Cloud-Scale Genomic Signals Processing Classification Analysis for Gene Expression Microarray Data

Benjamin Harvey-*IEEE Member* and Soo-Yeon Ji-*IEEE Member*

Abstract **- As microarray data available to scientists continues to increase in size and complexity, it has become overwhelmingly important to find multiple ways to bring inference though analysis of DNA/mRNA sequence data that is useful to scientists. Though there have been many attempts to elucidate the issue of bringing forth biological inference by means of wavelet preprocessing and classification, there has not been a research effort that focuses on a cloud-scale classification analysis of microarray data using Wavelet thresholding in a Cloud environment to identify significantly expressed features. This paper proposes a novel methodology that uses Wavelet based Denoising to initialize a threshold for determination of significantly expressed genes for classification. Additionally, this research was implemented and encompassed within cloud-based distributed processing environment. The utilization of Cloud computing and Wavelet thresholding was used for the classification 14 tumor classes from the Global Cancer Map (GCM). The results proved to be more accurate than using a predefined p-value for differential expression classification. This novel methodology analyzed Wavelet based threshold features of gene expression in a Cloud environment, furthermore classifying the expression of samples by analyzing gene patterns, which inform us of biological processes. Moreover, enabling researchers to face the present and forthcoming challenges that may arise in the analysis of data in functional genomics of large microarray datasets**.

I. Introduction

Cloud-Scale Genomic signal processing is a research modernization that combines the advantages of distributed processing and advanced signal processing methodologies for enhancing microarray data analysis. Microarrays provide a powerful method for simultaneous monitoring of gene expression levels [1]. In this paper, we present a methodology, which utilizes Wavelet-based threshold denoising in order to identify significant genes within a microarray dataset. Determination of differentially expressed genes is a widely used technique in feature selection for classification analysis. This research aims to uses Wavelet –based denoising to determine a threshold and compare the outcome of classification to the well-known differential expression significance determination. Furthermore, this will allow finding target genes that are expressed significantly with specific correlations between both genes and samples unproblematic. Iterative analysis of generated microarray frequencies by the Wavelet transformation becomes computationally expensive when dealing with highly dimensional feature vectors. Applications of Cloud computing have had a major impact on analysis performance in determining inter and intra genomic relationships among genes within samples. Although previous gene signals processing analysis techniques have been successfully applied to analyze selected genes disease susceptibility using expression profiles, there is still a need to carry out an analysis that can take into account expression profiles and signatures for a genome wide analysis [2]. Moreover, identifying multiple highly expressed correlated genes that can lead to disease causation. Therefore, we present an efficient and robust microarray data analysis methodology to identify abnormal genes by combining cloud computing, advanced signal processing techniques, and machine learning classification algorithms. The proposed methodology uses: 1) Cloud Computing and distributed processing - to increase the speed of data training, Wavelet coefficient generation, and threshold determination; 2) Wavelet-based denoising - applied to identify a threshold for differential gene expression in order to separate abnormal genes from normal genes for during microarray classifications analysis; and 3) Classification - applied to the selected features determined through Wavelet thresholding for significant feature expression classification of the GCM tumor and normal samples. Multiple classification techniques were applied to the selected features and their performance was compared.

This research study is partially supported by US Army Research Office (ARO) grant W911NF13110143.

B. S. Harvey is with Bowie State University, MD 20715 USA. He is also with the Department of Defense, National Security Agency, Fort Meade, MD 20755 USA. (phone: 904-662-6611; e-mail: bsharve@lgmail.com).

S. Ji is with Bowie State University, MD 20715 USA. (e-mail: sji@bowiestate.edu).

II. Global Cancer Map Microarray Data

The data that was used for this experiment was developed and created during a microarray experiment which analyzed 218 tumor samples, spanning 14 common tumor types, and 90 normal tissue samples by oligonucleotide microarray gene expression analysis [3]. The expression levels of 8934 genes were utilized in the initial experiment. Of the initial 314 tumor and 98 normal tissue samples processed, 218 tumor and 90 normal tissue samples passed quality control criteria and were used for subsequent data analysis. During quality assessment we reduced the number of samples from 308 to 279 due to reproducibility, identification of apparent outlier arrays, and computation of signal-to-noise ratio measures determined by the *Bioconductor arrayQualityMetrics* pckage [4]. The dataset was dowloaded in the form of .CEL files and uploaded into as an expression dataset through Bioconductor packages and were normalized using RMA [4].

III. Cloud Scale Distributed Processing

RHadoop is a collection of three *R* packages: *rmr, rhdfs* and *rhbase*. The *rmr* package provides Hadoop MapReduce

functionality in R, *rhdfs* provides HDFS file management in *R* and rhbase provides *HBase* database management from within *R* for analysis of Bioconductor datasets [5]. The functional mechanisms of the rmr programming model for expression/genotype does analysis by determining a set of microarray gene key/value input pairs and then reduces to another set of output key/value pairs [6]. The map function created a mapping of all microarray keys/values in the form of intermediate key/value pairs. The *rmr* library then routinely grouped all intermediate microarray key/value pairs together and passed them to the reduce function. The reduce function then merged together these processed microarray values to form a smaller set of values [7]. The map and reduce functions produced results from two different domains, furthermore the intermediate values of these domains were migrated. The map function created microarray key/value pairs of the gene and samples and maped them across all nodes on the Amazon EC2 cluster. The *rmr* specification object took in arguments of the names of the input and output files associated with the initial data to be computed [8]. The application of Cloud computing combined with expression analysis had a major impact on the analysis performance and determination of inter and intra genomic relationships among genes within samples. The integration of *rmr*, *rhdfs* and *rhbase* packages also aided with the distribution of processes involved with preprocessing and determining differential and Wavelet thresholds in the gene expression calculations from the expression data. The distribution of the processes decreased the total time of execution by a factor of the number of total processors (nodes).

IV. Microarray Signals Processing Methodology

This paper presents a microarray Wavelet threshold methodology that uses GSP to estimate the most efficient metric for significance based upon local features in the gene sample. We utilized two separate expression data sets, which included:

- Differentially expressed genes/features (**685**) based on a p-value that would support the Biological significance of the microarray data results relative to the problem under investigation
- Significantly expressed genes/features (**191**) based on a Wavelet Threshold value that would support the Biological significance of the microarray data results relative to the problem under investigation.

The steps in the overall methodology included:

A. Cloud Infrastructural Implementation.

RHadoop a collection of three *R* packages: *rmr, rhdfs* and rhbase were utilized on an Amazon EC2 cluster in order to enable the distribution of the processes and decreased the total time of execution by a factor of the number of total processors (nodes).

B. Import of raw Global Cancer Map Microarray data

Utilized the *R* affymetrix microarray package to import raw microarray samples

C. Quality Assessment

Was done before normalization utilizing package *arrayQualityMetrics* for convenience functions with automatic outlier detection

D. Microarray Preprocessing

Applied Background correction, normalization, and summarization to the GCM microarray data. Utilized RMA normalization (Robust Multi-Array Average expression measure form *R & Bioconductor* package *affy*)

E. Microarray Testing and Ranking

Calculation of the columns statistic, p-value and dm, the difference of the group means (only in the case of the t-test functions).

F. GSP Non-Specific Filtering

Created two separate expression datasets (1) differential expression with 685 features and (2) Wavelet Threshold value with 191 features, for analysis and comparison

F. Classification.

This refers to the selection of relevant algorithms used for classification of the data associated with each expression set to either Normal or Tumor.

- Compared the results from the following classifiers: randomForrest, Knn, DiagDA, nnet, rpart, lda, svm, qda, glm, ada, backboost, lvq, naiveBayes, bagging, slda, rda, and ksvm

G. Validation.

10-fold Cross Validation

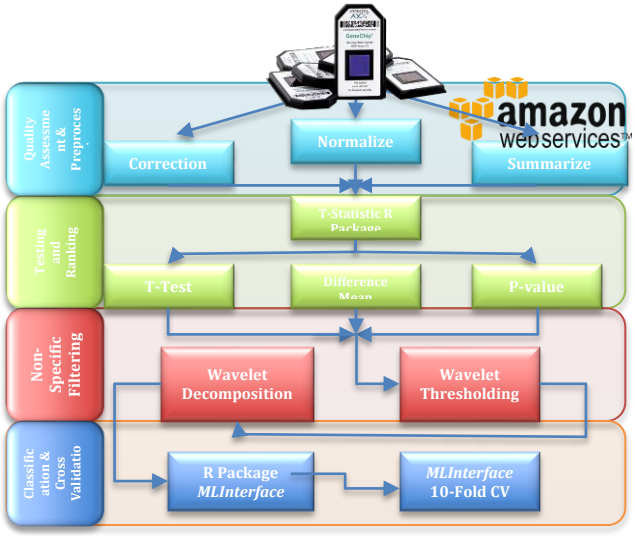

Figure 1. Gives an architectural workflow overview of the methodology and the steps associated with each component.

V. Wavelet Thresholding vs. Gene Differential Expression

Microarray heat maps are typically used in gene expression microarray analysis to represent the level of expression of many genes across a number of comparable

samples as they are obtained from a number of samples. In order to identify genes/features that were significant in mRNA expression during the GCM data, we determined a p-value for differential expression that would support the Biological significance of the microarray data results relative to the problem under investigation. We then compared the mRNA levels from each gene by using Bioconductor package *rowttests* which performed for each row a two-sided one-class t-test against the null hypothesis 'mean=0' [7]. We obtained the associated values per each row, which included the columns t-test *statistic*, *p-value* (optional in the case of the t-test functions) and *dm*, the difference of the group means (only in the case of the t-test functions.

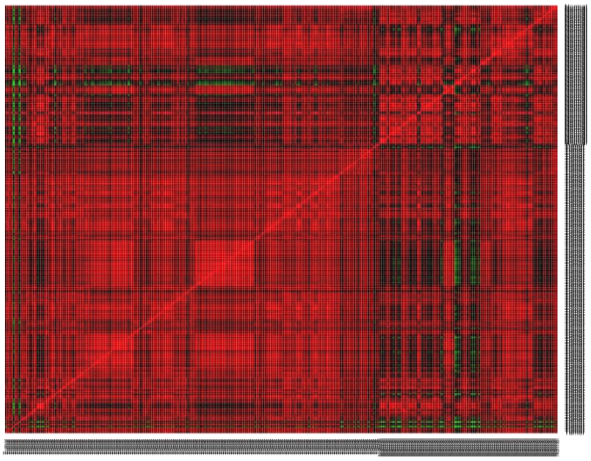

Figure 2: Heat map of Global Cancer Map (GCM) sample expression levels generated from Differentially Expressed method for genes/features selection.

Figure 2 depicts a Heat map representation of the expression set for 685 differentially expressed of the 279 GCM samples in order to show how experimental conditions influenced production (expression) of mRNA for a set of genes. The GCM Heat map shows the largest gene expression values which are displayed in red (hot), dispersed among the smallest values in green.

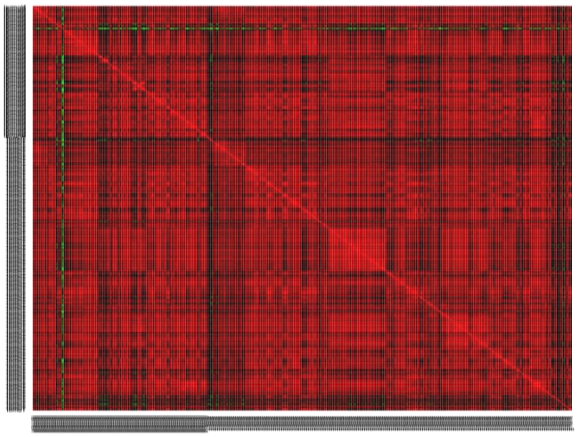

Figure 3: Heat map of Global Cancer Map (GCM) sample expression levels generated from Wavelet Thresholding method for genes/features selection.

Figure 3 shows the Heat map of the Wavelet-based threshold expression set (191 features/genes) which was created by Wavelet-based thresholding which denoised the expression signature *x* by using the 2-band wavelet system described by the Daubechies' wavelet filter using the traditional discrete wavelet transform (DWT) [4]. The GCM Heat map shows the largest gene expression values which are displayed in red (hot), segmented in separate clusters among the smallest values in green which increases overall the classification accuracy based upon the feature selection method.

VI. Wavelet Threshold Gene Differential Expression

Wavelet transform provides us with one of the methods for microarray image denoising Wavelet transform, due to its excellent localization property, has rapidly become an indispensable signal and microarray image processing tool for a variety of applications, including denoising and compression attempts to remove the noise present in the microarray signal while preserving the signal characteristics, regardless of its frequency content. In this paper we utilize Universal Thresholding for Wavelet Filtering, which involved three steps:

- a Linear forward Wavelet transform
- Nonlinear thresholding step and
- a linear inverse wavelet transform

Wavelet thresholding is a signal estimation technique that exploits the capabilities of wavelet transform for signal denoising. [9] It removed noise by killing coefficients that were insignificant relative to the determined threshold. In order to eliminate coefficients we incorporated a technique for choosing denoising parameters which encompassed using the Mean Absolute Deviation (MAD) applied to the noise variance estimator for universal threshold determination. Finally, we utilized *hard* thresholding which applied a "keep or kill" procedure to eliminate coefficients beyond the determined threshold generated by MAD noise variance estimation.

VII. Results

The expression sets (1) differential expression with 685 features and (2) Wavelet Threshold value with 191 features were used for classification. These expression sets were classified using multiple Machine Learning algorithms and the results were tabulated and compared.

TABLE I. Gene Classification

Classification and (Gene Feature Selection)		Predicted Outcome		Actual Outcome		Error		Cross Valid.
Machine Learning Algorithm	Genes Features	Norm	Tum	Norm	Tum	Norm	Tum	10-Fold
KNN (Wavelet Threshold)	191	47	112	22	104	0.468	0.071	84%
KNN (differentially expressed)	685	54	105	24	99	0.444	0.057	89%
SVM (differentially expressed)	685	44	115	23	108	0.523	0.061	78%
SVM (Wavelet Threshold)	191	45	114	22	106	0.489	0.070	90%
KSVM (differentially expressed)	685	41	118	22	110	0.537	0.068	87%
KSVM (Wavelet Threshold)	191	45	114	23	108	0.511	0.053	79%
rpart (differentially expressed)	685	54	105	23	98	0.426	0.067	94%
rpart (Wavelet	191	55	104	21	95	0.382	0.087	96%

Threshold)								
randFor (differentially expressed)	685	44	115	21	106	0.477	0.078	92%
randForrest(Wavelet Threshold)	191	39	120	20	110	0.513	0.083	87%
nnet (differentially expressed)	685	58	101	25	96	0.431	0.050	85%
nnet (Wavelet Threshold)	191	58	101	24	95	0.414	0.059	98%
gda (differentially expressed)	685	44	115	24	109	0.545	0.052	89%
gda (Wavelet Threshold)	191	54	105	23	98	0.426	0.067	79%
dla (differentially expressed)	685	54	105	19	94	0.352	0.105	88%
dla (Wavelet Threshold)	191	58	101	25	96	0.431	0.050	92%

a. Gene classification statistics for differential expression and Wavelet threshold feature selection

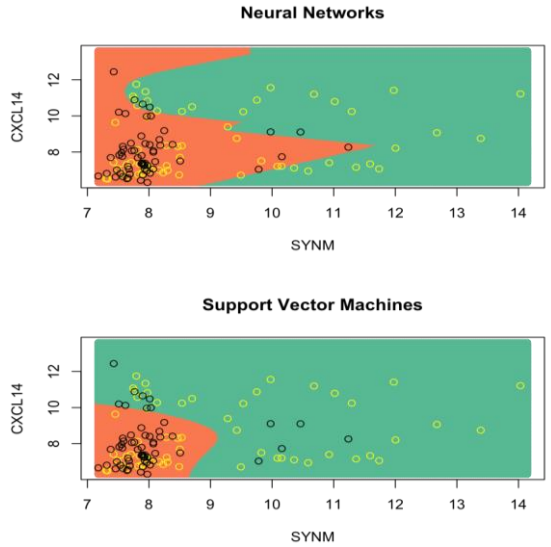

Figure 4: Shows the classification analysis of the selected expression set features produced by p-value Differential Expression. Shows the Classification boundaries presented by, nnet and svm on the plane dictated by two genes in an expression set

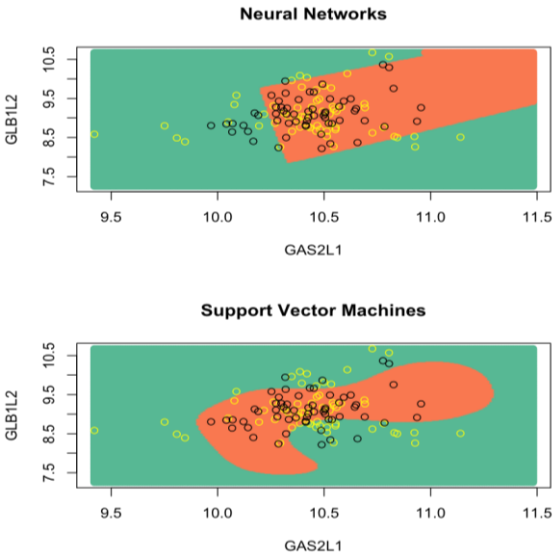

Figure 5: Shows the classification analysis of the selected expression set features produced by Wavelet based thresholding. Shows the Classification boundaries presented by nnet, and svm on the plane dictated by two genes in an expression set

VIII. Conclusion

Because differential expression analysis is a very popular method for gene expression extracting features for classification, we compared the results of the differential expression feature selection versus the Wavelet based thresholding method used for feature selection. Conclusion

The proposed method is robust and can be potentially used to identify genes, which have the same patterns or biological processes in similar datasets. Although there have been many studies dedicated to utilize differential expression for feature selection and GSP for microarray analysis, none have utilized Wavelet thresholding for feature selection and classification in a Cloud computing environment to-date. The wavelet method is more computationally expensive than other GSP methods, but it is more sensitive to detect sudden changes in input data. The novel methodology proposed within this research analyzed Wavelet based threshold features of gene expression in a Cloud environment, furthermore classifying the expression of samples by analyzing gene patterns, which inform us of biological processes. Furthermore, this research will help to face the present and forthcoming challenges that may arise in the analysis of genetics data in functional genomics.

References

[1] R. S. Istepanian, A. Sungoor, and J.-C. Nebel, "Comparative analysis of genomic signal processing for microarray data clustering," *NanoBioscience, IEEE Transactions on,* vol. 10, pp. 225-238, 2011.

[2] R. C. Taylor, "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics," *BMC Bioinformatics,* vol. 11 Suppl 12, p. S1, 2010.

[3] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, *et al.*, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences,* vol. 98, pp. 15149-15154, 2001.

[4] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, *et al.*, "Bioconductor: open software development for computational biology and bioinformatics," *Genome biology,* vol. 5, p. R80, 2004.

[5] V. Prajapati, *Big Data Analytics with R and Hadoop*: Packt Publishing Ltd, 2013.

[6] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM,* vol. 51, pp. 107-113, 2008.

[7] Q. Zou, X. B. Li, W. R. Jiang, Z. Y. Lin, G. L. Li, and K. Chen, "Survey of MapReduce frame operation in bioinformatics," *Brief Bioinform,* Feb 7 2013.

[8] K. Fan, X. Sun, Y. Tao, L. Xu, C. Wang, X. Mao, *et al.*, "High-Performance Signal Detection for Adverse Drug Events using MapReduce Paradigm," *AMIA Annu Symp Proc,* vol. 2010, pp. 902-6, 2010.

[9] D. L. Donoho, "De-noising by soft-thresholding," *Information Theory, IEEE Transactions on,* vol. 41, pp. 613-627, 1995.