

# The Analysis of Gene Markers and the Use of SAS Procedures to Examine MicroArray Data

Patricia B. Cerrito, University of Louisville and Jewish Hospital Center for Advanced Medicine, Louisville, KY

## ABSTRACT

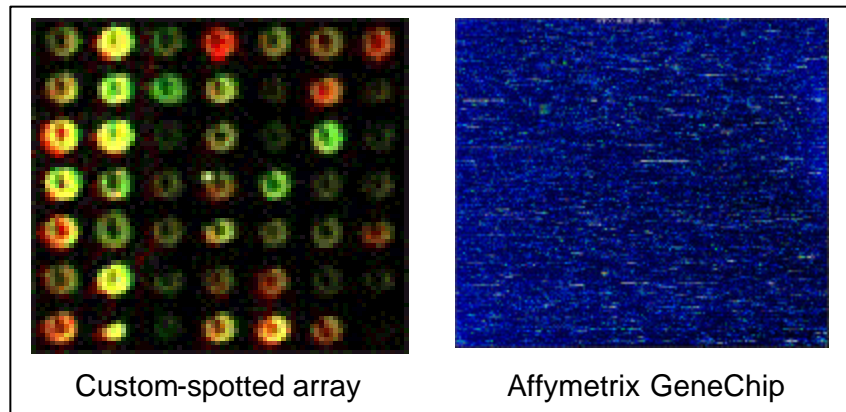
The purpose of this paper is to demonstrate SAS procedures that can be used to investigate gene marker data. Gene markers are often studied through microarray analysis where a chip containing identified genes is treated with a specific drug and then examined. The paper includes discussion of the use of SAS/Genetics and SAS/MicroArray as well as Proc KDE and Proc CALIS in SAS/Stat. Proc KDE is used to find the population distribution of gene markers for one treatment, or multiple treatments. Proc CALIS is used to develop structural equation models that can provide a means to examine the relationships between gene markers.

## INTRODUCTION

SAS/Genetics and MicroArray Solutions have been developed specifically to work with the analysis of gene markers. An examination of the current literature on microarray analysis indicates that few of these techniques are in common use to examine gene expressions. It is the purpose of this paper to examine a number of typical presentations of data results, and then to compare with the methodology developed in SAS.

A microarray consists of sets of miniaturized chemical reaction areas that may also be used to test DNA fragments, antibodies, or proteins. It is a snapshot to capture the activity pattern of thousands of genes simultaneously (Figure 1).<sup>4</sup>

Figure 1. Picture of a microarray



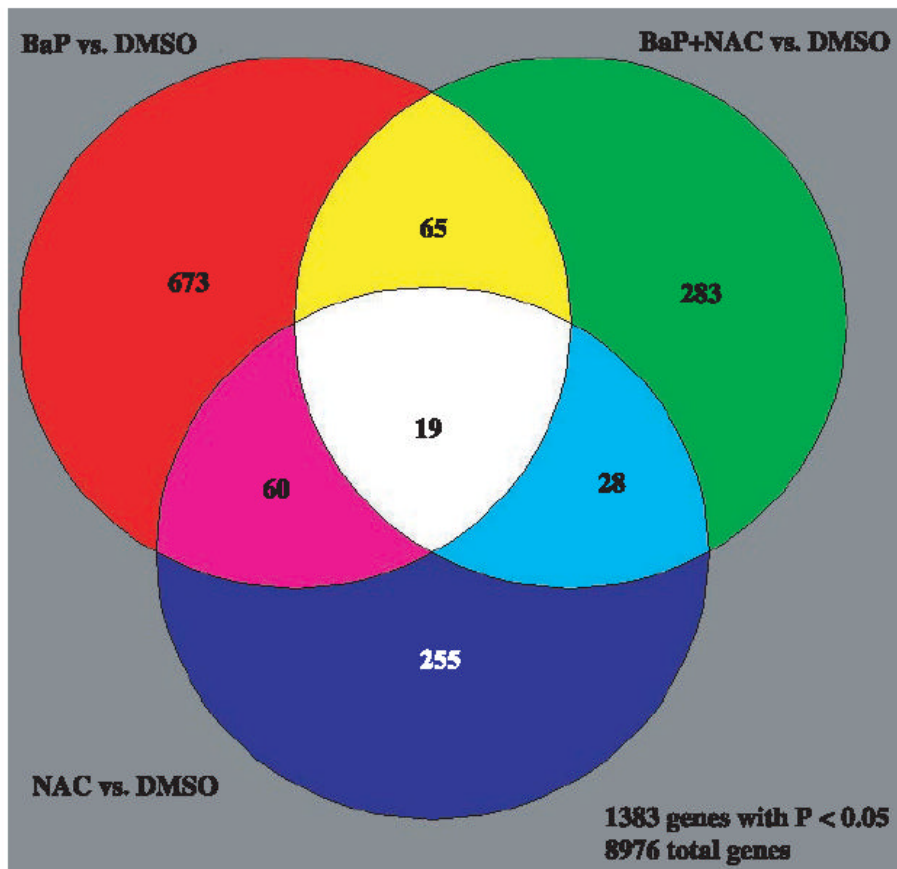
The advantage of using the Affymetrix GeneChip (Affymetrix, Inc.; Santa Clara, CA) is that it enables whole-genome surveys by simultaneously tracking nearly 1,500 genetic variations, known as single nucleotide polymorphisms (SNPs), dispersed throughout the genome. A number of different treatments (with replicates) are performed, each using a GeneChip. Each chip has a number of known gene markers (typically several thousand). The data are measured in light intensity values that are normalized for analysis purposes.

## PAIRWISE COMPARISONS FOR SIGNIFICANCE

The gene markers are compared pairwise using t-tests (Figure 2). The purpose is to find which gene markers can distinguish between treatments.<sup>5</sup> Consider that the total number of pairwise t-tests described in Figure 2 is equal to 10,359. Of that number, 1383 tests are statistically significant. The Venn Diagram indicates where significant differences occurred. This is an astonishing number of t-tests, with no attempt to correct for the accumulation of error. Typically, Bonferroni's correction is used. However, with so many t-tests, Bonferroni's is very inadequate. SAS/Genetics has a procedure, PSMOOTH that examines the entire set of p-values.<sup>6</sup> This procedure is new to SAS 9.1. Its purpose is to smooth p-values over windows of markers.

The basic SAS commands are

```
Proc Psmooth bw=20;  
Var testvalue;
```



Since the data usually come with the genes listed as observations and the treatments as fields, there is considerable pre-processing that has to be done in order to use Psmooth. The SAS code for this processing is available.<sup>3</sup> The statement

```
ods output diffs=d;
```

needs to be used in the pairwise comparisons to save the p-values so that they can be used in the Psmooth analysis.

## PREDICTORS

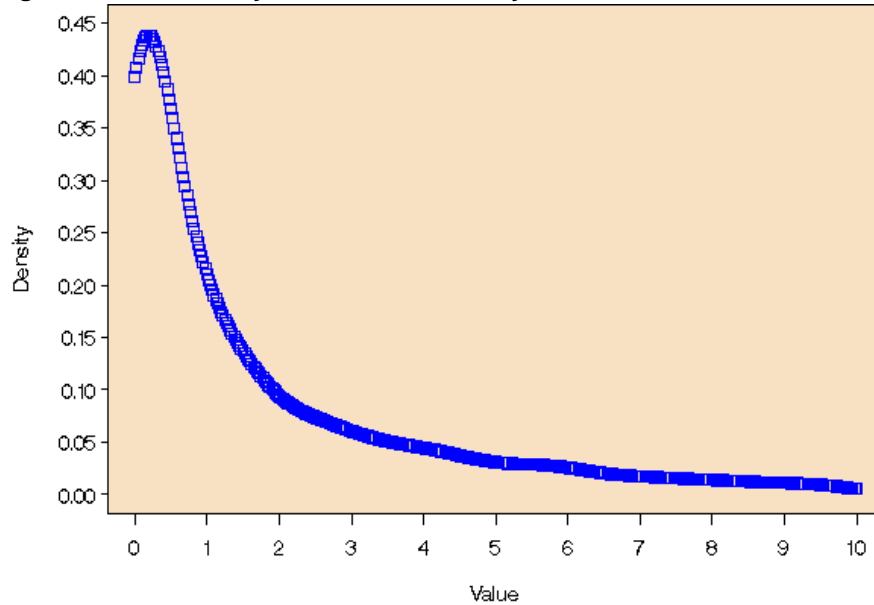
The treatments are compared across gene-markers. Since there are so many different gene markers, the total number of data points is too large for an ANOVA test. Therefore, the investigator typically reduces the number of gene markers in the ANOVA to a number under 500. There are several methods for reducing the number of markers, although almost all are related to an examination of correlations between genes. SAS/MicroArray Solution performs a gene-by-gene mixed models analysis of variance. The primary purpose is to find the optimal genetic markers that can be used to predict the results of a chosen gene. SAS/MicroArray Solution uses the standard Proc Mixed statements, included as a macro. However, it has also combined data visualization from JMP.

```
Class variables;
Model dependent=independent + 2-way and 3-way interactions/outp=residuals;
Random array;
Lsmeans 3-way interactions;
Estimate commands;
```

A second method for determining the best predictors for a particular gene is that of cluster analysis. SAS/MicroArray Solution uses hierarchical clustering procedures. Again, the graphics included are generated by JMP. The mixed models SAS code for performing gene-gene analyses are available online.<sup>3</sup> Statements to pre-process the data are also available online. Since the microarray data are usually listed in a spreadsheet so that the genes are listed as observations, and the treatments are identified as fields, the dataset has to be shifted so that

When comparing the probability distributions of two different treatments on the same genes, they are clearly exponential and it becomes difficult to distinguish them (Figure 3). The graph is substantially skewed, and the intensity levels can be larger than 10,000. Using the initial definition of a sample space, the set of gene markers such that the value is between a and b can be identified.

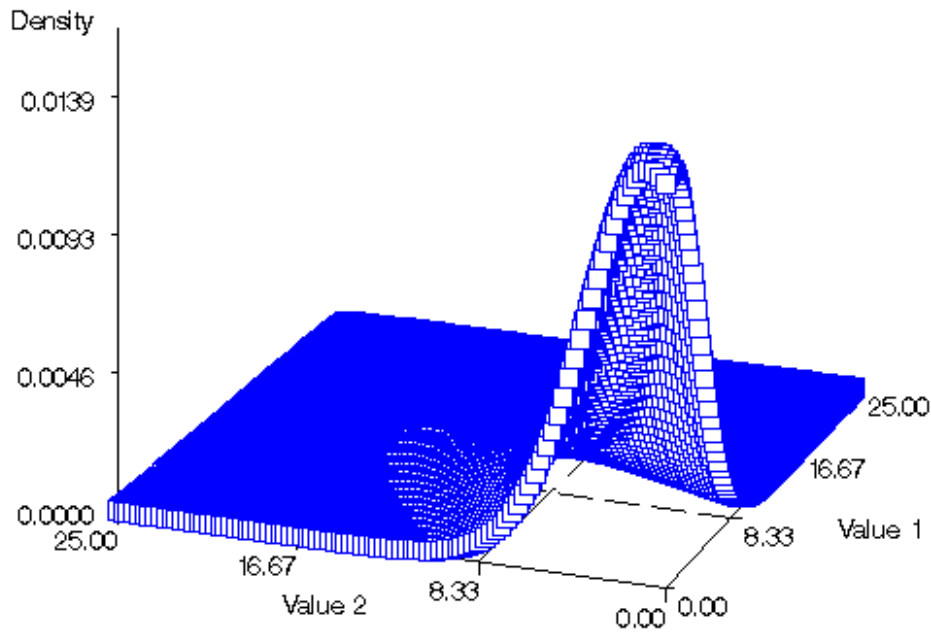
**Figure 3. Kernel Density Estimation of Intensity Levels for One Treatment**



Those same gene markers will correlate the most highly with any particular gene. Therefore, a set of predictors can be identified using the graphs of the intensity levels together with the subset of gene markers within a given radius. It is also possible to define a multidimensional density estimate. Replicates can be used to define one estimator per treatment, and the different treatments can give a set of predictor genes of radius  $k$  in the neighborhood of the selected gene.

Because of the exponential nature of the data, there are many gene markers that can be used to predict genes with intensity values in the low range. The key is to find predictors that can function at the high end. Consider an example with one control and one treatment (Figure 4).

**Figure 4. Bivariate Kernel Density Estimator of Gene Markers**



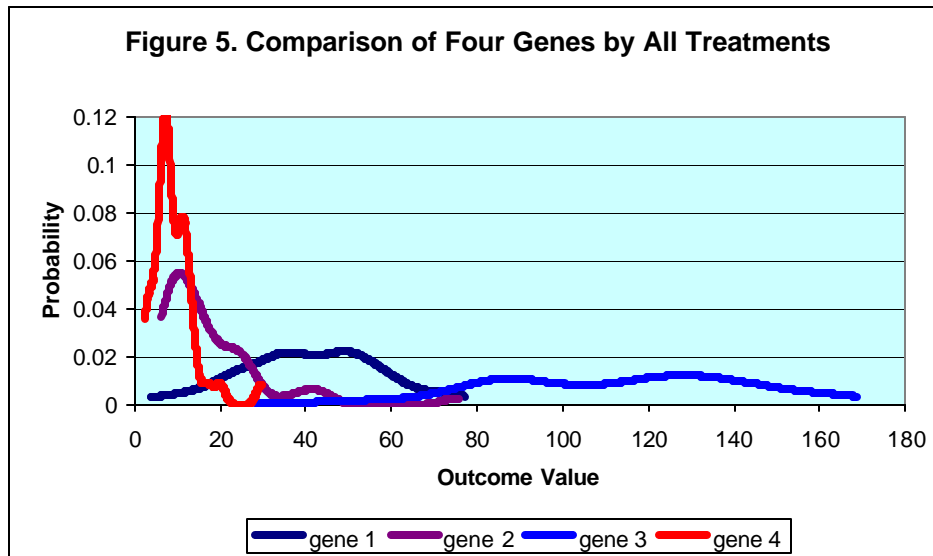
Using the where command, it is possible to filter down to the possible predictors. There are 40 gene markers with values greater than 100 on both the control and treatment. Those 40 markers then are optimal in predicting each other. The probability that the control is greater than 100 while the treatment is less than 100 is quite small. Although the distribution looks reasonably exponential, it is very highly skewed as the intensity value gets large.

Proc KDE in SAS/Stat (considerably revised in version 9.1) can perform both univariate and bivariate estimates of population distributions. Once computed, it is possible to estimate confidence widths. The basics for Proc KDE are

```
Proc KDE data=;
Univar var1/grid1=0 gridu=10 method=srot out=outkde;
Bivar var2 var3 / out=outkde2;
```

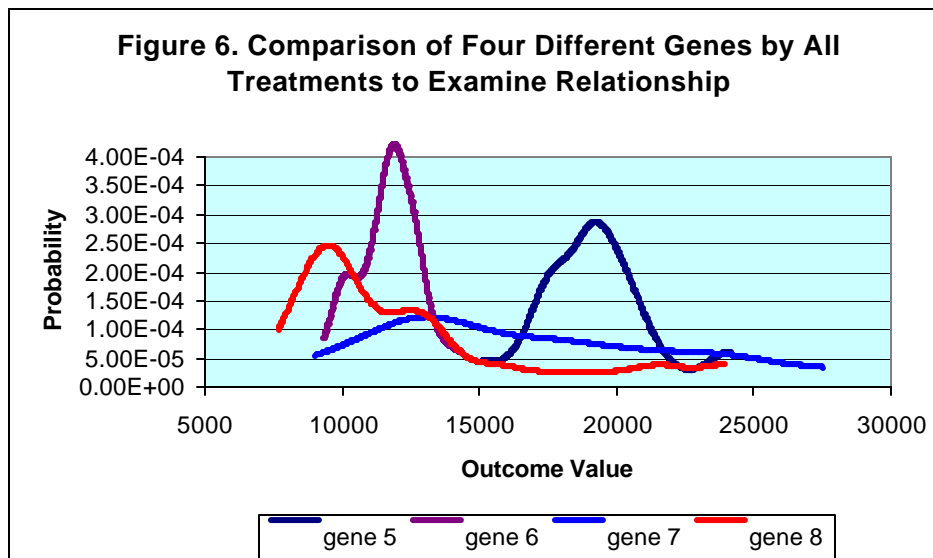
Nearest neighbors are equal to the set  $\{Genes | P(\text{predictedgene}) - \epsilon < P(\text{gene}) < P(\text{predictedgene}) + \epsilon\}$ .

It is also possible to transpose the microarray so that each gene represents a variable, and each treatment represents an observation. In this way, different gene patterns can be examined using PROC KDE, and the different density functions can be overlaid (Figure 5).



Note that these four genes behave very differently from each other. Gene 4 stays within a very narrow range across all treatments compared to genes 1 and 2. This indicates that their ability to predict one other is extremely limited. Consider also the fact that the behavior of each gene is not Gaussian.

This suggests that linear measures such as correlation may not be appropriate in defining predictors. A second set of four is also compared (Figure 6). In this case, gene 8 more closely resembles gene 6 for small values, and gene 7 for larger values. Again, the normal distribution is not a valid assumption for analysis.



It is possible to develop a batch process that will compare all combinations of genes to find the distributions that most closely resemble each other. However, it will be more efficient to use the nearest neighbor choices followed by a gene-by-gene analysis using kernel density estimation.

SAS code is provided in Wolfinger, et.al.<sup>3</sup> to examine the gene-gene interaction using mixed models. The code has been modified slightly for Affymatrix gene chips.

```
proc mixed data=sasuser.ramoscombined covtest cl lognote;
class treatment;
```

```

model outcome = treatment /
  outp=work.sudarp (keep=
    gene name treatment resid);
lsmeans treatment / diff cl;
run;

```

Gene represents the Affy ID; name the specific gene name associated with the Affy ID, and treatment is the treatment applied to the specific chip. The next step is to remove some genes that slow the analysis:

```

data work.sudarp;
  set work.sudarp;
  where gene not in ('EMPTY', 'NORF');
run;

proc sort data=work.sudarp;
  by gene treatment;
run;

```

The next step is to use an incomplete block design:

```

proc mixed data=work.sudarp;
  by gene;
  class treatment;
  model resid = treatment / outp=work.sudarr;
  lsmeans treatment / diff;
  ods output covparms=work.sudarc tests3=work.sudart
    diffs=work.sudard;
run;

```

The final step is a Bonferroni correction:

```

data work.sudard1;
  set work.sudard;
  if (probt = .) then delete;
  logp = log10(probt);
  neglogp = -logp - log10(6917*10);
  if (neglogp <= 0) then delete;
run;

```

The results provide the relationship between genes. The best predictors of a particular gene can be used to develop a network using structural equation modeling.

MicroArray Solution provides graphic representations of the mixed models results so that the best predictors can be identified easily.

## STRUCTURAL EQUATION MODELING

Once the primary predictors are located, the process can be repeated so as to create a network that related process to gene. This is an extremely difficult and time-consuming task. Once the network is diagrammed, the validity of the diagram must be examined. However, a method to determine validity has already been developed, although its application has been reserved primarily for the social and behavioral sciences; the method of structural equation modeling. The method is also called path analysis. The basic structural equation model has the form

$$\begin{array}{ccccccc} \text{Effect} & & \text{Structural} & \text{Causal} & & & \\ & = & \text{Sum} & \text{X} & + & \text{Disturbance} & \\ \text{Variable} & & \text{Coefficient} & \text{Variable} & & & \end{array}$$

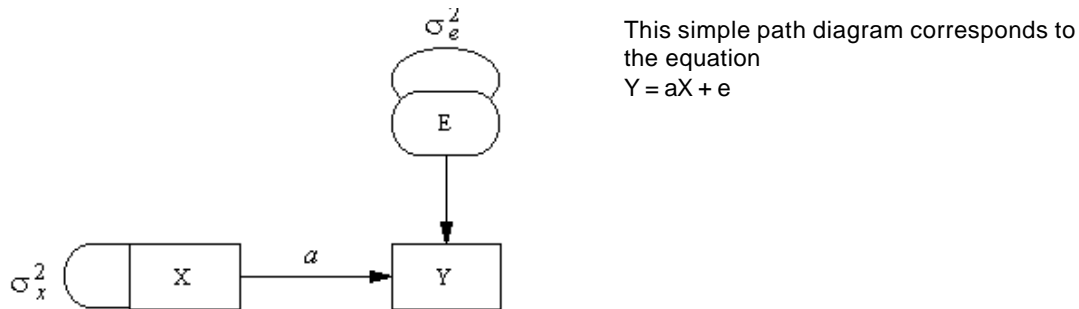
These are the following types of connections between variables to be considered:

1. Direct effect: Either X causes Y, Y causes X, or both.
2. Indirect effect: The relationship between X and Y is said to be indirect if X causes Z which in turn causes Y.
3. Spuriousness: The relationship between X and Y is said to be spurious if Z causes X and Y.

- Unexplained covariation: Both X and Y are exogenous and so variation between them is not explained by the model.

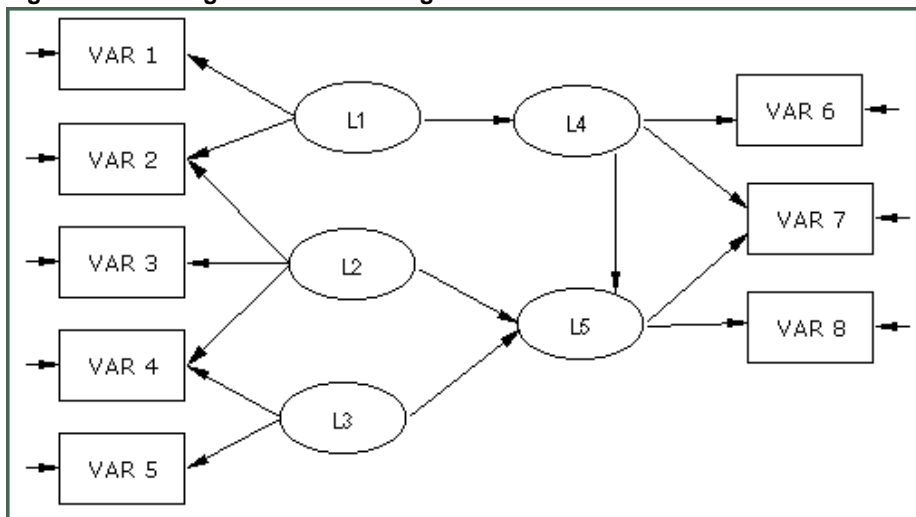
The process is to first identify the relationship between variables, and then to validate the model. Proc Calis in SAS/Stat is used to perform structural equation modeling. It is somewhat difficult to use since the entire variance-covariance estimate needs to be specified in developing the model. It is best to diagram the model prior to entering it into SAS. Proc Calis can codify a path diagram as shown in Figure 7.<sup>7</sup>

**Figure 7. Path Diagram**



However, path diagrams can become much more complicated and can include latent variables (Figure 8).<sup>8</sup>

**Figure 8. Path Diagram Modeled Using the LISREL Procedure in Proc Calis**



The application of path analysis to gene markers has been somewhat minimal.<sup>1</sup>

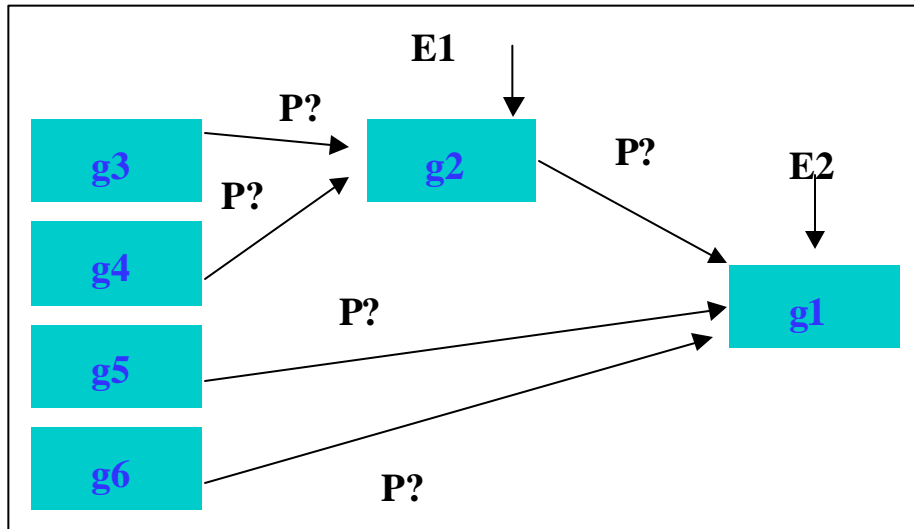
Consider the following terminology:

- An **Endogenous** variable is one whose variability is predicted to be causally affected by other variables in the model.
- An **Exogenous** variable is a construct influenced only by variables that lie outside of the causal model.
- A **manifest** variable is one that is directly measured or observed.
- A **latent** variable is a hypothetical construct not directly measured or observed.

In Figure 6, variables 1-5 are exogenous variables; variables 6-8 are endogenous. The hypothesis to be tested is that the data fit the model. All are manifest variables. Variables L1-6 are latent and hidden.

The hypothesis to be tested is that the model fits the data. There should be a minimum of 200 data values, with a minimum of 5 values per parameter. The connections between variables are linear. Only exogenous variables have covariances built into the model; endogenous variables have residuals. Consider the model in Figure 9, with the outline of the construction of a path diagram.

Figure 9. Construction of a Path Diagram for Gene Markers



The p values listed in Figure 9 represent coefficients that relate the gene markers, and E represents residuals so that  $g_2 = p_{23}g_3 + p_{24}g_4 + E_2$ . Other equations are defined similarly. The correlations are computed and entered in a data statement:<sup>2</sup>

```
Data d1 (type=corr);
Input _type_ $ _name_ $ g1-g6;
Cards;
N          .          value 1 value 2 ... value 6
Std        .
Corr g1    1.0000 . . . . .
Corr g2    .6742 1.0000 . . . .
...
```

followed by PROC CALIS

```
Proc calis;
Lineqs;
g1=pg1g2 g2 + pg1g5 g5 + pg1g6 g6 + E1,
g2=pg2g3 g3 + pg2g4 g4 + E2;
Std
E1=varE1,
E2=varE2,
g3=varg3,
g4=varg4,
g5=varg5,
g6=varg6;
COV
g2 g4=cg3g4,
g3 g5=cg3g5,
g3 g6=cg3g6,
g4 g5=cg4g5,
g4 g6=cg4g6,
g5 g6=cg5g6;
Var g1 g2 g3 g4 g5 g6;
Run;
```

It is also to use the raw data, without computing the correlations separately:

```
Proc corr data=dataset nomiss noprob;
Var g1 g2 g3 g4 g5 g6;
Run;
Proc Calis covariance corr residual modification;
```

The model is an ideal fit provided

1. Absolute value of entries in the normalized residual matrix do not exceed 2.0
2. The p-value associated with the model chi-square should not be statistically significant.

3. The comparative fit index and the non-normed fit index should both exceed 0.9.
4. The correlation obtained for each endogenous variable should be large  
The absolute value of the t statistics for each path coefficient should exceed 1.96 and the standardized path coefficients should be nontrivial in magnitude.

## TEXT ANALYSIS

A second interesting application of text mining is to use it to examine the literature to find papers on gene expressions in the medical literature (Figure 10). It is highly plausible to be able to extract new information from such collections of text documents because investigators still can only read a small subsample of the whole.<sup>9</sup> The results to date are very promising.<sup>10</sup>

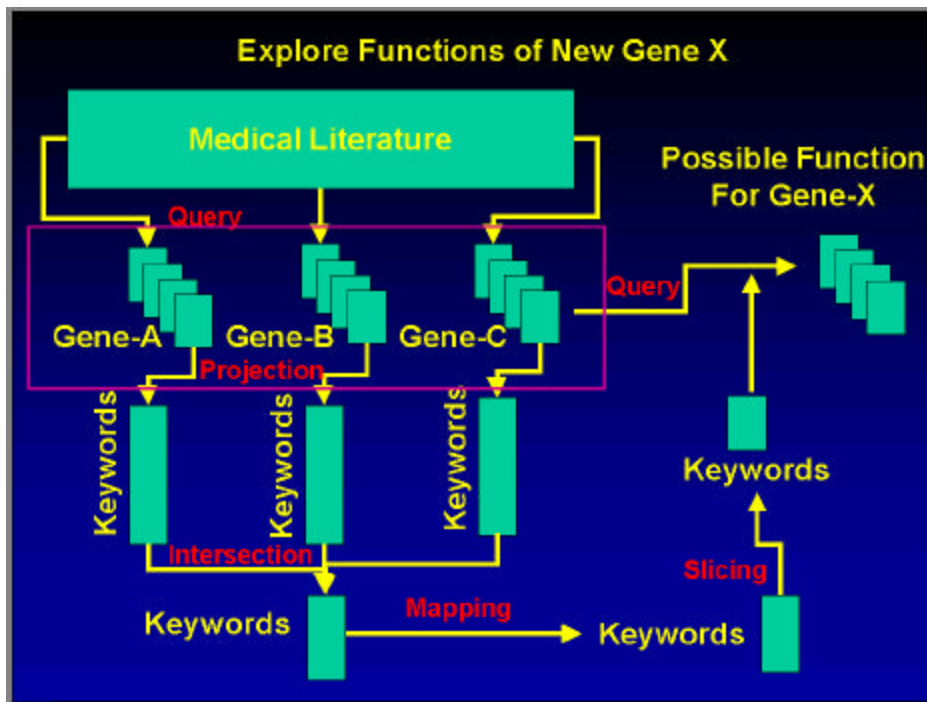


Figure 10. Using text analysis to find papers in the medical literature containing information about genes A,B,C.

The basics of text analysis are<sup>11</sup>

1. coding (involves simply determining the basic unit of analysis, and counting how many time each word appears.)
2. categorizing (I nvolves creating meaningful categories to which the unit of analysis (for example, "terms signifying 'cooperation' and terms signifying 'competition') can be assigned.)
3. classifying (involves verifying that the units of analysis can be easily and unambiguously assigned to the appropriate categories.)
4. comparing (involves comparing the categories in terms of numbers of members in each category.)
5. concluding (involves drawing theoretical conclusions about the content in its context.)

To demonstrate the technique, the keywords “diabetes” and “gene expression” were used in Medline. The search returned 147 documents, with the earliest dated 1996. Text Miner was used to investigate the returned abstracts, resulting in 4 clusters (Table 1).

Table 1. Clustering of Text Documents

Cluster Number	Descriptive Terms	Frequency
1	Factor, cell, type, gene, key, diabetes	16
2	Mouse, cell health, diabetes, science, result history	41
3	Level, expression, gene, science, health	53
4	Type, unique identifier, diabetes, history	32

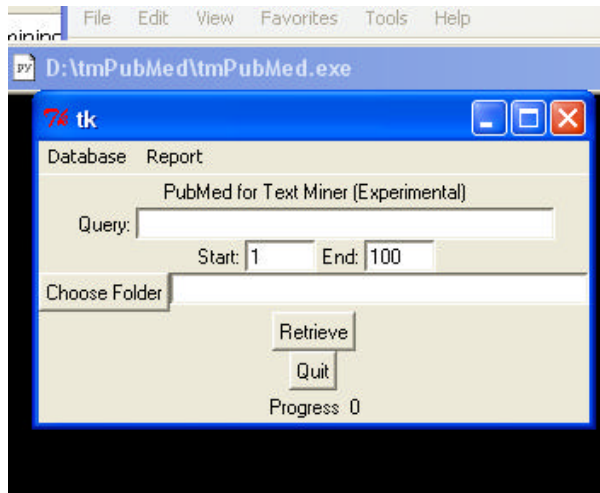
The unique identifiers are located in cluster 4, reducing the number of papers to be read to 32 from 147. Version 5.0 of Enterprise Miner can also develop concept maps that link terms to documents. The term “pancreatic beta-cells” is given in Figure 11.



To find documents on the internet linked to an initial web site, the tmfilter macro can be modified:

```
%tmfilter(url=http://directory.google.com/Top/Health/Conditions_and_Diseases/Infectious_Diseases/Viral/Influenza/?tc=1,  
depth=2,dir=c:\vaccine\dir,  
destdir=c:\vaccine\destdir,norestrict=1,  
dataset=work.vaccinewebcrawl);
```

The depth gives the depth of links that tmfilter will find. If the depth is too high, there will be too many irrelevant sites found, cluttering up the hard drive unnecessarily. The macro does not work well with a Google search since Google blocks the crawler and a key is required. To find abstracts in Medline, a program was developed by Russell Albright of the SAS Institute to search PubMed (Figure 13).



**Figure 13.** Executable file to search Medline using a keyword search, and depositing the abstracts into a directory on the hard drive.

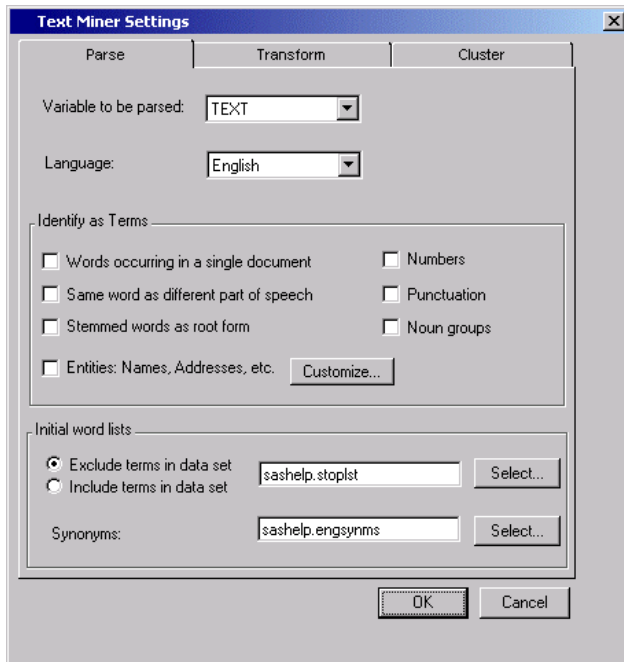
After completion, the following SAS code will return the abstracts into a SAS dataset:

```
filename fetch 'C:\search\fetch.xml';  
filename SXLEMAP 'C:\tmpubmed\pubmed.map';  
libname fetch xml xmlmap=SXLEMAP access=READONLY;
```

The dataset will have the name fetch.pubmedarticle.

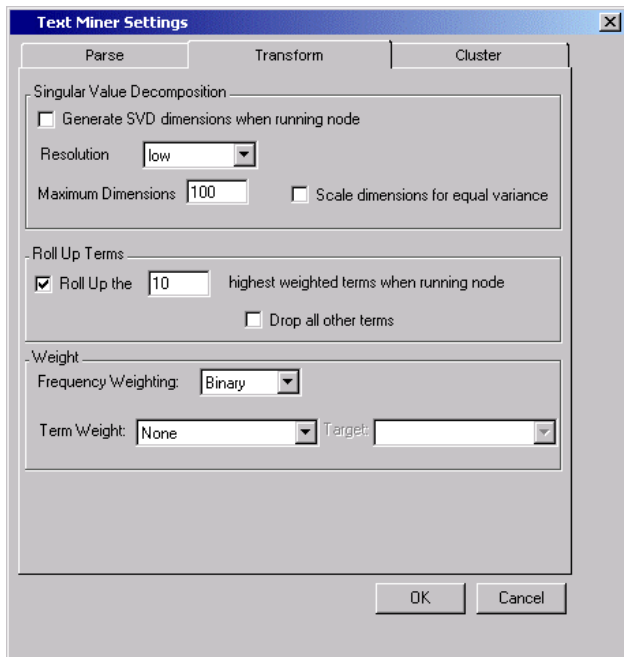
Enterprise Miner is used by connecting procedure icons. For Text Miner, the node contains three tabs (Figures 14,15,16).

Figure 14. Parsing Tab in Text Miner



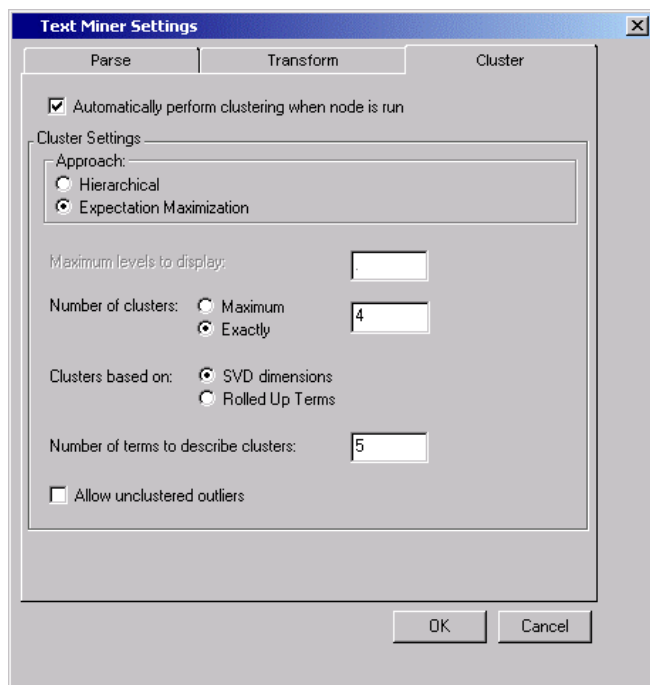
This tab allows specification of words and language. Numbers can be used, and a specific gene sequence can be included in the start list to identify only those documents containing that gene sequence.

Figure 15. Transform Tab in Text Miner



This tab allows for the user to change the weighting formula, and to specify the number of dimensions in the singular value decomposition.

**Figure 16. Cluster Tab in Text Miner**



This tab allows the user to specify the number of clusters, and the number of terms to list in each cluster.

It is possible to compare documents from two different sources. A field is defined to indicate document source. Then the two datasets are concatenated. Text Miner is used on the combined dataset, with a chi-square analysis on cluster and document source. A followup document search was conducted on the term “pancreatic beta cells” using Google (121 documents) and Medline (125 documents). Ten clusters were identified in the analysis ( $p < 0.0001$ , Table 2).

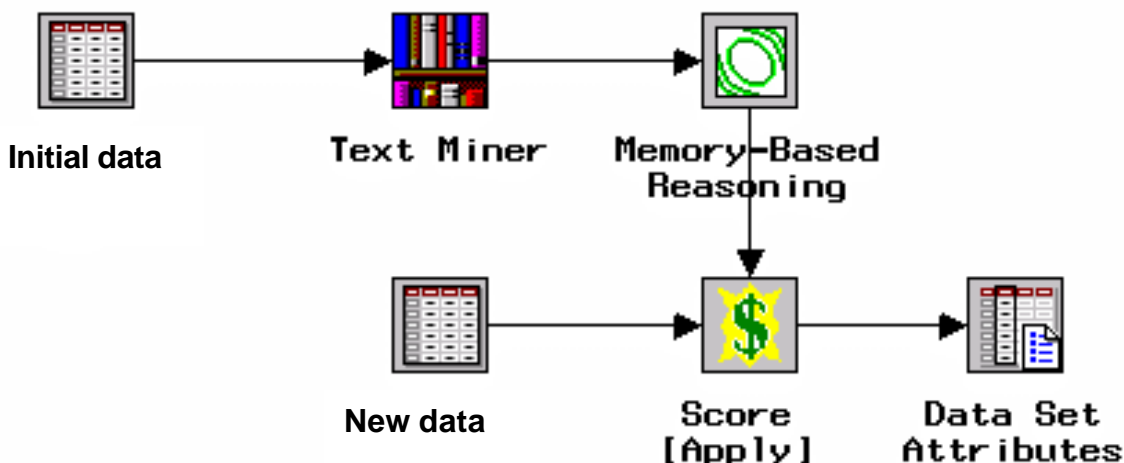
**Table 2. Clusters from Text Miner for “pancreatic beta-cells”**

Cluster Number	Descriptors	Medline	Google
1	Burst, figure, project, back, image, genetic, model, beta, research, pancreatic, novel, cell, diabetes	0	31
2	Growth, express, expression, function, embryonic, glucose-stimulated insulin secretion, gene, center, signal, genome, clinical alerts, clinicaltrials.gov	14	6
3	Amelioration, archive, contents, depancreatized, thymidine incorporation, national academy of sciences, replication, surgical	0	12
4	Insulin-producing, disease, expression, factor, pancreatic beta-cells, activation, molecular, individual, target	12	12
5	Endocrine, department, regulate, cellular, metabolites, find, study, act, links, help, pancreatic, show, biology, tutorial, limit, structure, batch, clinical alerts	10	6
6	Consortium, fund, effort, investigator, guide, grant, advance, embryonic, stem, award, transplant, national, future, kidney, establish, support, transplantation	0	10
7	Hormone-sensitive, metabolites, glucose-stimulated insulin secretion	0	7
8	Help, show, tutorial, index, batch, clinical alerts, clinicaltrials.gov, clipboard, consumer health, department of health, gateway	61	1
9	Help, show, links, nucleotide, batch, clinical alerts, clinicaltrials.gov, consumer health, department of health, gateway	28	5
10	Recent, secrete, bind, blood, target, endocrine, hormone, award, burst, potential, research, model, release, cellular, associate	0	31

Clearly, Medline documents tend to group in 4 of the 10 clusters; the Google documents are much more diverse. The table suggests that clusters 8 and 9, with similar defining words should be combined into one cluster.

In addition, if a target value is defined then once an initial set of text documents has been clustered, a new set can be scored (Figure 17).

Figure 17. Scoring process for new documents<sup>12</sup>



The new data can be the initial dataset, with the target variable changed to “rejected” and the dataset role is set to “score”. The memory based reasoning node is set to “Process or Score: Training, Validation, Test”. The score node is set to select “Apply training data score code to score data set”. The dataset labeled EMDATA.SD\_XXXXX contains the scored data. There is a predicted target variable as well as the actual variable. A chi-square analysis was performed to compare the two (Table 3).

Table 3. Chi-square analysis of scoring data

Document Source	Predicted Medline	Predicted Google
Medline	125 (100%)	0 (0%)
Google	11 (9%)	110 (91%)

Again, the chi-square analysis ( $p < 0.0001$ ) indicates that there is a clear separation between documents on the internet and documents in the medical literature that are returned on a keyword search of “pancreatic beta-cells”.

## CONCLUSION

It is feasible to investigate statistical methods that have been developed for other applications, and to use them to investigate gene markers and microarray data. While MicroArray Solution and SAS/Genetics have specifically developed tools to examine such data, other procedures in SAS/Stat and SAS/Enterprise Miner can be applied as well.

## REFERENCES

1. Spirtes P, Glymour C, Scheines R, Kauffman S, Aimale V, Wimberly F. *Constructing Bayesian Network Models of Gene Expression Networks from Microarray Data*. Washington DC: Carnegie Mellon University; 2000.
2. Hatcher L. *A step-by-step approach to using SAS for factor analysis and structural equation modeling*. Cary, NC: SAS Institute, Inc.; 1994.
3. Wolfinger R, Gibson G, Wolfinger E, et al. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* 8:625-637. Available at: <http://statgen.ncsu.edu/ggibson/Pubs.htm>, 2004.
4. Institute S. MicroArray Solutions Course Notes: a Training manual. 2003:244, Cary, NC.
5. Johnson C, Balagurunathan Y, Lu K, et al. Genomic profiles and predictive biological networks in oxidant-induced atherogenesis. *Physiological Genomics*. 2003;13(3):265-275.
6. Anonymous. SAS/Genetics User's Guide. 2003:120, Cary, NC.
7. Anonymous. Stat Soft Electronic Statistics Text. *Stat Soft, Inc*. Available at: <http://www.statsoftinc.com/textbook/stsepath.html>, 2003.
8. Anonymous. Using LISREL to create a GIF file with a conceptual path diagram in 10 easy steps. *SSI Scientific Software*. Available at: <http://www.ssicentral.com/lisrel/gifpath2.htm>, 2003.
9. Hearst MA. Untangling Text Data Mining. *University of California at Berkeley*. Available at: <http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html>, 2003.

10. Cerrito P, Badia A, Cerrito J, Cox J. Use of Text Miner to Automatically Abstract Patient Information from the Pharmacy Order Database. In: Pharmasug, ed. *Pharmasug Proceedings*. Miami: SAS Institute, Inc.; 2003:365-370.
11. Martens BVdV. IST 501: Research Techniques for Information Management. Available at: <http://web.syr.edu/~bvmarten/index.html>. Accessed 2002, 2002.
12. Anonymous. Text Mining Using Enterprise Miner, Course Notes. 2003:200.

## **ACKNOWLEDGMENTS**

The author wishes to acknowledge support from the Jewish Hospital Center for Advanced Medicine in the development of this paper.

## **CONTACT INFORMATION (HEADER 1)**

Patricia B. Cerrito  
Department of Mathematics and Jewish Hospital Center for Advanced Medicine  
University of Louisville  
Louisville, KY 40292  
Work Phone: 502-560-8534, 502-852-6826  
Fax: 502-852-7132  
Email: [pcerrito@louisville.edu](mailto:pcerrito@louisville.edu)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.