Statistical Quality Control of Microarray Gene Expression Data

Shen LU
University of Arkansas at Little Rock, Department of Computer Science,
Donaghey College of Engineering and Information Technology, Little Rock, AR 72204 USA

and

Richard S. SEGALL
Arkansas State University, Department of Computer & Information Technology,
College of Business, State University, AR 72467-0130 USA

ABSTRACT

This paper is about how to control the quality of microarray expression data. Since gene-expression microarrays have become almost as widely used as measurement tools in biological research, we survey microarray experimental data to see possibilities and problems to control microarray expression data. We use both variable measure and attribute measure to visualize microarray expression data. According to the attribute data's structure, we use control charts to visualize fold change and t-test attributes in order to find the root causes. Then, we build data mining prediction models to evaluate the output. According to the accuracy of the prediction model, we can prove control charts can effectively visualize root causes.

Keywords: statistical quality control, microarray, information product, control charts, root causes

1. INTRODUCTION

In microarray experiments, we compare patterns of expression across multiple samples hybridized to a particular array in order to discover new genes that, as drug targets, are responsible for a disease. For gene discovery and prediction, the quality of the data set usually affects the result of the experiment.

Microarray technology has found its applications in recent years in many fields of life science. Generally speaking, all the data analysis behind these applications can be characterized into two major categories: discovery and prediction. Discovery is to discover new knowledge, new genes involved in a pathway; prediction is to create predictive models to be used in such areas as toxicology and disease diagnosis. Fundamental to both discovery and prediction is the selection of genes that are differentially expressed (up or down) when comparing the samples of your interest to the control group.

Lu and Segall [11, 12] present preliminary research on medical record linkage and entity resolution methods as applied to bioinformatics. Segall [14] presented a chapter on data mining of microarray databases for biotechnology. Segall [14] performed data visualization and data mining of microarray databases for continuous numerical-valued Abalone fish data and discrete nominal-valued mushroom data using evolutionary algorithms specifically for neural networks and genetic algorithms. Segall [15] performed data mining of microarray databases for human lung cancer. Segall [16], Segall [17], and Segall [18] performed data mining of microarray databases of Leukemia cells using single SOM. This paper extends the methodology used in authors' previous research for microarray data analysis to those using statistical quality control techniques.

2. PHASE I: MEASURING THE QUALITY ATTRIBUTES

For two-color microarray experiments, as shown in Figure 1, one must decide what the most appropriate comparison is to be made with each array hybridization. The simplest comparisons can be separated into four general classes, such as direct comparison, reference design, balanced block design and loop design. In many ways, direct comparisons are the simplest conceptually; they are used when two distinct classes of experimental samples are to be compared, such as a treated sample and its untreated control. On each array, representatives of the two classes are paired and co-hybridized together such that the relative expression levels are measured directly on each array. The choice of appropriate pairing depends on the experimental question under study. For example, one can pair diseased and normal tissue from the same patient or randomly select animals from mutual and wild-type groups. The strategy to collect data for any given case is influenced by a wide range of factors, including the availability of samples, the quantity of RNA that can be obtained, the size of the study, and the logistical constraints in the laboratory.

For each gene, the process begins with defining an expression vector that represents its location in expression space. In this view of gene expression, each hybridization represents a separate distinct axis in space, and the log2(ratio) measured for that gene in that particular hybridization represents its geometric coordinate. In this way, expression data can be represented in m-dimensional expression space, where m is the number of hybridizations and where each gene expression vector is represented as a single point in that space. It should be noted that one could use a similar approach to representing each hybridization assay using a sample vector consisting of the expression

values for each gene; these define a sample space whose dimension is equal to the number of genes assayed in each array.

We collect Microarray experimental data and to see possibilities and problems about whether the data are sufficient and can be used to generate, evaluate, and improve the cancer-related prediction model and about whether the data can be used to select the proper pre-processing and modeling techniques. Several different data sets are considered. According to Babu [1], a microarray is typically a glass slide onto which DNA molecules are placed as spots. A microarray may contain thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene as shown in part A of Figure 1 from Babu [1].

For liver cancer [3], there are 17,400 genes and 179 samples, for lung cancer [2], there are 12,600 genes and 245 samples, for NIH cancer dataset [3], 12,196 genes and 240 samples, for prostate cancer [10], there are 26,260 genes and 103 samples. Most of Microarray data have a small size of samples in which the number of genes is large. Obviously, in comparison with the number of genes, we can make such a conclusion that most of Microarray experiments can neither supply enough samples to do statistical analysis, nor generate a prediction model. A wide range of methods for microarray data analysis have evolved, ranging from simple fold-change approaches to testing for differential expression, to many computationally demanding and complex techniques. In this paper, in order to control the quality of Microarray experimental data, we generate such a process that we collect Microarray experimental data, check the quality of the data, remove noise, and build a prediction model to evaluate the output.

Other than expression ratio value, we also use fold change and t-test as attribute measures. Fold change and t-test can be used to identify best distinguish genes between the sample classes.

The student's t-test can be used to test whether a difference is significant, which is an assessment of signal-to-noise ratio for the particular gene in question.

$$t = (\text{signal} / \text{noise}) =$$
$$(\text{difference between groups} / \text{variability of groups})$$

A large value for the t statistics indicates that the populations representing measurements of a gene for condition A and B are well separated. It can be used to estimate how likely that a gene is differentially expressed between conditions.

Fold change is a mathematical operation describing how much two variables differs. It is the ratio of the final value and initial value (B/A), if the final value is larger.

For t-test, a *p*-value is normally calculated to quantify the significance. And the most common interpretation for a *p*-value of 0.05 is that there is a 5% probability that the observed difference in expression may simply due to chance, not independent. Calculating a fold change is straightforward, although one does have to decide which of the three methods to use to calculate an average (arithmetic, geometric, harmonic).

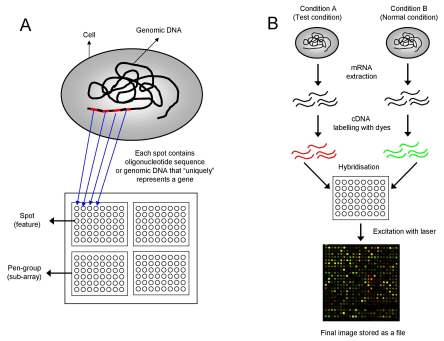

Figure 1. Illustration of a microarray that may contain thousands of "spots" of genomic data [1]

3. PHASE II: ANALYZE THE INFORMATION PRODUCT

Statistical process for microarray expression data includes the following steps:

1. Pre-processing: because of experimental errors, some values of expression data are missing. We use K Nearest Neighbor (KNN) algorithm to automatically impute missing values first.
2. Sample selection: since microarray expression data set is not very big, we can use total data for any experiments and applications. However, regarding to the different number of treated samples and untreated samples, we randomly generate data sets in which both treated and untreated classes have the same number of samples.
3. Feature selection: even if data mining analysis can be performed, it is still extremely useful to reduce the data set to those genes that are best distinguished between the sample classes.

Before statistical analysis of microarray expression data, we have to decide the attribute data which can be used to analyze root causes. For an information system, there are two types of data: one is variable data; the other is attribute data. Variable data can be measured, primarily continuous in nature. Attribute data are observed to be either present or absent, conforming or non-conforming. The effectiveness of charts depends on the attribute data's structure. The problem is how to choose attribute data. The critical is if the categories of non-conforming are sufficiently focused, so that there is likely to be only one assignable cause per category. For microarray expression data, since the difference and independence of difference samples are significant, we use fold change and t-test to measure the quality of the sample. For microarray expression data, since it belongs to binomial distribution and each gene is tested in the same number of experiments, we use np chart to visualize the fold change and pValue.

Our work is based on liver cancer [3] data. We choose total data and balanced data to generate sample sets. For fold

change equal to 1.5 or 2.0, and pValue equal to 0.05 and 0.01, we generate different data sets, as shown in Table 1. For each data set, we use np chart and moving range chart to visualize fold change and pValue, as shown in Figure 2 and Figure 3.

	Fold change	pValue
Total data	1.5	0.01
	1.5	0.05
	2.0	0.01
	2.0	0.05
Balanced data	1.5	0.01
	1.5	0.05
	2.0	0.01
	2.0	0.05

Table 1. Data Sets with Difference Values of Fold Change and pValue

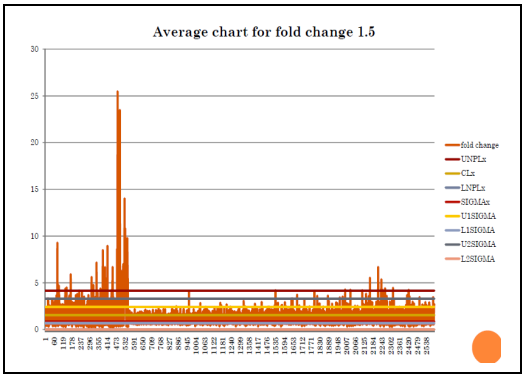

Figure 2. NP Chart for Total Data with Fold Change Set to 1.5

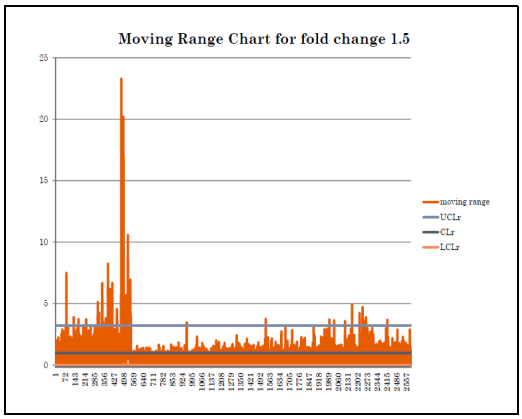

Figure 3. Range Chart for Random Data

In Total Quality Management (TQM), an organization would follow certain guidelines to scope an IQ project, identify critical issues, and develop procedures and metrics for continuous analysis and improvement [19]. Control charts provide us an easy way to compare the observed subgroup averages and subgroup ranges against the predicted limits. As shown in Figures 2, 3, we use upper and low control limits and 1 sigma and 2 sigma control limits to analyze the root cause. A sigma unit is a measure of scale for the data. Roughly 60% to 75% of the data will be located within a distance of one sigma unit on either side of the average. Usually 90% to 98% of the data will be located within a distance of two sigma units on either side of the average. Approximately 99% to 100% of the data will be located within a distance of three sigma units on either side of the average. Figures 4(a)-4(d) show us the relationship between empirical rules and control charts. Any point outside of the upper and lower control limits is a clear example of a special cause variation. The other forms of special cause variation are called runs. Trends are special forms of a run. According to the average chart, we can see some data are out of upper and lower 2 sigma limits, and even out of upper and lower control limits, they are definitely out of control. That means those observations are not consistent with predictions and can make the process unstable. Since the average chart and range chart are for root cause analysis, out of range data are special causes of variation, we need to take actions to identify and remove them.

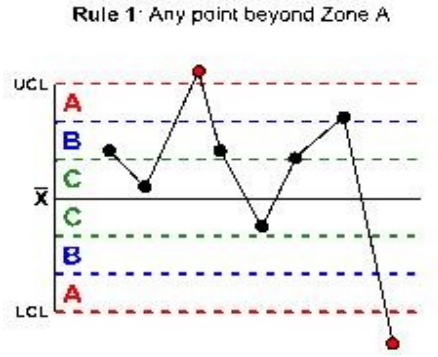

Figure 4(a). Interpretation of Western Electronic Rule 1 [20]

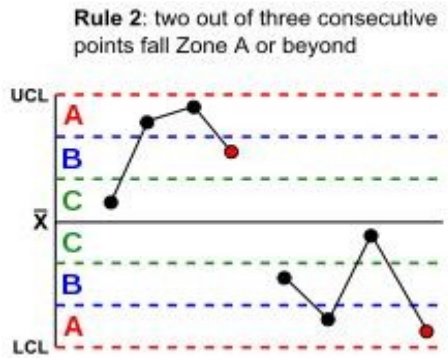

Figure 4(b). Interpretation of Western Electronic Rule 2 [20]

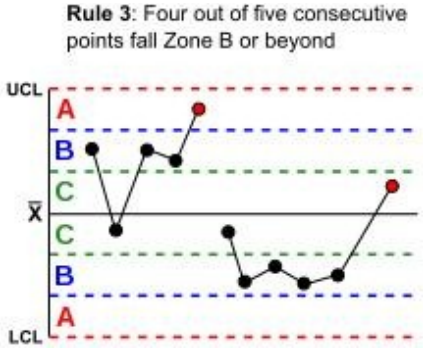

Figure 4(c). Interpretation of Western Electronic Rule 3 [20]

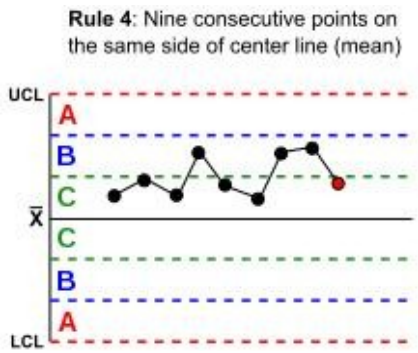

Figure 4(d). Interpretation of Western Electronic Rule 4 [20]

Table 2 lists the number of genes and the percentage of deletion by using fold change and t-test feature selection. In next section, we use prediction model to evaluate the output after removing special causes.

	dataset	sample	genes	removing pe	total orde	sub orde
1	total	179	19536			
2	total_pvlaue005	179	6660	0.659090909	24	8
3	total_pvlaue001	179	4383	0.775644963	18	6
4	total_fc20	179	2717	0.860923423	12	4
5	total_fc20_pvalue005	179	1270	0.93499181	6	2
6	total_fc20_pvalue001	179	772	0.96048321	3	1
7	total_fc15	179	5795	0.703368141	19	7
8	total_fc15_pvalue005	179	2768	0.858312858	14	5
9	total_fc15_pvalue001	179	2181	0.888359951	9	3
10	balance3	50	19536			
11	balance3_pvlaue005	50	5864	0.699836	21	7
12	balance3_pvlaue001	50	3939	0.798372236	16	6
13	balance3_fc20	50	2785	0.85744267	15	5
14	balance3_fc20_pvalue005	50	886	0.95464783	4	2
15	balance3_fc20_pvalue001	50	743	0.961967649	2	1
16	balance3_fc15	50	5900	0.697993448	22	8
17	balance3_fc15_pvalue005	50	2403	0.876996314	10	4
18	balance3_fc15_pvalue001	50	2135	0.890714578	7	3
19	balance2	50	19536			
20	balance2_pvlaue005	50	5948	0.695536446	23	8
21	balance2_pvlaue001	50	4006	0.79494267	17	6
22	balance2_fc20	50	2739	0.859797297	13	5
23	balance2_fc20_pvalue005	50	898	0.954033579	5	2
24	balance2_fc20_pvalue001	50	739	0.9621724	1	1
25	balance2_fc15	50	5851	0.700501638	20	7
26	balance2_fc15_pvalue005	50	2422	0.876023791	11	4
27	balance2_fc15_pvalue001	50	2148	0.890049144	8	3

Table 2. The Number of Genes Present after Sample Selection and Feature Selection

4. PHASE III: IMPROVE THE INFORMATION PRODUCT

After generating different data sets, as the output of the process, we use data mining analysis to evaluate them. In TQM, knowledge has been created for Information Quality (IQ) practice [5, 6]. Precision model building includes two steps: model building and model validation. Model building involves in training data selection. Model validation involves in testing the built model with testing samples and measuring the precision and recall of the output of the generated model.

We use K-Nearest Neighbor (KNN) [4], Random Forest (RF) [7], Multipass-LVQ (MPL) [8], and Self-Organizing Map (SOM) [9] algorithms to calculate the precision and recall on different data sets. KNN is based on the direct comparison of the distance between two neighbors. This algorithm is good for high dimensional vectors. Random Forest is based on decision tree theory. Since the best features are selected to build decision trees, the significance of different features are considered in this algorithm. Multipass-LVQ and SOM belong to neural network algorithm. Since samples can be randomly selected as input for many times, these algorithms are good for high-dimensional small size data sets, such as microarray expression data. Precision and recall of these algorithms on different data sets are shown in Figured 5(a) to 5(g) below..

In the below Figures 5(a)-5(e) and 5(g), KNN is used for #1-4, RF for #5-8, SOM for #9-12, and MPL for #13-16. In Figures 5(f) and 5(h), SOM is used for #9-16, and MPL for #17-20.

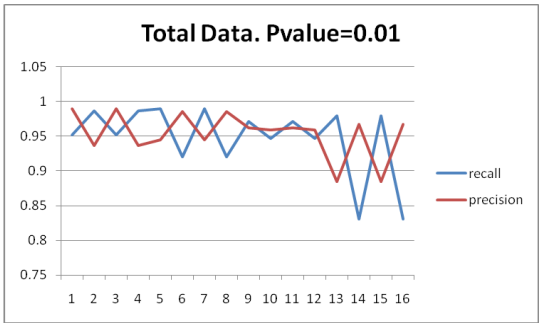

Figure 5(a)

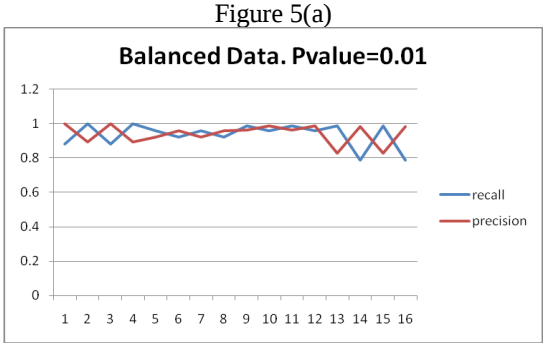

Figure 5(b)

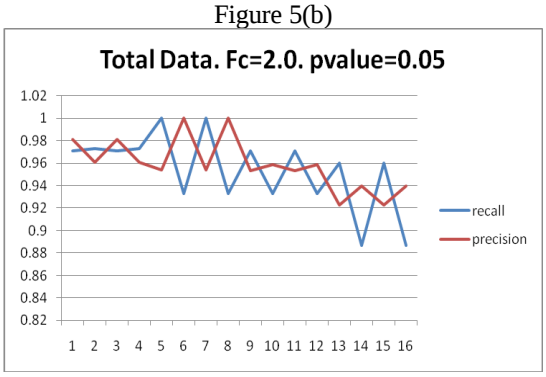

Figure 5(c)

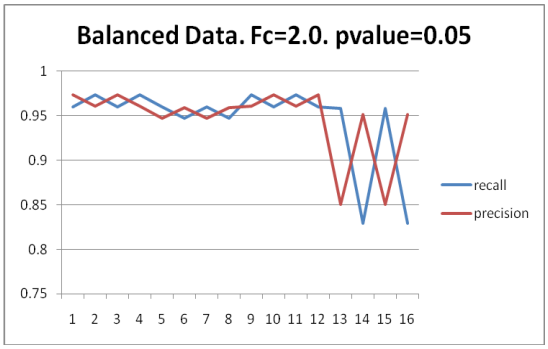

Figure 5(d)

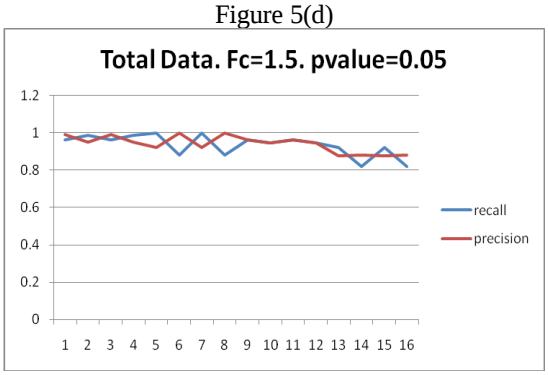

Figure 5(e)

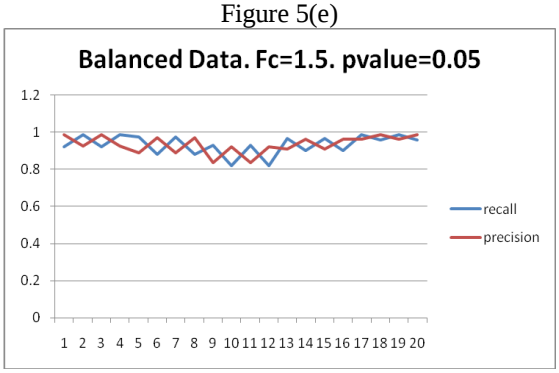

Figure 5(f)

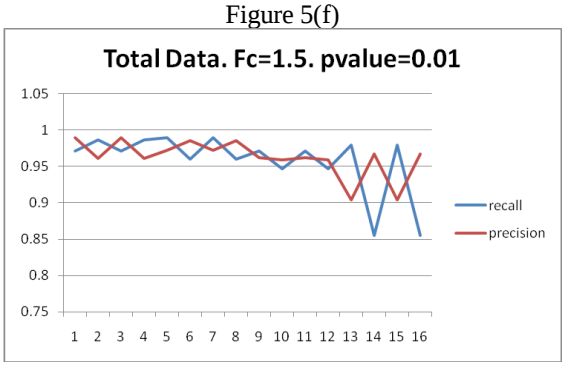

Figure 5(g)

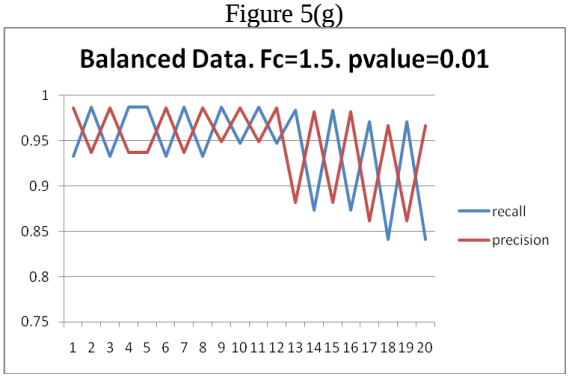

Figure 5(h)

5. CONCLUSIONS

This paper introduced the significance of data quality control in microarray experiments. According to different microarray comparisons, we collected data in different ways. A formal method was given to measure the possibility and problems about whether data are sufficient and can be used to generate, evaluate, and improve prediction model. We used T-test and fold change to select samples and genes, and used control chars to visualize the quality of the output. Four data mining algorithms, such as KNN, SOM, Random Forest, Multipass-LVQ, were used to build prediction models and to evaluate the quality of the data. The performance of the output showed us control charts are useful for the visualization of the root cause variation of the data. Selection of appropriate charts to visualize the output is very important for data quality control. Empirical root cause rules and analysis can be used to explain control charts and ensure control charts will yield very few false conclusions.

6. REFERENCES

[1] M. M. Babu, "An Introduction to Microarray Data Analysis", Chapter 11 in **Computational Genomics** by R. Grant, Editor, Horizon Press, UK, 2004, pp. 225-249, http://www.mrclmb.cam.ac.uk/genomes/madanm/microarray/chapter-final.pdf

[2] L. Breiman, "Random Forests", **Machine Learning** Volume 45, Number 1, 2001, pp. 5-32.

[3] X. Chen, S. T. Cheung, S. So, S. T. Fan, C. Barry, J. Higgins, K. M. Lai, J. Ji, S. Dudoit, I.O. Ng, et al., "Gene Expression Patterns in Human Liver Cancers", **Molecular Biology of the Cell**. Vol. 13, Issue 6, pp. 1929-1939, June 2002.

[4] T . M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification." **IEEE Transactions of Information Theory**, Volume IT-13, Number 1, 1967, pp. 21-27.

[5] P. Cykana, A. Paul, and M. Stern, "DoD Guidelines on Data Quality Management", **Proceedings of the 1996 Conference on Information Quality**, Cambridge, Mass., 1996, pp. 154–171.

[6] C.P. Firth and R.Y. Wang, **Data Quality Systems: Evaluation and Implementation**, Cambridge Market Intelligence Ltd., London, UK 1996.

[7] G. J. Gordon et al., "Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma", **Cancer Research**,Vol. 62, No. 17, pp. 4963–4967, September 2002.

[8] T. K. Kohonen, J. Kangas, J. Laaksonen, and K. Tokkola, **LVQ-PAK**: The Learning Vector Quantization Program Package, Helsinki University of Technology,

Finland, 1992.

[9] T. Kohonen, "The Self-Organizing Map", **Proceedings of the IEEE**, Volume 78, Issue. 9, 1990, pp. 1464-1480.

[10] J. Lapointe, C. Li, M. van de Rijn, J. P. Huggins, E. Bair, et al., "Gene Expression Profiling Identifies Clinically Relevant Subtypes of Prostate Cancer", **Proceeding of the National Academy of Sciences.** Vol. 101, No.3, pp. 3811-816, January 2002.

[11] S. Lu and R. S. Segall and S. Atiff, "Using Entity Resolution to Discover Scientific Numbers," Poster Abstract in **Proceedings of 15th International Conference on Information Quality (ICIQ-2010)**, University of Arkansas at Little Rock (UALR), Little Rock, AR, November 10-12, 2010.

[12] S. Lu and R. S. Segall, "Linkage in Medical Records and Bioinformatics Data", submitted for publication to **Proceedings of 2011 MidSouth Computational Biology and Bioinfomatics Society (MCBIOS), Special Issue of MBC Bioinformatics**, College Station, TX, April 1-2, 2011.

[13] A. Rosenwald, G. Wright, W. C. Chan, et al., "The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-cell Lymphoma", **The New England Journal of Medicine**. Vol. 346, pp. 1937-1947, June 2002.

[14] R. S. Segall, "Data Mining of Microarray Databases for Biotechnology," **Encyclopedia of Data Warehousing and Mining**, Edited by John Wang, Montclair State University, USA; Idea Group Inc., 2006, ISBN 1-59140-557-2, pp.734-739.

[15] R. S. Segall, and Q. Zhang, "Data Mining of Microarray Databases for Human Lung Cancer", **Proceedings of the Thirty-eighth Annual Conference of the Southwest Decision Sciences Institute**, vol. 38, no. 1, March 15-17, 2007, San Diego, CA.

[16] R. S. Segall and R. M. Pierce, "Data Mining of Leukemia Cells using Self-Organized Maps," **Proceedings of 2009 Conference on Applied Research in Information Technology**, sponsored by Acxiom Laboratory of Applied Research (ALAR), University of Central Arkansas (UCA), Conway, AR, February 13, 2009, pp. 92-98.

[17] R. S. Segall and R. M. Pierce, "Advanced Data Mining of Leukemia Cell Micro-Arrays", **Proceedings of 13th World Multi-Conference on Systemics, Cybernetics and Informatics: WMSCI 2009**, Orlando, FL, July 10-13, 2009.

[18] R. S. Segall and R. M. Pierce, "Advanced Data Mining of Leukemia Cell Micro-Arrays", **Journal of Systemics, Cybernetics and Informatics** (JSCI), Volume 7, Number 6, 2009, pp.60-66.

[19] R. Wang, "A Product Perspective on Total Data Quality Control", **Communications of the ACM**, Vol. 41, No. 2, 1998, pp. 58-66.

[20] Wikipedia, Western Electric Rules, 2011, http://en.wikipedia.org/wiki/Western_Electric_rules