Data Mining of Microarray Databases for the Analysis of Environmental Factors on Plants Using Cluster Analysis and Predictive Regression

Richard S. Segall

Arkansas State University, Department of Computer & Information Technology
College of Business, State University, AR 72467-0130
(870)972-3989 (Phone)   870-910-8187 (Fax)
rsegall@astate.edu

ABSTRACT

This paper is a continuation of previous research of author as appeared in Segall et al. (2003, 2004a, 2004b) pertaining to the applications of data mining of microarray databases of representative ingredients that could be used for the manufacture of Plant-Made-Pharmaceuticals (PMPs). This paper extends the previous research by first providing an extensive review of current research on data mining of microarray databases by other researchers, as well as more precisely discussing what a microarray database is. This paper uses SAS Enterprise Miner as in the previous research, but instead uses the prediction regression module and advanced cluster analysis techniques for the analysis of the data mining of the microarrays of plant data pertaining to environmental factors such as drought, chilling, and salinity. Conclusions and future directions of the research are presented.

BACKGROUND

What Is A "Microarray"?

A microarray is a huge collection of spots that contain massive amounts of compressed data. Researchers in the life sciences for genetics use microarrays because DNA (Deoxyribonucleic Acid) contains so much information on a micro-scale. Each spot of a microarray thus could contain a unique DNA sequence.

Figure 1 of this paper, cited from Kennedy and Wilson (2004) illustrates the data capture process for construction of microarrays. According to Kennedy and Wilson (2004) "the original and most common form of microarray technology today is glass slide microarrays, also known as two-color or two-channel microarrays. Using a robotic spotter, DNA fragments can be printed onto a glass slide in a defined compact design that can contain up to 80,000 spots." The image scanned by lasers is then quantified according to the color generated by each individual spot and the results organized as a text file that can be then subjected to analyses such as data mining.

According to Kennedy and Wilson (2004), a wide variety of microarray analysis opportunities now exist for plant functional genomics such as stress responses, discovering genes and gene functions, comparative genomics, and regulatory networks.

Previous Related Research On Microarray Databases By Author

The reader is referred to previous research of the author in Segall et al. (2003, 2004a, 2004b) on the application of some cluster analysis data mining techniques to microarray databases specifically for the analysis of environmental stresses on plants. This paper extends this previous research of the author by providing the results of new analysis of some of the same microarray databases as well as additional other new ones from the Osmotic Stress Microarray Information Database (OSMID) databases by using data mining with predictive regression and other different cluster analyses techniques.

As indicated above, the plant data used in this paper utilize both same and different microarray databases from the same data warehouse of Osmotic Stress Microarray Information Database (OSMID) that were studied in the previous research of Segall et al. (2003, 2004a, 2004b). In the latter it was discussed that the microarray databases in the OSMID data warehouse are considered to be representative of those that could be used for biotech application such as the manufacture of plant-made-pharmaceuticals (PMP) and genetically modified (GM) foods.

As indicated in Segall et al. (2003, 2004a, 2004b) the OSMID data warehouse contains the results of approximately 100 microarray experiments performed at the University of Arizona as part of a National Science Foundation (NSF) funded project named the "The Functional Genomics of Plant Stress".

As discussed in Segall et al. (2003, 2004a, 2004b), the OSMID database is available for public access on the web, and the OSMID contains information about the more than 20,000 ESTs (Experimental Stress Tolerances) that were used to produce these arrays. These 20,000 ESTs could be considered as components of data warehouse of plant microarray databases that could be subjected to data mining. The plants represented in the OSMID database include rice, barley, maize or corn, ice plant and Arabidopsis. Specifically, the OSMID microarray database contains 4,000 ESTs for maize or corn, ice plant and rice, and 2,000 ESTs for barley, and 9,000 ESTs for Arabidopsis.

Segall et al. (2003, 2004a, 2004b) also referred to the fact that the web page of the OSMID, the Stress Genomics Consortium as funded by NSF (National Science Foundation) utilizes a variety of techniques to investigate the responses of plants and certain microbes to environmental factors of stress such as drought, chilling, and salinity. Hence the data provided in the OSMID database that could be used in the data mining include that for the variables of treatments with environmental factors of salt, cold, and drought.

Previous Research On Data Mining Of Microarray Databases By Other Authors

Other researchers have recently performed data mining of plant data. Bluck and Menting (2003) discussed that NASA (National Aeronautical and Space Administration) utilized data mining on satellite and other global scientific data to reach conclusions about looking for ecological disasters by watching for changes in the amount of absorption of sunlight by green plants.
Bartley and Ishida (2002) performed data mining in the TIGR Tomato Gene Index and concluded that mining the TIGR Tomato EST (Expressed Sequence Tags) is considerably useful

by helping to reveal new directions for determining the development process involved in tomato fruit ripening. Torto et al. (2000) studied EST (Expressed Sequence Tags) data mining for Oomycete plants, and Zimback et al. (2004) for Eucalyptus plants.

Rosenquist et al. (2001) performed data mining of the plant Arabidopsis Genome that revealed fifteen 14-3-3 genes which are the “key regulators of primary metabolism and membrane transport.”

The Alberta Research Council (2000) of Canada utilized data mining in data from Alberta Agriculture that included applications such as studying the ability of tracing cattle from farm to consumer.

Spiliopoulou (2004) presented data mining for business applications that included that for Iris plant database including factor of petal width. Hsu and Joehanes (2003) discussed graphical models of gene regulation from microarray data and proposed “a synthesis of traditional clustering techniques with those for learning graphical models from data.”

Wotawa and Kinston (2003) compiled a data dictionary of the definitions and cell values of the data tables displayed on the web site of the National Park Service (NPS) specific to “Data Mining to Populate NPSpecies.” Wotawa and Kinston (2003) lists data for vascular plants at several universities including Arkansas State University, and other organizations such as Brooklyn Botanic Garden. This website of the National Park Service (2003) also contains web pages on “NPSpecies Data Mining and Geo-Reference Tools”.

Bio1inform newsletter by GenomeWeb LLC (2003) discusses the embedded data-mining tools of ORACLE 9i software claiming “ORACLE hopes to strike bioinformatics gold” with their introduction. ORACLE 9i Data Mining includes such features as naïve Bayes, decision trees, association rules, hierarchical k-means clustering, and O-cluster. ORACLE 9i Data Mining is in competition to other database-embedded data mining tools such as IBM’s DB2 Intelligent Miner and Teradata Warehouse Miner.

Shah (2002) presented data mining of GeneChips® in which he discusses its technology and cDNA microarrays that are scanned using laser microscopy.

Abdullah et al. (2003) have also recognized data mining of plants as being used in recent studies by agricultural researchers in Pakistan. In this study by Abdullah et al. (2003) it was shown “how data mining integrated agricultural data including pest scouting, pesticide usage and meteorological recordings is useful for optimization (and reduction) of pesticide usage.”

Uppsala Monitoring Centre (2004) used data mining tools including Bayesian theory and neural networks for World Health Organization (WHO) databases to signal adverse drug reactions. Embrechts (2002) presented a 3-day Bioinformatics Workshop funded by NSF (National Science Foundation) at Rensselaer Polytechnic Institute on visual explorations in clustering and data mining that included datasets from amino acids and gene expression arrays for Leukemia and preparing microarrays with StripMiner^{TM}.

Several textbooks have recently been written in this revolutionary of bionformatics and analysis of microarrays. Some of these that are highly recommended are Baldi et al. (2003) DNA Microarray and Gene Expression: From Experiments to Data Analysis and Modeling, Bergeron (2003) Bioinformatics Computing, Berrar et al. (2003) A Practical Approach to Microaray Data Analysis, Causton et al. (2003) Microarray Gene Expression Data Analysis: A Beginner's Guide, Claverie and Notredame (2003) Bioinformatics for Dummies, Hardiman (2003) Microarray Methods and Applications: Nuts & Bolts, Lesk (2003) Introduction to Bioinformatics, Schena (2003) Microarray Analysis, and Stekel (2004) Microarray Bioinformatics, .

REVIEW OF DESCRIPTION OF PLANT DATABASES USED FOR DATA MINING

This section provides a brief review of the description of the plant databases used for data mining in the previous research as in Segall (2003, 2004a, 2004b), and will be again utilized in the new research presented later in this paper. The microarray database of corn is used as representative of those ingredients used for the manufacture of plant-made pharmaceuticals (PMPs). The selection of corn as the crop to be examined is based on the following three observations from the current state of the plant biotech industry (Monsanto (2002)):

1. corn in the US is one of the most researched products in the food and feed system, and its genetic as well as agronomic properties are well documented;
2. corn is a safe and stable medium for genetic expression;
3. corn has been shown to express and accumulate high levels of monoclonal antibodies (proteins) not achieved in other plants.

According to the web page of the OSMID, the Stress Genomics Consortium as funded by NSF utilizes a variety of techniques to investigate the responses of plants and certain microbes to environmental factors of stress such as drought, chilling, and salinity. Hence the data provided in the OSMID database that could be used in the data mining include that for the variables of treatments with environmental factors of salt, cold, and drought.

The OSMID allows users to search for a gene of interest by name, id or by DNA or protein sequence. This microarray database is normalized by a uniform method of local iterative linear regression that minimizes the effect of spatial variation.

Microarray databases assembled by Wang et al. (2003) are available on OSMID website for corn and maize. A mircoarray database for corn is also provided in Wang et al. (2003) as log(2) normalized mean net signal pixel intensities. This log(2) normalization means that the local mean background has been subtracted from each spot prior to its normalization and log transformation.

ADVANCED DATA MINING APPLIED TO REPRESENTATIVE PLANT BIOTECHNOLOGY

Data Mining Of Microarray Databases

As discussed in Segall et al. (2003, 2004a, 2004b) microarray databases contain so much data that one cannot know in advance of any patterns in the data would appear upon selection of the variables of interest of the investigator. Data mining can identify patterns upon selected conditions using techniques such as clustering as shown in this paper.

Data mining in this paper as also in previous research of Segall et al. (2003, 2004a, 2004b) is used for the factor of salinity only and is one representative example of other possible factors that could have been used such as drought and temperature as previously mentioned.

Overview Of Data Mining Of Representative Ingredient Of Corn Using Normalized Data

This part of this paper present some advanced data mining using both predictive regression and clustering for the plant ingredient of corn that is representative for data mining of other plant biotech databases that could be used for either biotechnology analysis or manufacture of biopharmaceuticals using plants. The databases selected for the data mining presented are those representing the intensity of the ESTs (Experimental Stress Tolerances) for corn. Separate databases for corn for the factors are created by log (2) transformation ratios, and data mining for these are also performed and contrasted with the normal databases for ingredient of corn.

The software used for the data mining results presented in this paper is SAS Enterprise Miner using both the predictive and the cluster analysis modules respectively.

Data Mining Using Predictive Regression Of Ingredient Of Corn

Predictive modeling using regression with SAS Enterprise Miner generates charts for the numeric variables that include box and whisker plots, histograms, and descriptive statistics. The variables investigated for the corn microarray database included several input variables indicating the duration of the amount of salt in the controlled environment which were for durations of 1, 3, 6, 12, 24 and 72 hours. Other variables of interval measurement include prediction variable of P_HR_LOG2_SALT_CONTROL, residual variable of R_ HR_LOG2_SALT_CONTROL, and input variable for validation of
V_ HR_LOG2_SALT_CONTROL .

The frequency distributions of the data for each of these variables subjected to a logarithmic transformation as shown in Figures 2 (a) thru (g) are all approximately normally distributed as illustrated in the histogram figures. The percentage distributions for these same logarithmic transformations used in Figure 2 are shown in Figures 3 (a) thru (g). These observations of approximation to normality of all of these distributions of both Figures 2 and 3 for the frequency and percentage distributions respectively are as to be expected because the logarithmic transformation is often used to normalize data.

Comparing the statistics of these histograms of Figures 2 and 3, one more easily determine the modes or most frequent values using Figure 2. It can be noted that Figure 2 (g) with 72 hours log 2 of salt/control has the largest value of the frequency modes of about 2500, followed by that of Figure 2(d) of about 2000. One can determine the dispersion of the data values using the data values on the horizontal axes of Figure 3, which are also listed in Table 1. Using these horizontal axis data of Figure 3 or Table 1, one can conclude that the 3 hour log 2 of salt/control of Figure 3 (b) has the largest range of 5.4 and largest dispersion of .4865, with Figure 3 (e) of 24 log 2 of salt/control having the next largest range of 4.97 but the second smallest dispersion of .2583. Figure 3 (a) of 1-hour log 2 of salt/control of Figure 3 (a) has the smallest range of 2.5 and smallest dispersion of .2387.

Table 2 (a) and (b) provides a list of the variables used in the data mining. Table 2 (a) provides the definition of the variable names, and where Table 2 (b) indicates the model role and measurement type of the variable as either interval or unary. Table 2 (b) indicates that 1-hour log 2 salt/control has been used as the target variable of the data mining performed.

Figure 4 illustrates the convergence of the training and validation curves when the number of leaves equals 8 for the predictive regression performed. This is also indicated in Table 3 with the list of numerical values used in the plotting of Figure 4 in which the Training value is 0.0448 and Validation value is 0.0449.

Table 4 presents the predicted values of amount of salt in the controlled environment which were for durations of 1, 3, 6, 12 and 24 hours generated for each of the thirty (30) iterations of predictive regression performed using the predictive regression module of SAS Enterprise Miner. The values presented in Table 4 agree to be within those histogram bars visual within Figures 3 (a) thru (e) respectively.

Figure 5 provides the SAS Enterprise Miner output for the stepwise predictive regression that includes the Analysis of Variance (ANOVA) table. model fitting information, and Data Mining Regression (DMREG). The F-value pf 1349.98 and each of the t values such as –7.2 are all significant at the 99% level. This indicates that each of the intercepts of _hr_log2_salt_control_0 for 3 hours, _hr_log2_salt_control_1 for 6 hours, _2_hr_log2_salt_control_0 for 72 hours are significant at the 99% level.

Table 5 provides the training and validation values for each of the nineteen listed fit statistics. Table 6 provides the parameter estimates and Effect T-scores for each of the intercepts of the six effects of salt/control on plant of corn. As you will note from Table 6, two of the effects of _2hr_log2_salt_control for 12 hours and _4hr_log2_salt_control for 24 hours have no parameter estimates and corresponding Effect T-scores indicating that they are not present in the fitted data mining regression model. The intercept value as indicated by Table 6 is –0.02 that is near the origin.

Figure 6 provides a scatter plot of the predicted data mining regression data values versus the actual data values for the single effect 1_hr_log2_salt_control for the plant of corn. Note that a straight line could be fitted to the scatter that is not the same as that presented by the SAS

Enterprise Miner for predictive regression for the multiple effects simultaneously as discussed below.

Figure 5 of the SAS Enterprise Miner output has several stages of Step 0: Intercept entered, Step 1: Effect _hr_log2_salt_control_0 for 3 hours entered, Step 2: _hr_log2_salt_control_1 for 6 hours entered, Step 3: _2_hr_log2_salt_control_0 for 72 hours entered. In Step 3, the SAS Enterprise Miner output indicates that no additional effects other than the three listed met the 99.9% significance level for entry into the model and provides the analysis of parameter estimates for the intercept and the three effects as summarized in Table 6 of intercept = -0.0241, 2_hr_log2_salt_control_0 for 72 hours= 0.0617, _hr_log2_salt_control_0 for 3 hours=.1833 and _hr_log2_salt_control_1 for 6 hours= .0605.

Data Mining Using Advanced Cluster Analysis Of Ingredient Of Corn

Data mining using cluster analysis module of SAS Enterprise Miner was performed on the microarrays of corn data using techniques that were more advanced from those as presented in Segall (2003, 2004a, 2004b). The results are shown in Figures 7 and 8 respectively and Tables 7 and 8. Figure 7 shows the proportion correctly classified for the number of leaves and Figure 8 shows total leaf impurity, which is also refereed to as the Gini Index for the number of leaves. From Figure 7 we can conclude that the maximum proportion correctly classified increases as the number of leaves increases with a peak with the maximum of leaves of seven. Conversely, Figure 8 illustrates that the total leaf impurity decreases as the number of leaves increases and is a minimum when there is a maximum number of leaves of seven.

Table 7 provides the clustering criterion and other statistics related to the formation of six clusters as result of the data mining. These other statistics presented in Table 7 include maximum distance from cluster seed and distance to nearest cluster, and frequency of cluster. There is a uniform clustering criterion and uniform maximum relative change in cluster seeds. It can be noted from Table 7 that the maximum distance from cluster seed occurs with cluster number 6, which is the closest to cluster 2.

Table 8 provides the results on stress measurements on the plant of corn as subjected to each of the six (6) clusters for twenty-four (24) different intensities of and time durations of environmental stress factors of salt and controls ranging in treatment times from 1 hour, 3 hours, 6 hours, 12 hours, 24 hours, and 72 hours. Once can come to numerous conclusions by looking at the numerical values of Table 8 for the six clusters. Significant conclusions that can be made include the fact that cluster 6 has the largest numerical values for most of the treatments to corn. This is followed by cluster 3. Cluster 4 appears to the minimum numerical values for all of the treatments. Significant treatments are 6-hour control cy3 and 72-hour control cy3 as indicated by the huge numerical values.

CONCLUSIONS

This paper continues previous research conducted by the author on the applications of data mining to microarray databases, which is an exciting new area of data analysis because of the confluence of disciplines for biotechnology, information systems, and gene expression analysis or genomics. According to Hardiman (2003, p.38), the use of microarray databases have revolutionized the way in biomedical research has been conducted in the sense that "high-density arrays of specified DNA sequences can be fabricated onto a single glass slide or 'chip' ".

This research brings forth an essential point that state-of-the-art of database technology had lead to a new era in the sense that as stated by Hardiman (2003, p.38):

> "In the last decade with the drawing of the post genomic era researchers have progressed from analysis of gene expression, one gene at a time to the analysis of entire transcriptomes. Biological questions can now be asked at a qualitatively different scale than has been possible previously."

This paper has presented a description of what a microarray database is the reader, and brief summaries of previous related research on microarray databases both by the author of this paper as well as other authors within the last five (5) years.

This paper reviews a description of the plant databases as utilized in Segall (2003, 2004a, 2004b) that are representative of ingredients that could be used in the manufacture of plant-made-pharmaceuticals (PMPs), as available to the public domain on the Internet as the data warehouse of Osmotic Stress Microarray Information Database (OSMID) from a funded project of the National Science Foundation (NSF) named the "The Functional Genomics of Plant Stress."

Data mining was performed on the microarray databases for the plant of corn subjected to various levels of environmental stress and salt. Each of the frequency and percentage distributions of the logarithmic distributions of the microarray databases of corn ingredient at each of the levels of environmental factors yielded an approximate normal distribution as to be expected because the logarithmic distribution is used to normalize data. As stated previously, the 3-hour_log2 of Figure 2(b) has the largest dispersion of data, and the most intense duration of salt/control of 72 hours of Figure 2(f) yielded the maximum frequency of about 3000 corn plants that illustrated effects of the environmental factors in contrast to a maximum frequency of about 1600 corn plants for each of the other durations of the same salt/control factor except for that of 12 hours of Figure 2(d) of about 2000 corn plants.

The data mining was performed using SAS Enterprise Miner in both the predictive regression and cluster analysis modules. In the predictive regression module, the factor of 1-hour log2_salt/control was used as the target variable in the data mining performed, and in the cluster analysis module no target variable is required to be identified. The convergence of the predictive regression in eight (8) iterations as illustrated in Figure 4 indicates that convergence was not too difficult to achieve with this moderate number of iterations. The histograms predicted by SAS Enterprise Miner using the predictive regression module given by Figures 3(a)-3(g) approximate the frequency histograms given by Figures 2(a)-2(g) thus indicating a visual sense of agreement with the actual frequency data with that predicted by SAS Enterprise Miner predictive regression module. Thus, it again should be noted that Figure 3(g) has of the longest duration of 72 hours of the log2(salt/control) has the maximum percentage of almost 50% than any of the other five treatment levels presented.

One can conclude from the output from SAS Enterprise Miner for the corn microarray databases that each iteration introduced an additional log2(salt/control) factor that was significant at the

99.99% level. This is evidenced on the SAS Enterprise Mine output where Pr>|t| is given as <.0001 meaning the area in the normal tails are less than .01%. The final iteration of the SAS Enterprise Miner predictive regression module provides estimates for the parameters of intercept of –0.0241, _2hr_log_2_salt_control_0 for 72 hours as .0617, _hr_log_2_salt_control_0 for 3 hours of 0.1833, and _hr_log_2_salt_control_1 for 6 hours of 0.0605; which are all small numerical values that were obtained with a very significant level of accuracy of 99.99%. This level of accuracy is also confirmed by the very small value of the Average Squared Error (ASE) of 0.04474 as listed in Table 5. The regression of the individual selected factor of 1-hr_log2(control/salt) was not significantly linear as can be visually concluded from Figure 6 indicating that additional variables need to be included in the predictive regression model to improve linearity and minimize sum of squares of error, as were successfully done in multiple regression model obtained as described above in Figure 5. Hence the use of data mining technique of predictive regression was useful in the analysis of microarray databases of corn data for environmental stress factors.

Finally the use of SAS Enterprise Miner for advanced cluster analysis indicated that early convergence of the training and validation cluster analysis iterations for the six (6) clusters created were possible with seven (7) leaves when the proportion correctly classified was only about 20% and when the minimum leaf impurity or Gini Index was only about 89.5% as illustrated in Figures 7 and 8 respectively. Cluster 6 was the most distance from the next nearest cluster and also had the maximum distance from the cluster seed. As previously discussed, clusters 6 and 3 had overall the most significant factor results. Once again, we can conclude that the data mining technique of cluster analysis was useful in the analysis of the microarray databases of corn data for environmental stress factors.

FUTURE DIRECTIONS

The future directions of the research include additional analysis of other microarray of data different from corn that are abundant available on the website of OSMID. Other investigations that already have been started by the author include the construction of the box and whisker plots, density histograms, and descriptive statistics of moments and quantiles. Using SAS Enterprise Miner with the “Interactive Training” option by plotting two salt/control factor data simultaneously, and the density by branch and overall density for each of the salt/control factors already have indicated preliminary promise in new research conclusions. The use of the data mining tool of decision trees for each factor as a control with and without salt also has been preliminarily tested as a future direction of the research. Also the use of SAS Enterprise Miner with both the linear and logistic regression model options, and the plotting of the residuals of other environmental factors also have produced some promising results for future research. Comparisons of the results for the other OSMID plants with that obtained using the corn data could provide some interesting results in the future research.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the funding provided by a block grant from the Arkansas Biosciences Institute (ABI) as administered by Arkansas State University (ASU) to encourage development of a focus area in Biosciences Institute Social and Economic and Regulatory

Studies (BISERS) for which he served as Co-Investigator (Co-I) in 2003, and with which funding the analyses of the Osmotic Stress Microarray Information Database (OSMID) discussed within this paper were performed.

The author also wishes to acknowledge a three-year software grant from SAS Incorporated to the College of Business at Arkansas State University for SAS Enterprise Miner that was used in the data mining of the OSMID microarrays discussed within this paper.

REFERENCES

Abdullah, A., Brobst, S., Pervaiz, I., Umer M., and Nisar A. (2003), "Learning dynamics of pesticide abuse through data mining, The Australasian Workshop Data Mining and Web Intelligence, Dunedin, New Zealand, *Conferences in Research and Practice in Information Technology*, v. 32, Editors Hogan, J., Montague, P., Purvis M., and Steketee, C.

Alberta Research Council (2000), Data Mining, http://www.arc.ab.ca/agric/DataMining.asp

Baldi, P, Hatfield GW, and Hatfield WG (2003), DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling, Cambridge University Press.

Bartley, GE and Ishida BK (2002), "Digital fruit ripening: data mining in the TIGR tomato gene index, Plant Molecular Biology Reporter, v. 20, p. 115-130, June 2002.

Bergeron, B (2003), Bioinformatics Computing, Prentice Hall.

Berrar DP, Dubitzky, W amd Granzow M, A Practical Approach to Microarray Data Analysis, Kluwer Academic Publishers, 2003.

Bluck, JG and Menting, AM, "NASA data mining reveals a new history of natural disasters," *NASA News*, http://amesnews.arc.nasas.gov/releases/2003/03_51AR.html , July 8, 2003.

Causton, HC, Quackenbush, J and Brazma, A (2003), Microarray Gene Expression Data Analysis: A Beginner's Guide, Blackwell Publishers.

Claverie, JM and Notredame, C, (2003), Bioinformatics for Dummies,

Embrechts, MJ (2002), "Visual Explorations in Clustering and Data Mining," Bioinformatics Workshop, July 24-26, 2002, Rensselaer Polytechnic Institute, Troy, NY, http://www.drugmining.com .

GenomeWeb LLC (2003), "Oracle hopes to strike bioinformatics gold with embedded data-mining tools," bio1INFORM, v.7, n8, February 24, 2003, http://otn.oracle.com/industries/life_sciences/ pdf/ls_bio1nf0rm.pdf

Hardiman, G (2003), Microarrays Methods and Applications: Nuts & Bolts, DNA Press, LLC, Eagleville, PA, www.dnapress.net.

Hsu, WH and Joehanes, R. (2003), "Learning the structure of graphical models of gene regulation from microarray data: survey and experiments," Laboratory for Knowledge in Databases, Kansas State University, Manhattan, KS.

Kennedy, GC and Wilson, IW (2004), "Plant functional genomics: opportunities in microarry databases and data mining," *Functional Plant Biology*, v. 31, pp. 295-314.

Lesk, AM (2003), Introduction to Bioinformatics, Oxford University Press.

Monsanto (2002), Plant-Made Pharmaceuticals: A New Way to Make Medicine.

National Park Service, U.S. Department of Interior, NPSpecies Data Mining and Geo-Reference Tools (2003),http://science.nature.nps.gov/im/apps/npspp/DataMine.htm

Rosenquist, M, Alsterfjord, Larsson, C, and Sommarin, M. (2001), "Data mining the arabidopsis genome reveals fifteen 14-3-3 genes. Expression is demonstrated for two out of five novel genes," Plant Physiology, v. 127, n.1, pp. 142-149, September 2001.

Schena, M (2003), Microarray Analysis, Wiley Publishing Company.

Segall, RS, Guha GS, and Nonis S (2003), Data mining for analyzing the impact of environmental stress on plants-A case study using OSMID, submitted and in revision status to *Kybernetes: International Journal of Systems and Cybernetics.*

Segall, RS and Nonis S (2004a) Data mining for analyzing the impact of environmental stress on plants-A case study using OSMID, accepted for publication in *Acxiom Working Paper Series of Acxiom Laboratory of Applied Research (ALAR)* and presented at Acxiom Conference on Applied Research and Information Technology, University of Arkansas at Little Rock (UALR), February 27, 2004.

Segall, RS, Guha GS, and Nonis S (2004b), Data mining for assessing the impact of environmental stresses on plant geonomics, *Proceedings of the Thirty-Fifth Meeting of the Southwest Decision Sciences Institute*, May 4-6, 2004, Orlando, FL, pp. 23-31.

Shah, D., "Data Mining of GeneChips", INT3470 Course Project, http://www.atsweb.neu.edu/physics/b.barbiellini/int3470/devang/Data%20Mining%20of%20GneChips%AE.ppt, March 14, 2002.

Spiliopoulou, M. (2004), Data mining for business applications: Data preparation, http://omen.cs.unimagdeburg.de:8080/iti_kmd/lehre/WS03/Slides/DM_Marketing/KDDmarkt_dataprep.pdf .

Stekel, D (2004), Microarray Bioinformatics, Cambridge University Press.

Torto T., Styer, A., and Kamoun, S., "EST data mining: Novel extracellular proteins from the Oomycete plant pathogen *PHYTOPHTHORA INFESTANS,* (2000), http://www.biosci.ohio-state.edu/~plantbio/ seminar/2000abstracts/Torto.html

Uppsala Monitoring Centre (2004), Projects, "The use of data-mining for signaling adverse drugs reactions: Monitoring of herbal medicines", http://www.who-umc.org/projects.html .

Wang et al. (2003) "A maize QTL for silk maysin levels contains duplicated Myb-homologous genes which jointly regulate flavone biosynthesis, " *Journal of Plant Molecular Biology*, v. 52, n. 1, pages 1-15, May 2003.

Wotawa MA and Kinston, S. (2003), "Data Dictionary: Data Mining to Populate NPSpecies", National Park Service, Inventory and Monitoring Biological Inventories,http://science.nature.nps.gov/im/apps/npspp/DataMine.htm .

Zimback L, Mori ES, de Morase ML, Rosa DD, Furtado EL, Marino, CL, Wilken CF, Velini ED, Guerrini AI, Maia IG, and Camargo LE (2004), "Data mining of Eucalyptus ESTs involved in the mechanism non hormonal growth genes," International Plant & Animal Genomes XII Conference, San Diego, CA, January 10-14, 2004, http://63.141.253.172/12/abstracts/P01_PAG12_52.html

Tables and Figures are available from author upon request.