Data Mining Supercomputing with SAS JMP® Genomics

Dr. Richard S. SEGALL*
Arkansas State University, Department of Computer & Information Technology
State University, AR 72467-0130, USA, rsegall@astate.edu

Dr. Qingyu ZHANG*
Arkansas State University, Department of Computer & Information Technology
State University, AR 72467-0130, USA, qzhang@astate.edu

and

Ryan M. PIERCE
Arkansas State University, Student Affairs Technology Services,
State University, AR 72567-0348, USA, rmpierce@astate.edu

ABSTRACT

JMP® Genomics is statistical discovery software that can uncover meaningful patterns in high-throughput genomics and proteomics data. JMP® Genomics is designed for biologists, biostatisticians, statistical geneticists, and those engaged in analyzing the vast stores of data that are common in genomic research (SAS, 2009).

Data mining was performed using JMP® Genomics on the two collections of microarray databases available from National Center for Biotechnology Information (NCBI) for lung cancer and breast cancer. The Gene Expression Omnibus (GEO) of NCBI serves as a public repository for a wide range of high-throughput experimental data, including the two collections of lung cancer and breast cancer that were used for this research. The results for applying data mining using software JMP® Genomics are shown in this paper with numerous screen shots.

Keywords: Microarray databases, Lung Cancer, Breast Cancer, Data Mining, Supercomputing, Gene Expression Omnibus (GEO), SAS JMP® Genomics.

1. BACKGROUND

The software used in this research is JMP® Genomics from SAS Institute, Inc. of Cary, NC that according to Product Brief of SAS (2009) dynamically links advanced statistics with graphics to provide a complete and comprehensive picture of results, whether the data comes from traditional microarray studies or data summarized from next-generation technologies. Preliminary work done by the authors for the visualization by supercomputing data mining using JMP® Genomics from SAS for similar data was presented in Segall et al. (2010) and (2009).

Some of the previous research that has been performed by others in the area of applications of supercomputing to data mining include those of Zaki et al. (1996) for parallel data mining, Thoennes and Weems (2003) for performance of data mining on complex microprocessors, and data mining of large datasets with geospatial information by the image spatial data analysis group (2009) and University of Illinois at Urbana-Champaign, and Wilkins-Diehr and Mirman (2009) for on-demand supercomputing for emergencies that includes discussions for applications to breast cancer diagnosis.

2. DATA

The Gene Expression Omnibus (GEO) is a public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community. These data include single and dual channel microarray-based experiments measuring mRNA, miRNA, genomic DNA (including arrayCGH, ChIP-chip, and SNP), and protein abundance, as well as non-array techniques such as serial analysis of gene expression (SAGE), and various types of next-generation sequence data. In addition to data storage, a collection of web-based interfaces and applications are available to help users query and download the experiments and gene expression patterns stored in GEO.

The data sets used in the research presented in this paper are those from the Gene Expression Omnibus (GEO) from the National Center of Biotechnology Information (NCBI). One set of data is that of expression data for lung cancer that was made public on August 30, 2008; and the other is that for gene expression profiling in breast cancer that was made public in February 2006.

Lung Cancer Data Used in This Paper

According to NCBI (2007), the detection, treatment, and prediction of outcome for lung cancer patients increasingly depend on a molecular understanding of tumor development and sensitivity of lung cancer to therapeutic drugs.

NCI (2007) states that the application of genomic technologies, such as microarray, is widely used to monitor global gene expression and has built up invaluable information and knowledge, which is essential to the discovery of new insights into the mechanisms common to cancer cells, resulting in the identification of unique, identifiable signatures and specific characteristics. According to NCBI (2007) it is likely that application of microarray may revolutionize many aspects of lung cancer being diagnosed, classified, and treated in the near future. NCBI (2007) used microarrays to detail the global gene expression patterns of lung cancer.

The overall design of NCBI (2007) as used in this paper consisted of adjacent normal-tumor matched lung cancer samples that were selected at early and late stages for RNA extraction and hybridization on Affymetrix microarrays. A total of 66 samples were used for microarray analysis in NCBI (2007), including pairwise samples from 27 patients, who underwent surgery for lung cancer at the Taipei Veterans General Hospital, two tissue mixtures from the Taichung Veterans General Hospital, two commercial human normal lung tissues, one immortalized, nontumorigenic human bronchial epithelial cell line, and 7 lung cancer cell lines.

Breast Cancer Data Used in This Paper

The breast cancer data set used in this research was obtained on the web from NCBI (2006), which analyzed microarray data from 189 invasive breast carcinomas and from three published gene expression datasets from breast carcinomas. NCBI (2006) identified differentially expressed genes in a training set of 64 estrogen receptor (ER)-positive tumor samples by comparing expression profiles between histologic grade 3 tumors and histologic grade 1 tumors and used the expression of these genes to define the gene expression grade index. The data set for the figures generated in this paper consisted of over 22,000 rows representing different variables.

The breast cancer data presented by NCBI (2006) was from 597 independent tumors were used to evaluate the association between relapse-free survival and the gene expression grade index in a Kaplan-Meier analysis. All statistical tests performed by NCBI (2006) were two-sided. The overall design of NCBI (2006) was 64 microarray experiments from primary breast tumors used as training set to identify genes differentially expressed in grades 1 and 3. NCBI (2006) design included 129 microarray experiments from primary breast tumors of untreated patients used as validation set to validate the list of genes and its correlation with survival.

3. RESULTS

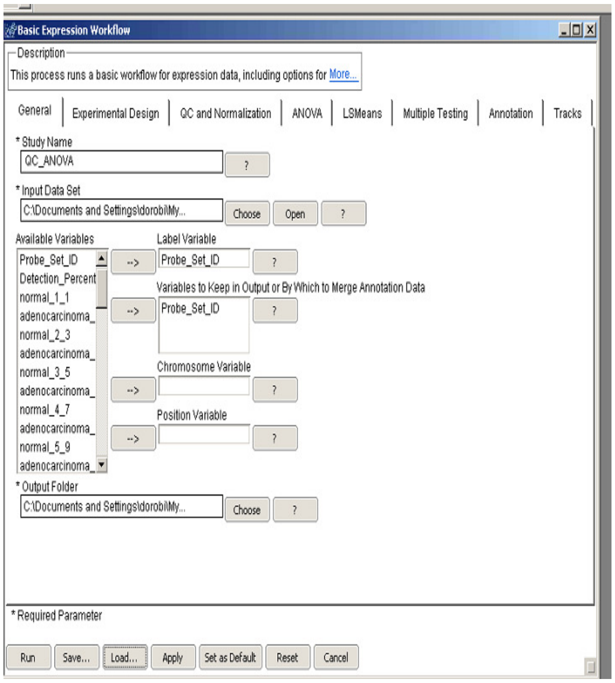
Data Mining Performed Using Sas Jmp® Genomics For Lung Cancer Data

Figure 1 shows the window called “basic expression workflow” that is the process that runs a basic workflow for expression data used to select variables of interest.

The data used for the lung cancer and its associated tumors consisted of over 22,000 rows representing genes and 54 columns representing samples as shown in Figure 2.

Our research using SAS JMP® Genomics yielded distributions plots of conditions, patients and characteristics; correlation analysis of principle components as shown in Figure 3 which shows “normal” versus “cancer” in the scatterplots , and dendograms of hierarchical clustering as shown in Figure 4. Figure 5 shows a Volcano plot of the summary plot of individual genes and their differences in condition of cancer from normal tissues.

Our research performed some predictive modeling using SAS JMP® Genomics that yielded one-way analysis plots of fitting a selected gene number 1773 by condition and also by patient as shown in Figure 6.



Description

This process runs a basic workflow for expression data, including options for More...

* Study Name: QC_ANOVA

* Input Data Set: C:\Documents and Settings\dorobi\My...

Available Variables: Probe_Set_ID, Detection_Percent, normal_1_1, adenocarcinoma_, normal_2_3, adenocarcinoma_, normal_3_5, adenocarcinoma_, normal_4_7, adenocarcinoma_, normal_5_9, adenocarcinoma_

Label Variable: Probe_Set_ID

Variables to Keep in Output or By Which to Merge Annotation Data: Probe_Set_ID

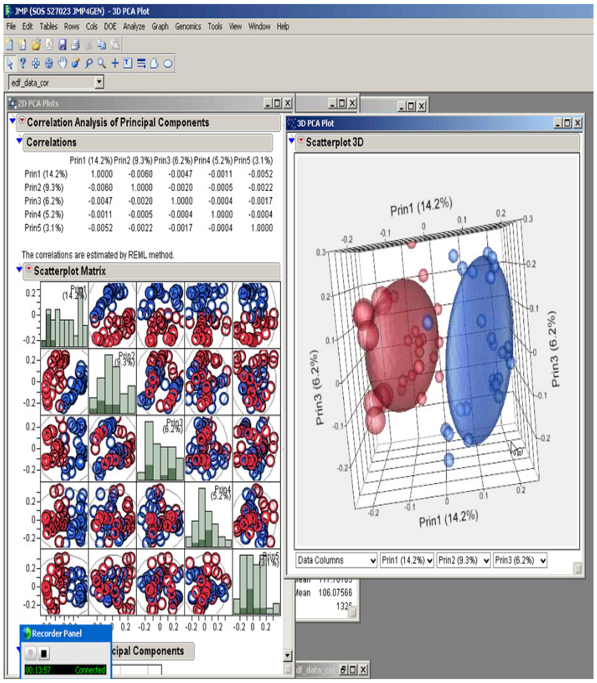
Chromosome Variable

Position Variable

* Output Folder: C:\Documents and Settings\dorobi\My...

* Required Parameter

Figure 1. Basic expression workflow



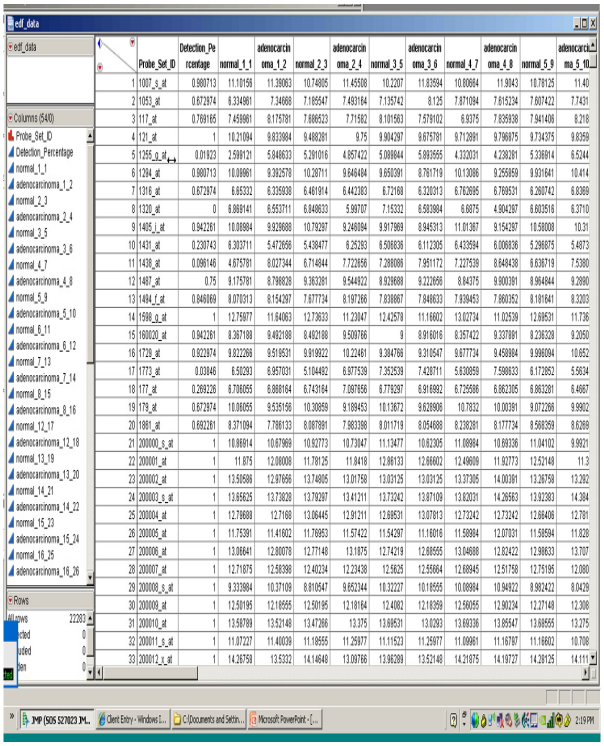
Correlation Analysis of Principal Components

Correlations

	Prin1 (14.2%)	Prin2 (9.3%)	Prin3 (6.2%)	Prin4 (5.2%)	Prin5 (3.1%)
Prin1 (14.2%)	1.0000	-0.0060	-0.0047	-0.0011	-0.0052
Prin2 (9.3%)	-0.0060	1.0000	-0.0020	-0.0005	-0.0022
Prin3 (6.2%)	-0.0047	-0.0020	1.0000	-0.0004	-0.0017
Prin4 (5.2%)	-0.0011	-0.0005	-0.0004	1.0000	-0.0004
Prin5 (3.1%)	-0.0052	-0.0022	-0.0017	-0.0004	1.0000

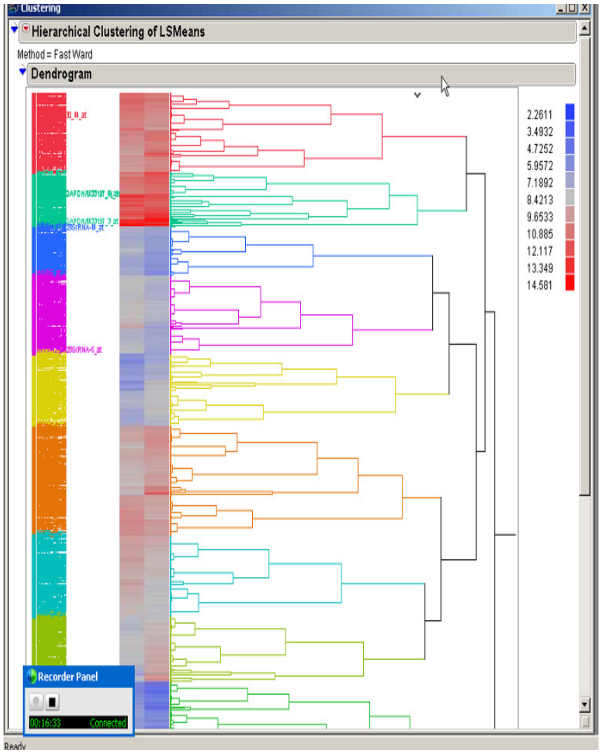
The correlations are estimated by REML method.

Figure 3. Correlation analysis of principle components



	Probe_Set_ID	Detection_Percentage	normal_1_1	adenocarcinoma_1_2	normal_2_3	adenocarcinoma_2_4	normal_3_5	adenocarcinoma_3_6	normal_4_7	adenocarcinoma_4_8	normal_5_9	adenocarcinoma_5_10
1	1007_s_at	0.980713	11.10156	11.39063	10.74805	11.45508	10.2207	11.83594	10.80664	11.9043	10.78125	11.40
2	1053_at	0.672974	6.334961	7.34668	7.185547	7.493164	7.135742	8.125	7.871094	7.615234	7.607422	7.7431
3	117_at	0.769165	7.459961	8.175781	7.686523	7.71582	8.101563	7.579102	6.9375	7.835938	7.941406	8.218
4	121_at	1	10.21094	9.833984	9.488281	9.75	9.904297	9.675781	9.712891	9.796875	9.734375	9.8359
5	1255_g_at	0.01923	2.599121	5.848633	5.291016	4.857422	5.089844	5.893555	4.332031	4.238281	5.336914	6.5244
6	1294_at	0.980713	10.09961	9.392578	10.28711	9.646484	9.650391	8.761719	10.13086	9.255859	9.931641	10.414
7	1316_at	0.672974	6.65332	6.335938	6.461914	6.442383	6.72168	6.320313	6.762695	6.769531	6.260742	6.8369
8	1320_at	0	6.869141	6.553711	6.848633	5.99707	7.15332	6.583984	6.6875	4.904297	6.603516	6.3710
9	1405_i_at	0.942261	10.08984	9.929688	10.79297	9.246094	9.917969	8.945313	11.01367	9.154297	10.58008	10.31
10	1431_at	0.230743	6.303711	5.472656	5.438477	6.25293	6.506836	6.112305	6.433594	6.006836	5.296875	5.4873
11	1438_at	0.096146	4.675781	8.027344	6.714844	7.722656	7.288086	7.951172	7.227539	8.648438	6.636719	7.5380
12	1487_at	0.75	9.175781	8.798828	9.363281	9.544922	8.929688	9.222656	8.84375	9.900391	8.964844	9.2890
13	1494_f_at	0.846069	8.070313	8.154297	7.677734	8.197266	7.838867	7.848633	7.939453	7.860352	8.181641	8.3203
14	1598_g_at	1	12.75977	11.64063	12.73633	11.23047	12.42578	11.16602	13.02734	11.02539	12.69531	11.736
15	160020_at	0.942261	8.367188	9.492188	8.492188	9.509766	9	8.916016	8.357422	9.337891	8.236328	9.2050
16	1729_at	0.922974	9.822266	9.519531	9.919922	10.22461	9.384766	9.310547	9.677734	9.458984	9.996094	10.652
17	1773_at	0.03846	6.50293	6.957031	5.104492	6.977539	7.352539	7.428711	5.630859	7.598633	6.172852	5.5634
18	177_at	0.269226	6.706055	6.868164	6.743164	7.097656	6.779297	6.916992	6.725586	6.862305	6.863281	6.4667
19	179_at	0.672974	10.06055	9.535156	10.30859	9.189453	10.13672	9.628906	10.7832	10.00391	9.072266	9.9902
20	1861_at	0.692261	8.371094	7.786133	8.087891	7.983398	8.011719	8.054688	8.238281	8.177734	8.568359	8.6269
21	200000_s_at	1	10.86914	10.67969	10.92773	10.73047	11.13477	10.62305	11.08984	10.69336	11.04102	9.9921
22	200001_at	1	11.875	12.08008	11.78125	11.8418	12.86133	12.66602	12.49609	11.92773	12.52148	11.3
23	200002_at	1	13.50586	12.97656	13.74805	13.01758	13.03125	13.03125	13.37305	14.00391	13.26758	13.292
24	200003_s_at	1	13.65625	13.73828	13.79297	13.41211	13.73242	13.87109	13.82031	14.26563	13.92383	14.384
25	200004_at	1	12.79688	12.7168	13.06445	12.91211	12.69531	13.07813	12.73242	12.73242	12.66406	12.781
26	200005_at	1	11.75391	11.41602	11.76953	11.57422	11.54297	11.16016	11.58984	12.07031	11.58594	11.828
27	200006_at	1	13.06641	12.80078	12.77148	13.1875	12.74219	12.68555	13.04688	12.82422	12.98633	13.707
28	200007_at	1	12.71875	12.58398	12.40234	12.23438	12.5625	12.55664	12.68945	12.51758	12.75195	12.080
29	200008_s_at	1	9.333984	10.37109	8.810547	9.652344	10.32227	10.18555	10.08984	10.94922	8.982422	8.0429
30	200009_at	1	12.50195	12.18555	12.50195	12.18164	12.4082	12.18359	12.56055	12.90234	12.27148	12.308
31	200010_at	1	13.58789	13.52148	13.47266	13.375	13.69531	13.0293	13.69336	13.85547	13.68555	13.275
32	200011_s_at	1	11.07227	11.40039	11.18555	11.25977	11.11523	11.25977	11.09961	11.16797	11.16602	10.708
33	200012_x_at	1	14.26758	13.5332	14.14648	13.09766	13.96289	13.52148	14.21875	14.19727	14.28125	14.111

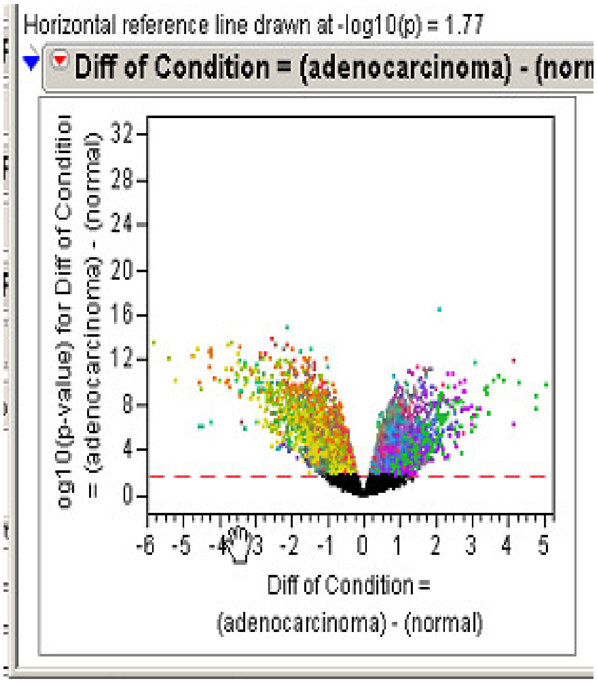
Figure 2. Adenocarcinoma Cancer Data



Hierarchical Clustering of LSMeans

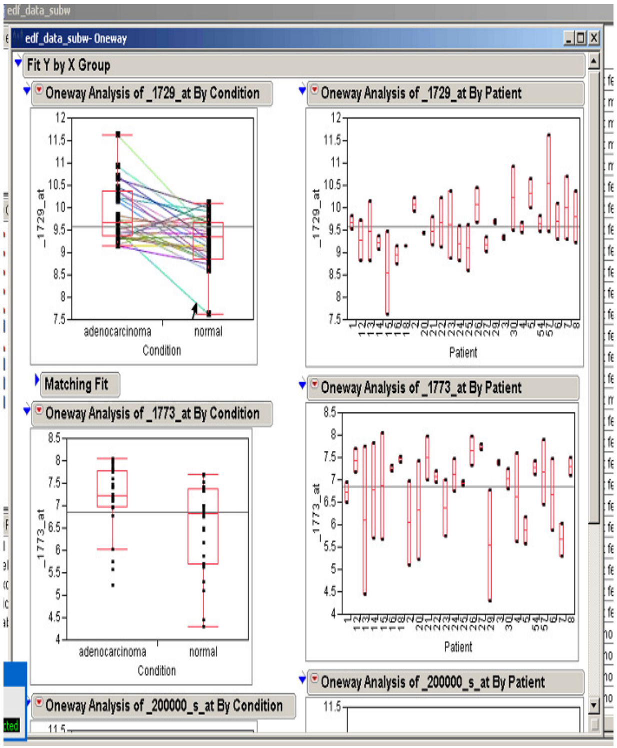
Method = Fast Ward

Figure 4. Dendograms of hierarchical clustering



Horizontal reference line drawn at -log10(p) = 1.77

Figure 5. Volcano plot



edf_data_subw
edf_data_subw- Oneway
Fit Y by X Group
Oneway Analysis of _1729_at By Condition
Oneway Analysis of _1729_at By Patient
_1729_at
adenocarcinoma
normal
Condition
Patient
Matching Fit
Oneway Analysis of _1773_at By Condition
Oneway Analysis of _1773_at By Patient
_1773_at
Oneway Analysis of _200000_s_at By Condition
Oneway Analysis of _200000_s_at By Patient

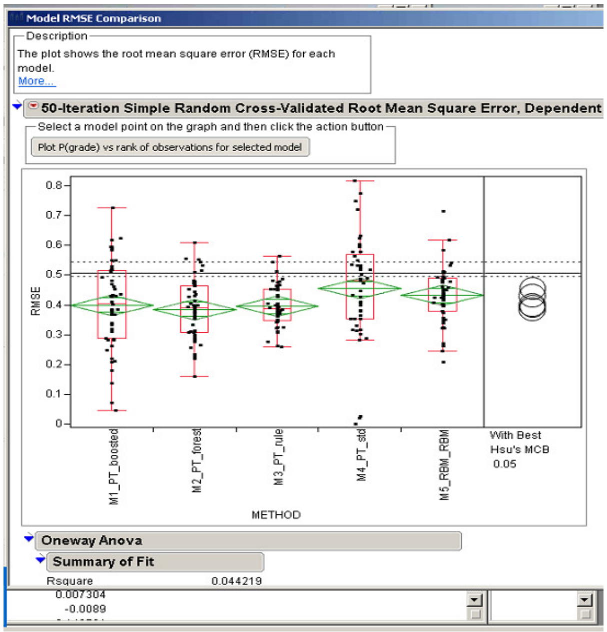
Figure 6. One-way analysis plots

Data Mining Performed Using Sas Jmp® Genomics For Breast Cancer Data

Box plots of a 50-iteration simple random cross-validation root mean square error (RMSE) are shown I Figure 7 for five different models. In this Figure 7, the dependent variables is "grade" for level of severity of cancer tumors in breast cancer, and the predictor continuous variables is "age". Cross validation was performed that on predictive model settings selected and compares the results.

Figure 8 shows the 235 predictors ranked for each of the models used as training set data. Figure 9 shows the Heat Map and Dendograms for breast cancer data which uses colors to indicate the intensity of correlation. The lower right corner of Figure 9 Heat Map is in red indicating highly correlated microarrays.

The frequency distributions are shown in Figure 10 that were obtained by highlighting the selected portion of Figure 9 Heat Map and indicate no grade 3 tumors. Partitioning the decision trees as shown in Figure 11 shows contingency analysis of predicted class by grade of tumor, and also the distribution data by true grade of tumor, actual probabilities, and correct predictions.



Model RMSE Comparison

Description

The plot shows the root mean square error (RMSE) for each model.
More...

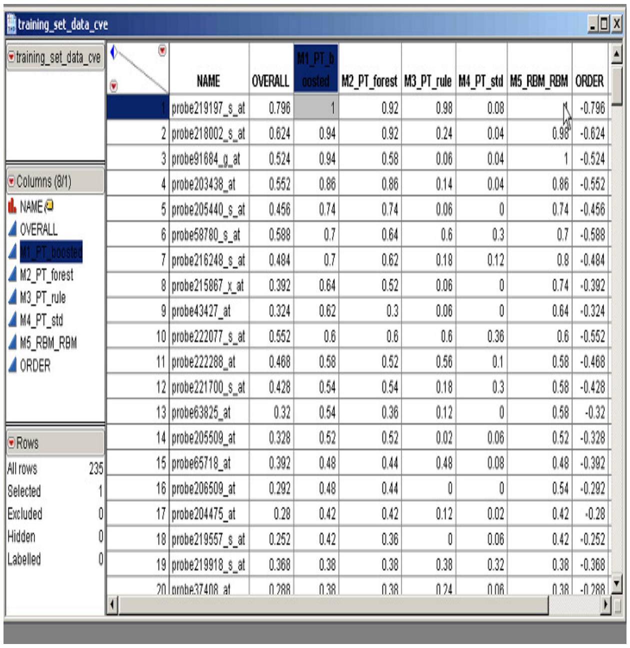
50-Iteration Simple Random Cross-Validated Root Mean Square Error, Dependent

Select a model point on the graph and then click the action button

Plot P(grade) vs rank of observations for selected model

Rsquare 0.044219
0.007304
-0.0089

Figure 7 Five different models



training_set_data_cve

training_set_data_cve

Columns (8/1)

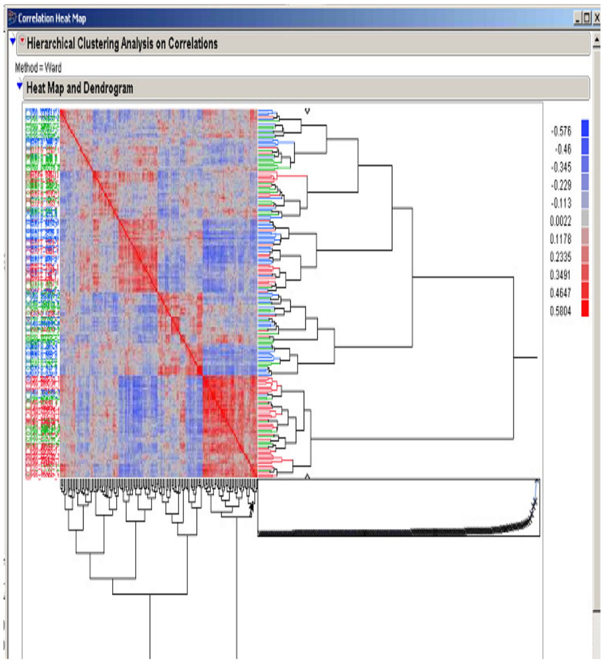
NAME
OVERALL
M1_PT_boosted
M2_PT_forest
M3_PT_rule
M4_PT_std
M5_RBM_RBM
ORDER

Rows

All rows	235
Selected	1
Excluded	0
Hidden	0
Labelled	0

	NAME	OVERALL	M1_PT_boosted	M2_PT_forest	M3_PT_rule	M4_PT_std	M5_RBM_RBM	ORDER
1	probe219197_s_at	0.796	1	0.92	0.98	0.08	1	-0.796
2	probe218002_s_at	0.624	0.94	0.92	0.24	0.04	0.98	-0.624
3	probe91684_g_at	0.524	0.94	0.58	0.06	0.04	1	-0.524
4	probe203438_at	0.552	0.86	0.86	0.14	0.04	0.86	-0.552
5	probe205440_s_at	0.456	0.74	0.74	0.06	0	0.74	-0.456
6	probe58780_s_at	0.588	0.7	0.64	0.6	0.3	0.7	-0.588
7	probe216248_s_at	0.484	0.7	0.62	0.18	0.12	0.8	-0.484
8	probe215867_x_at	0.392	0.64	0.52	0.06	0	0.74	-0.392
9	probe43427_at	0.324	0.62	0.3	0.06	0	0.64	-0.324
10	probe222077_s_at	0.552	0.6	0.6	0.6	0.36	0.6	-0.552
11	probe222288_at	0.468	0.58	0.52	0.56	0.1	0.58	-0.468
12	probe221700_s_at	0.428	0.54	0.54	0.18	0.3	0.58	-0.428
13	probe63825_at	0.32	0.54	0.36	0.12	0	0.58	-0.32
14	probe205509_at	0.328	0.52	0.52	0.02	0.06	0.52	-0.328
15	probe65718_at	0.392	0.48	0.44	0.48	0.08	0.48	-0.392
16	probe206509_at	0.292	0.48	0.44	0	0	0.54	-0.292
17	probe204475_at	0.28	0.42	0.42	0.12	0.02	0.42	-0.28
18	probe219557_s_at	0.252	0.42	0.36	0	0.06	0.42	-0.252
19	probe219918_s_at	0.368	0.38	0.38	0.38	0.32	0.38	-0.368
20	probe37408_at	0.288	0.38	0.38	0.24	0.06	0.38	-0.288

Figure 8 Training set data



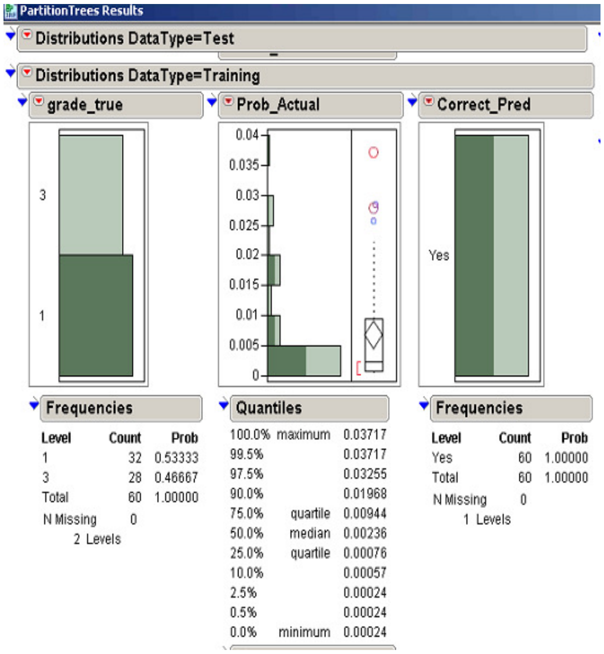
Correlation Heat Map

Hierarchical Clustering Analysis on Correlations

Method = Ward

Heat Map and Dendrogram

Figure 9 Heat Map



PartitionTrees Results

Distributions DataType=Test

Distributions DataType=Training

grade_true

Frequencies

Level	Count	Prob
1	32	0.53333
3	28	0.46667
Total	60	1.00000
N Missing	0	

2 Levels

Prob_Actual

Quantiles

100.0%	maximum	0.03717
99.5%		0.03717
97.5%		0.03255
90.0%		0.01968
75.0%	quartile	0.00944
50.0%	median	0.00236
25.0%	quartile	0.00076
10.0%		0.00057
2.5%		0.00024
0.5%		0.00024
0.0%	minimum	0.00024

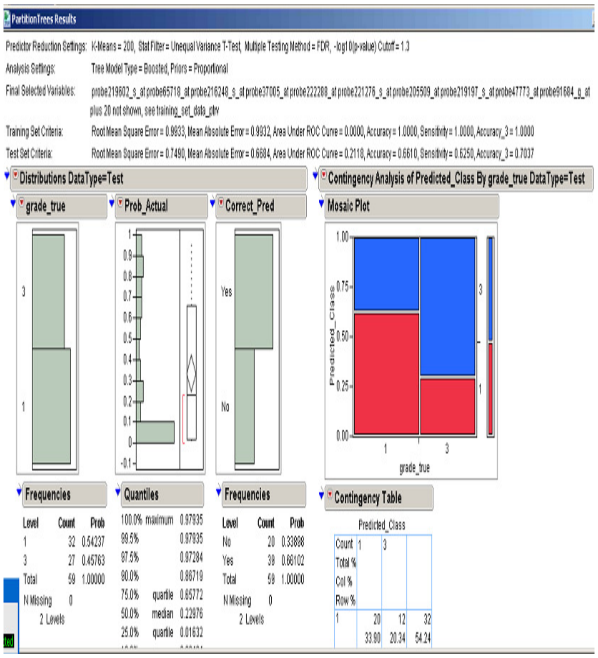
Correct_Pred

Frequencies

Level	Count	Prob
Yes	60	1.00000
Total	60	1.00000
N Missing	0	

1 Levels

Figure 10 Frequency distributions



PartitionTrees Results

Predictor Reduction Settings: K-Means = 200, Stat Filter = Unequal Variance T-Test, Multiple Testing Method = FDR, -log10(p-value) Cutoff = 1.3

Analysis Settings: Tree Model Type = Boosted, Priors = Proportional

Final Selected Variables: probe219602_s_at probe65718_at probe216248_s_at probe37005_at probe222288_at probe221276_s_at probe205509_at probe219197_s_at probe47773_at probe91684_g_at plus 20 not shown, see training_set_data_ptrv

Training Set Criteria: Root Mean Square Error = 0.9933, Mean Absolute Error = 0.9932, Area Under ROC Curve = 0.0000, Accuracy = 1.0000, Sensitivity = 1.0000, Accuracy_3 = 1.0000

Test Set Criteria: Root Mean Square Error = 0.7490, Mean Absolute Error = 0.6684, Area Under ROC Curve = 0.2118, Accuracy = 0.6610, Sensitivity = 0.6250, Accuracy_3 = 0.7037

Distributions DataType=Test

grade_true

Frequencies

Level	Count	Prob
1	32	0.54237
3	27	0.45763
Total	59	1.00000
N Missing	0	

2 Levels

Prob_Actual

Quantiles

100.0%	maximum	0.97935
99.5%		0.97935
97.5%		0.97284
90.0%		0.86719
75.0%	quartile	0.65772
50.0%	median	0.22976
25.0%	quartile	0.01632

Correct_Pred

Frequencies

Level	Count	Prob
No	20	0.33898
Yes	39	0.66102
Total	59	1.00000
N Missing	0	

2 Levels

Contingency Analysis of Predicted_Class By grade_true DataType=Test

Mosaic Plot

Contingency Table

Predicted_Class

Count Total % Col % Row %	1	3	
1	20 33.90	12 20.34	32 54.24

Figure 11 Partitioning the decision trees

4. CONCLUSIONS AND SUMMARY

This paper emphasizes the usefulness of SAS JMP® Genomics with supercomputing and data mining. This research illustrates genetic visualization for the analysis and modeling of microarray databases for both lung and breast cancer as a tool for better understanding of the consequences of these diseases and for potential strategies for their treatments

5. ACKNOWLEDGEMENTS

The first two authors* would like to acknowledge funding to support this research from a 2009 Summer Faculty Research Grant as awarded by the College of Business at Arkansas State University, and also the technical support of SAS Inc. for that provided for their software of JMP® Genomics.

6. REFERENCES

1. Image Spatial Data Analysis Group (2009), National Center for Supercomputing, University of Illinois at Urbana-Champaign, http://isda.ncsa.illinois.edu

2. NCBI (2007), "Expression data from Lung Cancer", Gene Expression Omnibus (GEO), Series GSE7670, **National Center for Biotechnology Information**, http://www.ncbi.nlm.nih.gov/geo/

3. NCBI (2006), "Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade to Improve Prognosis, Gene Expression Omnibus (GEO), Series GSE2990, **National Center for Biotechnology Information**, http://www.ncbi.nlm.nih.gov/geo/

4. SAS (2009), **JMP® Genomics 4.0 Product Brief**, http://www.jmp.com/software/genomics/pdf/103112_jmpg4_prodbrief.pdf

5. Segall, Richard S., Zhang, Qingyu and Pierce, Ryan M.(2009), "Visualization by Supercomputing Data Mining", **Proceedings of the 4th INFORMS Workshop on Data Mining and System Informatics**, San Diego, CA, October 10, 2009.

6. Segall, Richard S., Zhang, Qingyu and Pierce, Ryan M.(2010), "Data Mining Supercomputing with SAS JMP® Genomics: Research-in-Progress" submitted to **Proceedings of 2010 Conference on Applied Research in Information Technology**, sponsored by Acxiom Laboratory of Applied Research (ALAR),University of Central Arkansas (UCA), Conway, AR, April 9, 2010.

7. Sotiriou C, Wirapati P, Loi S, Harris A et al. (2006), Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. **J National Cancer Institute**, February 15; volume 98, number 4, pp. 262-72. PMID: 16478745

8. Su LJ, Chang CW, Wu YC, Chen KC et al.(2007), "Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme", **BMC Genomics** 2007 June 1; volume 8, number 140. PMID: 17540040

9. Thoennes, M.S., and Weems, C.C. (2003), "Exploration of the performance of a data mining application via hardware based monitoring", **The Journal of Supercomputing**, volume 26, pp. 25-42.

10. Wilkins-Diehr, N. and Mirman, I. (2008), "On-Demand supercomputing for emergencies", **Design Engineering Technology News Magazine**, February 1, http://www.deskeng.com/articles/aaagtk.htm

11. Zaki, M.J., Ogihara, M., Parthasararthy, S., and Li, W. (1996), "Parallel data mining for association rules on shared-memory multi-processors", **Proceedings of the 1996 ACM/IEEE Conference on Supercomputing**, Pittsburgh, PA, Article Number 43, ISBN 0-89791-854-1.