, choose an initial threshold slightly to the right of the median edge-weight value. We then enumerate all maximal cliques in the unweighted graph, and check to see that every subject is in at least one clique. Assuming this is true, we incrementally select higher and higher thresholds until we generate an unweighted graph that has at least one subject that is not in any clique. At this point, we use the previous threshold and analyze the data by testing the supposition that all cliques of significant size (e.g. $k>5$) are uniform (contain only cancer patients or only normal subjects).

When the above-mentioned iterative process is carried out on the Harvard dataset, without the use of any previous knowledge pertaining to their classifications, we are able to separate the subjects into cancerous cliques and normal cliques almost flawlessly. In fact, out of the 585 cliques of size greater than three in the resulting graph, only three cliques have both cancerous and normal subjects, and these are very small (one each of size 4, 5, and 6). The resultant frequency distribution of these cliques is depicted in Figure 4.
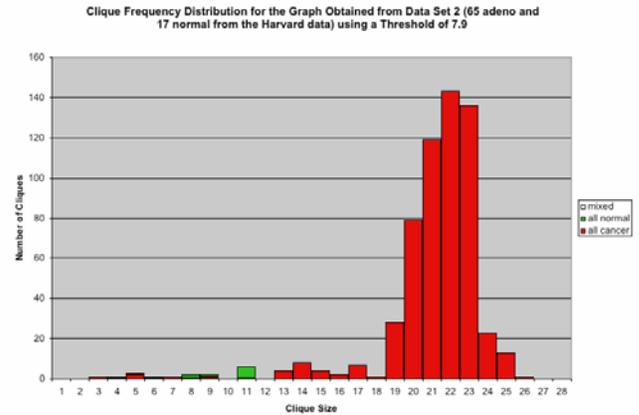

Figure 4. Cliques found using 75 genes (Harvard data)

7. CONCLUSIONS

There is no apparent consensus as to the best approach to mining microarray data.   Popular methods in current use include Bayesian networks [11,14], hierarchical clustering, or scale-free networks [13], to name a few.   We posit that the new methodology we have described here may both complement these well-known techniques as well as be of independent interest. Not only does it serve as a potential disease screening tool, but it also aids in the identification of sets of genes with good predictive quality, that is, those that are good discriminators between adenocarcinoma and normal tissues.   Our novel clique enumeration and dominating set based approach allows us to

distinguish these genes more effectively than would be possible with either approach alone.

Because our algorithm, once trained on a consistent data set for a particular type of cancer, does not require any additional knowledge to discriminate between types for which it has been trained, we feel that it may be utilized together with other methods to strengthen disease prediction through the use of microarray analysis. With the high-fidelity that the resulting cliques partition adenocarcinoma and normal samples, as shown in Figure 1, our algorithm should be a highly reliable tool for cancer prediction.

As a further proof of principle, several of the seventy-five genes we have identified as discriminators are known or suspected to play a role in oncogenesis. Among these are: CYP4B1, a cytochrome P450 enzyme that has been implicated in both bladder and lung cancer in humans[4,10]; FHL1, shown to have cytotoxic effects on melanoma cell lines and to possibly play a role in cellular differentiation[15]; the p85 alpha subunit of phosphoinositide-3-kinase, which plays a role in human breast cancer[12]; and tetranectin, which has already been shown to have to prognosticative value for survival rates at certain stages of ovarian cancer[9].

REFERENCES

1. Abu-Khzam FN, Langston MA, Shanbhag P. Scalable Parallel Algorithms for Difficult Combinatorial Problems: A Case Study in Optimization. *Proceedings, International Conference on Parallel and Distributed Computing and Systems*, 2003, to appear.

2. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA*. 98 (24), 13790-13795, 2001.

3. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* 9 (816), 816-824, 2002.

4. Czerwinski M, McLemore TL, Gelboin HV, Gonzalez FJ. Quantification of CYP2B7, CYP4B1, and CYPOR messenger RNAs in normal human lung and lung tumors. *Cancer Res*. 54(4):1085-91, 1994.

5. Das R, Mahabeleshwar GH, Kundu GC. Osteopontin stimulates cell motility and nuclear factor kappaB-mediated secretion of urokinase type plasminogen activator through phosphatidylinositol 3-kinase/Akt signaling pathways in breast cancer cells. *J Biol Chem*. 278(31):28593-606, 2003.

6. R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Springer-Verlag. 1999.

7. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comput Biol*. 7(3-4):601-20, 2000.

8. Garey MR, Johnson DS. Computers and Intractability. W. H. Freeman, New York, 1979.

9. Hogdall CK, Norgaard-Pedersen B, Mogensen O. The prognostic value of pre-operative serum tetranectin, CA-125 and a combined index in women with primary ovarian cancer. *Anticancer Res*. 22(3):1765-8, 2002.

10. Imaoka S, Yoneda Y, Sugimoto T, Hiroi T, Yamamoto K, Nakatani T, Funae Y. CYP4B1 is a possible risk factor for bladder cancer in humans. *Biochem Biophys Res Commun*. 277(3):776-80, 2000.

11. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4(2): 249-264. 2003.

12. Mahabeleshwar GH, Kundu GC. Syk, a protein-tyrosine kinase, suppresses the cell motility and nuclear factor kappa B-mediated secretion of urokinase type plasminogen activator by inhibiting the phosphatidylinositol 3'-kinase activity in breast cancer cells. *J Biol Chem*. 278(8):6209-21, 2003.

13. del Rio G, Bartley TF, del-Rio H, Rao R, Jin KL, Greenberg DA, Eshoo M, Bredesen DE. Mining DNA microarray data using a novel approach based on graph theory. *FEBS Letters* 509(2):230-4, 2001.

14. Sok JC, Kuriakose MA, Mahajan VB, Pearlman AN, DeLacure MD, Chen FA. Tissue-specific gene expression of head and neck squamous cell carcinoma in vivo by complementary DNA microarray analysis. *Arch Otolaryngol Head Neck Surgery* 129(7):760-70, 2003.

15. de Vries JE, Meyering M, van Dongen A, Rumke P. The influence of different isolation procedures and the use of target cells from melanoma cell lines and short-term cultures on the non-specific cytotoxic effects of lymphocytes from healthy donors. *Int J Cancer*. 15(3):391-400, 1975.