Cloud-Scale Genomic Signals Processing for Robust Large-Scale Cancer Genomic Microarray Data Analysis

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

***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 forth oncological inference to the bioinformatics community through the analysis of large-scale cancer genomic (LSCG) DNA and mRNA microarray data that is useful to scientists. Though there have been many attempts to elucidate the issue of bringing forth biological interpretation by means of wavelet preprocessing and classification, there has not been a research effort that focuses on a cloud-scale distributed parallel (CSDP) separable 1-D wavelet decomposition technique for denoising through differential expression thresholding and classification of LSCG microarray data. This research presents a novel methodology that utilizes a CSDP separable 1-D method for wavelet-based transformation in order to initialize a threshold which will retain significantly expressed genes through the denoising process for robust classification of cancer patients. Additionally, the overall study was implemented and encompassed within CSDP environment. The utilization of cloud computing and wavelet-based thresholding for denoising was used for the classification of samples within the Global Cancer Map, Cancer Cell Line Encyclopedia, and The Cancer Genome Atlas. The results proved that separable 1-D parallel distributed wavelet denoising in the cloud and differential expression thresholding increased the computational performance and enabled the generation of higher quality LSCG microarray datasets, which led to more accurate classification results.**

***Index Terms*—Cancer, cloud, decomposition, distributed, genomics, parallelization, processing, scale, thresholding signal.**

I. Introduction

CLOUD-SCALE genomic signal processing is a research modernization that combines the advantages of distributed, parallel, and scalable processing as well as advanced signal processing methodologies for enhancing microarray data analysis. Microarrays provide a powerful method for simultaneous monitoring of gene expression levels [1]. In this study, a novel methodology is presented, which utilizes wavelet-based thresholds for denoising to remove noise from microarrays while retaining gene significance within the microarray dataset. Determination of differentially expressed genes is a widely used technique in selection of significant features, which can impact accuracies within a classification analysis. Moreover, this study presents a novel methodology in which differentially expressed genes are utilized to determine the threshold during wavelet-based denoising, with the ultimate goal of safeguarding important genes during the wavelet transformation process. This study also provides a denoising mechanism through wavelets that can assure the fortification and safekeeping of inter- and intragenomic relationships that are expressed by specific correlations between both genes and samples throughout the transformation and denoising process. Iterative analysis of generated microarray frequencies by the wavelet transformation becomes computationally expensive when dealing with highly dimensional feature vectors [1]. Applications of Cloud computing have had a major impact in enabling highly dimensional analysis allowing elucidation and determination of inter- and intragenomic relationships among genes. Unfortunately, there has not been a research effort to implement a cloud-based separable 1-D wavelet-based transformation which can be applied on single genes instead of an entire 2-D representation of the microarray data. Although signals processing techniques have been successfully applied to analyze selected gene disease susceptibility using expression profiles in previous studies, there is still a need to carry out an analysis that can sustain genomic relationships that may be removed throughout the denoising process [2]. Moreover, this study aims to denoise the Global Cancer Map (GCM) large-scale cancer genomic (LSCG) dataset while retaining the significance of multiple highly expressed correlated genes that can lead to disease classification and causation. Therefore, we present an efficient and robust microarray data analysis methodology to classify tumor and normal samples by combining cloud computing, advanced signal processing, and machine learning techniques. The proposed methodology uses:

1) cloud-scale distributed parallel (CSDP) environment to increase the speed of data training, wavelet transform coefficient generation, and threshold determination;
2) wavelet-based feature decomposition, thresholding, and denoising to determine a genetically based threshold by

Manuscript received May 2, 2015; revised September 2, 2015; accepted October 14, 2015. Date of publication November 3, 2015; date of current version January 31, 2017. This work was supported in part by the U.S. Army Research Office Grant W911NF-13-1-0143.

B. S. Harvey is with the Computer Science Department, Bowie State University, Bowie, MD 20715 USA, and also with the Department of Defense, National Security Agency, Fort Meade, MD 20755 USA (e-mail: bsharve@gmail.com).

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

Digital Object Identifier 10.1109/JBHI.2015.2496323

utilizing differential gene differential expression in order to improve the identification of significant gene features and eliminating of noise to increase the effectiveness and efficiency of the analysis;

3) classification—R/bioconductor *MLInterfaces* *k*-NN, support vector machine (SVM), and neural network were applied to denoised microarray datasets and their performance was compared.

II. LSCG Microarray Data

The microarray data used in this paper was developed and created for the purpose of large-scale analysis of cancer microarray experiments. This research utilizes three different LSCG microarray datasets which included the GCM, Cancer Cell Line Encyclopedia (CCLE), and The Cancer Genome Atlas (TCGA). The GCM dataset consisted of 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 tumors and 90 normal tissue samples passed quality control criteria and were used for subsequent data analysis. During quality assessment, the number of samples decreased from 308 to 279 due to the identification of apparent outlier arrays. The dataset was downloaded in the form of .CEL files which were ingested and represented as an expression dataset using R/Bioconductor packages and were normalized using RMA [4]. The CCLE describes a conglomeration of gene expression, chromosomal copy number, and sequence data in accordance with 947 different human cancer cell lines and 36 different tumor types with over more than 50 000 different gene features represented. This information is combined with the pharmacological drug information for 24 anticancer drugs across 479 of the cancer cell lines [5]. TCGA Pan-Cancer project assembled data from thousands of patients with primary tumors occurring in different sites of the body, covering 12 tumor types. This research specifically analyzed the ovarian carcinoma cells of 22 277 gene features across 598 samples [6].

III. Cloud-Scale Distributed Parallel

In order to implement an environment in the cloud, we utilized a parallel, distributed, scalable wavelet transformation with *R* and a software package called *Parallel* in order to parallelize wavelet transformations that analyzed microarray gene expression profiles along the rows of the microarray data structure. In addition, utilized *Hadoop in* Amazon Work Station (AWS), Elastic Compute Cloud (EC2), and Elastic MapReduce (EMR). For security purposes, we implemented the instances within VMWare and deployed those virtual nodes to the cloud before utilization. This enabled secure analysis through Hadoop in order to parallelize functions directly utilizing Hadoop Distributed File System for genomic signal processing. Moreover, providing the ability to write code that transparently manipulates MapReduce jobs in R using Hadoop Streaming and the Parallel package [7]. Additionally, the usage *StarCluster* and *R Parallel* enabled the capability to allow for the parallelization of wavelet functions within a given job among multiple processors in AWS to be implemented.

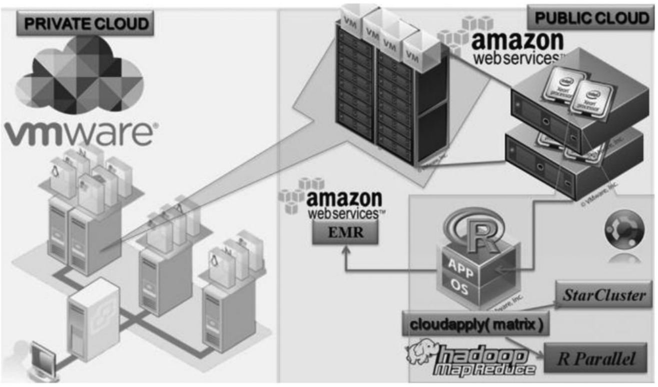


Fig. 1. Usage of VMWare to deploy instances to the AWS public cloud where 20 nodes were implemented with a *cloudGSPApply* function that enabled the parallelization of wavelet transformations in the cloud.

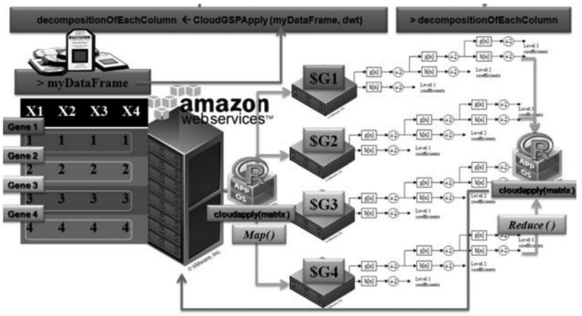


Fig. 2. Cloud-scale parallel wavelet transformation within the CSDP environment in AWS distributing data to each of the nodes for analysis.

StarCluster is a utility for creating and managing distributed computing clusters hosted on Amazon's Elastic Compute Cloud (EC2). *StarCluster* utilizes Amazon's EC2 web service to create and destroy clusters of Linux virtual machines on demand. *StarCluster* provides a convenient way to quickly set up a cluster of machines to run analytics on data with parallel jobs using a distributed memory framework. *R Parallel* utilizes a *parLapply* function from the Parallel package to run a task on the cluster without further configuration, but does not have the ability to run wavelet transformation with selected parameters. This research study utilized a parallel framework on a cluster with 20 AWS/StarCluster nodes and implemented the *CloudGSPApply* function similar to the *parLapply*, but allowed for 1-D separable wavelet-based parallel transformations in a CSDP environment, which partitioned each of the gene expression profiles within the microarray to separate nodes and did a 1-D wavelet transform on each.

IV. CSDP Wavelet Transformation and Classification for GCM

In previous studies, there has been the utilization of 2-D wavelet transformations thresholding method for analyzing GSP microarray data. The proposed method takes a microarray image does a series of steps, which include the following:

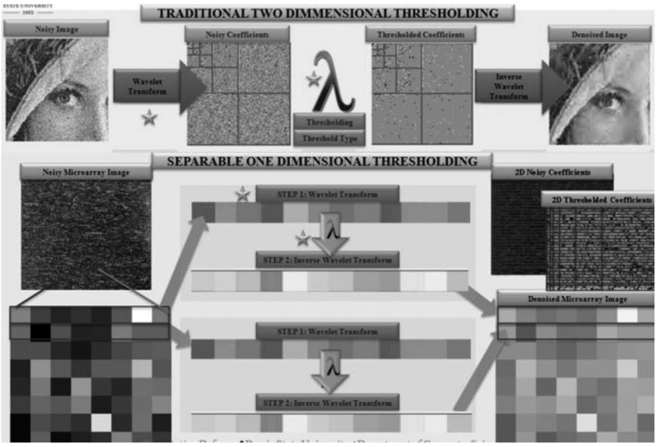

TRADITIONAL TWO DIMMENSIONAL THRESHOLDING
Noisy Image
Wavelet Transform
Noisy Coefficients
λ
Thresholding
Threshold Type
Thresholded Coefficients
Inverse Wavelet Transform
Denoised Image
SEPARABLE ONE DIMENSIONAL THRESHOLDING
Noisy Microarray Image
STEP 1: Wavelet Transform
STEP 2: Inverse Wavelet Transform
2D Noisy Coefficients
2D Thresholded Coefficients
Denoised Microarray Image

Fig. 3. Difference between the traditional 2-D wavelet transformation and thresholding compared to the novel separable 1-D wavelet transformations and thresholding.

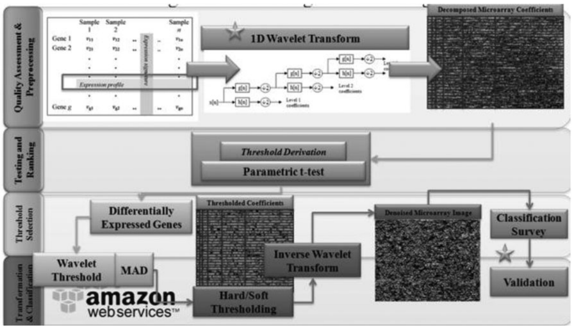


Fig. 4. Methodology for the novel separable 1-D wavelet: (1) transformation, (2) thresholding, (3) classification and validation.

Step 1: Utilized one low-pass filter which is represented in the x-direction and high-pass filter which is represented in the y-direction $\psi^1(x,y) = \varphi(x)\psi(y)$, which yields a corresponding detail coefficient D^x.

Step 2: Utilized one is low-pass filter that is represented in the y-direction and high-pass filter which is represented in the x direction as $\psi^2(x,y) = \psi(x)\varphi(y)$ yielding detail coefficients D^y.

Step 3: Utilized one final high-pass filter represented in both x- and y-directions which can be represented as $\psi^3(x,y) = \psi(x)\psi(y)$ yielding detail coefficient D^{xy}.

This established a horizontal, vertical, and diagonal orientation for details that are now represented within the microarray image. This can be visualized within Fig. 4 as the Lena image has been decomposed at four wavelet levels thresholded λ and transformed back to its original feature space. The novel separable 1-D wavelet-based denoising methodology transforms only one low-pass filter, which is represented in the x-direction, and one high-pass filter, which is represented in the y-direction $\psi^1(x,y) = \varphi(x)\psi(y)$, which yields a corresponding detail coefficient D^x for each gene in the microarray image represented in the lower region of Fig. 4.

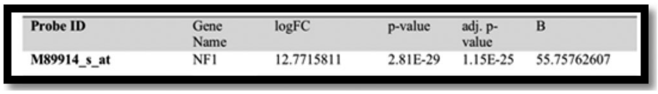
Fig. 4 represents the methodology for the novel separable 1-D wavelet transformation and thresholding. The steps in the separable 1-D wavelet thresholding and classification are as follows:

Step 1: Ingest the microarray data into the AWS master node

Step 2: Apply the *cloudGSPApply* function to the data and specify the following parameters for the wavelet-based denoising:

a) Levels—1–8.
b) Type—Daubechies.
c) Thresholding—(0) MAD, (1) parametric t-test, (2) nonparametric t-test.

Step 3: Apply k-NN, SVM, and NN classification to the denoised dataset.



Probe ID	Gene Name	logFC	p-value	adj. p-value	B
M89914_s_at	NF1	12.7715811	2.81E-29	1.15E-25	55.75762607

Fig. 5. Parametric and nonparametric p-value for gene NF1.

V. Cloud-Scale Wavelet Thresholding Utilizing Gene Differential Expression

Wavelet transformation and thresholding provides us with an important mechanism 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 [3]. In this paper we utilized mean absolute deviation (MAD) as well as a parametric and nonparametric t-test for wavelet-based thresholding for denoising, which involved three steps:

1) a linear forward wavelet transform;
2) identification of differentially expressed genes;
3) differentially expressed thresholding step;
4) a linear inverse wavelet transform.

Wavelet thresholding is a signal estimation technique that exploits the capabilities of wavelet transform for signal denoising [8]. It removes noise by killing coefficients that were insignificant relative to the determined threshold. In order to eliminate coefficients in microarray problems we incorporated two different techniques for choosing a threshold. This encompassed the following:

1) Apply thresholding to insignificant genes and then utilize the MAD as the universal threshold for values deemed not to be differentially expressed between the normal and abnormal groups.
2) Utilize the p-value of each gene as method to threshold on a gene by gene basis. For example, in Fig. 5 for the gene *NF1*, utilize the p-value and the adjusted p-value as the threshold for gene.

We utilized a parametric p-value and nonparametric p-value to threshold the 1-D gene representation of each gene. Finally, we utilized *soft* thresholding which applied a “keep or bring up to threshold” procedure to eliminate coefficients beyond the

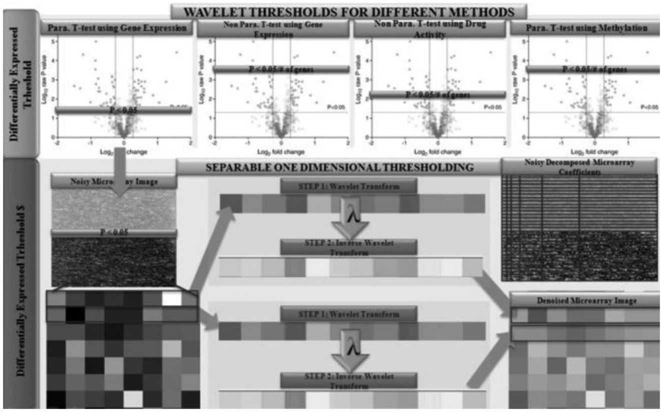

WAVELET THRESHOLDS FOR DIFFERENT METHODS
Differentially Expressed Trheshold
Para. T-test using Gene Expression
Non Para. T-test using Gene Expression
Non Para. T-test using Drug Activity
Para. T-test using Methylation
Log10 raw P value
Log2 fold change
P<0.05
P<0.05/# of genes
Differentially Expressed Trheshold $
Noisy Microarray Image
P<0.05
SEPARABLE ONE DIMENSIONAL THRESHOLDING
STEP 1: Wavelet Transform
λ
STEP 2: Inverse Wavelet Transform
Noisy Decomposed Microarray Coefficients
Denoised Microarray Image

Fig. 6. Wavelet-based thresholding using differentially expressed genes.

determined threshold generated by median absolute deviation MAD noise variance estimation.

VI. Sustainment of Inter- and Intragenomic Relationships Throughout the Transformation and Denoising Process

In order to test the fortification and safekeeping of inter- and intragenomic relationships that are expressed by specific correlations between both genes and samples throughout the transformation and denoising process, we utilized the following techniques:

1) Determine thresholds based upon the point at which genes were expressed significantly for each perspective GCM, CCLE, and TCGA dataset.
2) Evaluate the magnitude of genes that were differentially expressed or significant throughout (before and after) the transformation and denoising process.
3) Conduct a canonical correlation analysis (CCA) to examine specific gene and sample correlations between the 1-D, 2-D, and thresholding methodologies.
4) Gauge classification accuracies throughout the wavelet decomposition and thresholding process in order to utilize accuracy as a mechanism to determine genomic relationship sustainment.

Figs. 7 and 8 depict a CCA plot of the 218 samples (right) and ~8934 genes in the GCM dataset. The proposed CCA strives to plot the cross-correlation matrix between the original 1-D and 2-D denoised expression and samples thresholded by the differentially expressed MAD. This provides a mechanism, which allows the visual representation of the root-mean-squared error between values (comparison 1), covariance (comparison 2), and correlation (comparison 3) between the two datasets. The CCA plots of GCM genes and samples shown in Figs. 7 and 8 depict the genomic relationships that exist between samples. In Fig. 8, the CCA shows the separable 1-D transformation technique, which utilized MAD thresholding in which the denoised expression (light) is more highly correlated with the original GCM sample expression (dark). In addition, this shows that each of the sample's gene expression has a small error between the original and denoised values for the separable 1-D technique. Fig. 9 depicts a CCA for the results of the traditional 2-D technique on GCM samples.

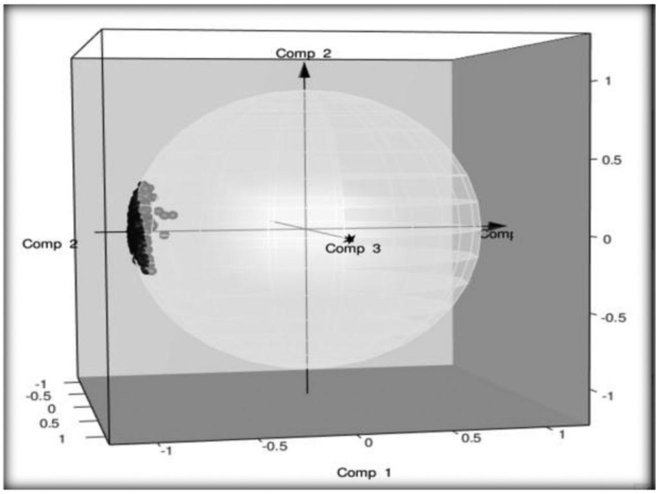


Fig. 7. Separable 1-D wavelet decomposition of GCM samples using MAD/STD thresholding for denoising at level 6 (light) compared to original GCM gene expression (dark).

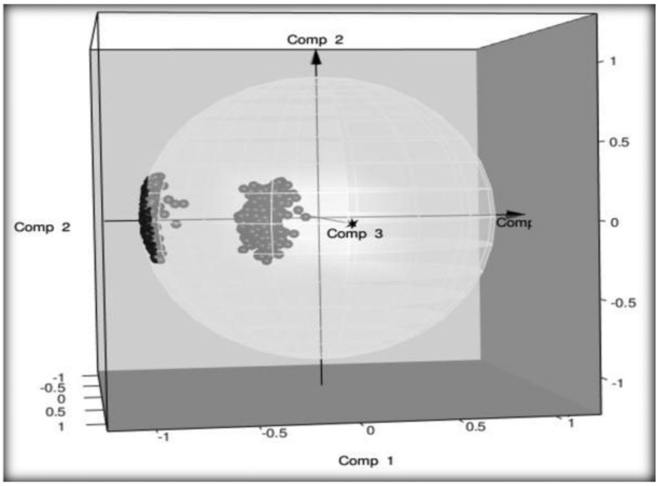
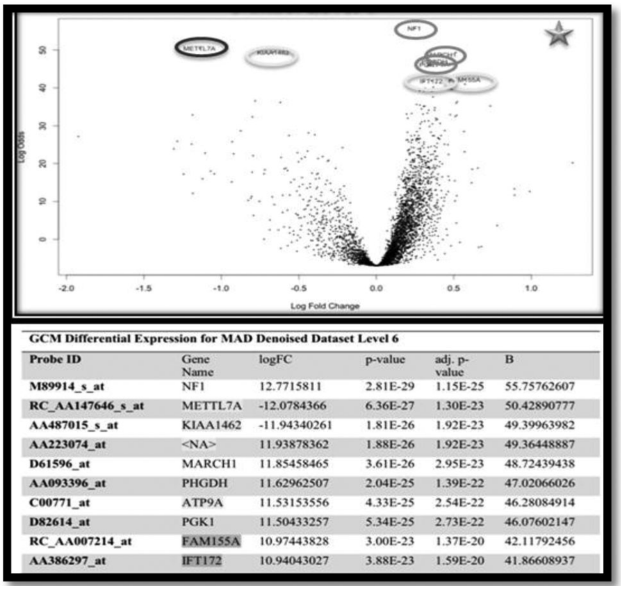


Fig. 8. Traditional 2-D wavelet decomposition of GCM samples using MAD/STD thresholding for denoising at level 6 (light) compared to original GCM gene expression (dark).

The plot shows that the GCM expression samples are not highly correlated and have low variance (comparison 2) because of the error or distance between the samples. Furthermore, this shows that there was a large difference in the intensity values represented in the 1-D MAD after denoising versus the original GCM intensity values in the expression before the denoising process. This pattern of high correlation and low error was apparent for genes and samples in each of the three datasets including GCM, CCLE, and TCGA generated by the separable 1-D technique.

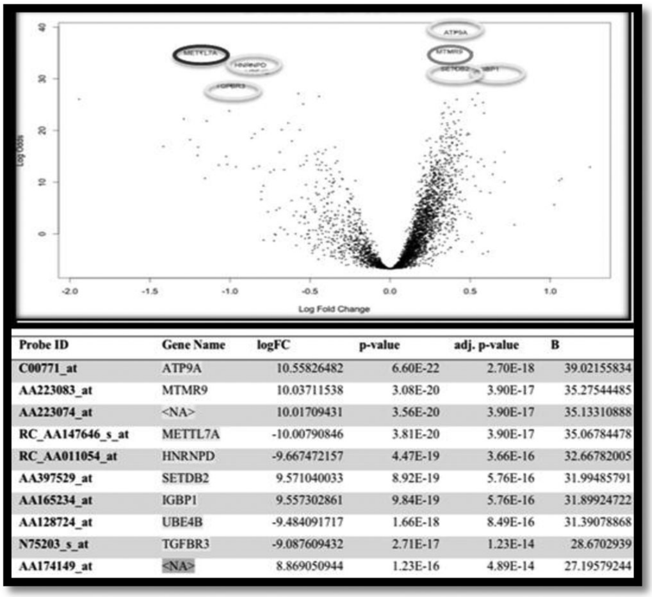
In the next technique, the goal was to generate the GCM top differentially expressed genes produced by the separable 1-D technique and compare those to the top differentially expressed genes produced by the traditional 2-D technique for denoising. First, we generated a listing of the top differentially expressed gene values and compared those gene values before and af-



GCM Differential Expression for MAD Denoised Dataset Level 6

Probe ID	Gene Name	logFC	p-value	adj. p-value	B
M89914_s_at	NF1	12.7715811	2.81E-29	1.15E-25	55.75762607
RC_AA147646_s_at	METTL7A	-12.0784366	6.36E-27	1.30E-23	50.42890777
AA487015_s_at	KIAA1462	-11.94340261	1.81E-26	1.92E-23	49.39963982
AA223074_at	<NA>	11.93878362	1.88E-26	1.92E-23	49.36448887
D61596_at	MARCH1	11.85458465	3.61E-26	2.95E-23	48.72439438
AA093396_at	PHGDH	11.62962507	2.04E-25	1.39E-22	47.02066026
C00771_at	ATP9A	11.53153556	4.33E-25	2.54E-22	46.28084914
D82614_at	PGK1	11.50433257	5.34E-25	2.73E-22	46.07602147
RC_AA007214_at	FAM155A	10.97443828	3.00E-23	1.37E-20	42.11792456
AA386297_at	IFT172	10.94043027	3.88E-23	1.59E-20	41.86608937

Fig. 9. GCM differential expression for MAD traditional 2-D denoised genes at level 6. *Dark Gray*—newly identified differentially expressed genes. *Red Black*—genes that lost their significance through denoising. *Yellow Light Gray*—genes that are present and highly significant in the original and denoised datasets.



Probe ID	Gene Name	logFC	p-value	adj. p-value	B
C00771_at	ATP9A	10.55826482	6.60E-22	2.70E-18	39.02155834
AA223083_at	MTMR9	10.03711538	3.08E-20	3.90E-17	35.27544485
AA223074_at	<NA>	10.01709431	3.56E-20	3.90E-17	35.13310888
RC_AA147646_s_at	METTL7A	-10.00790846	3.81E-20	3.90E-17	35.06784478
RC_AA011054_at	HNRNPD	-9.667472157	4.47E-19	3.66E-16	32.66782005
AA397529_at	SETDB2	9.571040033	8.92E-19	5.76E-16	31.99485791
AA165234_at	IGBP1	9.557302861	9.84E-19	5.76E-16	31.89924722
AA128724_at	UBE4B	-9.484091717	1.66E-18	8.49E-16	31.39078868
N75203_s_at	TGFBR3	-9.087609432	2.71E-17	1.23E-14	28.6702939
AA174149_at	<NA>	8.869050944	1.23E-16	4.89E-14	27.19579244

Fig. 10. GCM differential expression for MAD separable 1-D denoised genes at level 6. *Dark Gray*—newly identified differentially expressed genes. *Black*—genes that lost their significance through denoising. *Light Gray*—genes that are present and highly significant in the original and denoised datasets.

ter 1-D or 2-D wavelet decomposition and MAD denoising as seen in Figs. 9 and 10. The separable 1-D method allowed for more genes to remain consistent between original and denoised, enabling genes which were significantly expressed before and after denoising to remain highly expressed throughout the process (see Fig. 10). The traditional 2-D method did not allow for genes to remain consistent throughout the denoising process

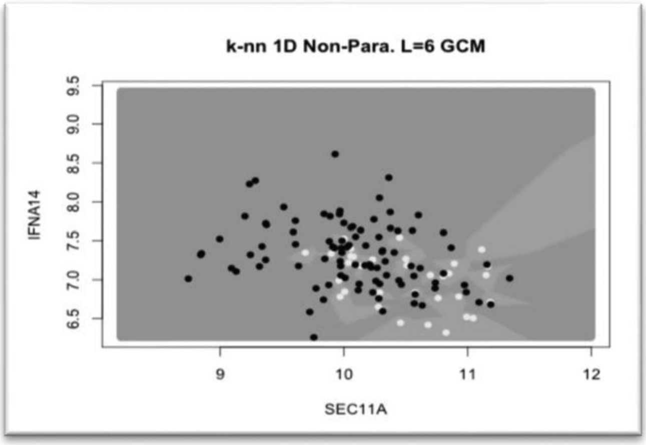


Fig. 11. Classification a nonparametrically thresholded at level 6 expression set features produced by separable 1-D wavelet-based thresholding. Classification boundaries presented by *k*-NN on the plane dictated by two genes in the expression set.

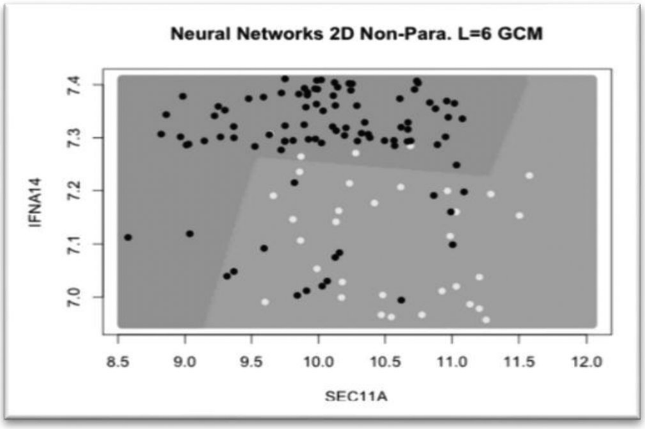


Fig. 12. Classification a nonparametrically thresholded at level 6 expression set features produced by traditional 1-D wavelet-based thresholding. Classification boundaries presented by neural networks on the plane dictated by two genes in the expression set.

and caused genes that were significantly expressed before to lose their significance after the denoising process (see Fig. 9).

1) Genes highlighted in light gray depict genes that are highly expressed in the original and denoised datasets throughout the denoising process.
2) Genes highlighted in dark gray depicts newly identified differentially expressed genes.
3) Genes that are highlighted in black depict genes that lost their significance throughout the denoising process.

VII. Results

In this section, classification results are presented. The results include denoised expression sets for 1-D and 2-D decomposition as well as all wavelet levels (i.e., levels 1–8) for MAD, parametric and nonparametric t-test thresholding. The expression sets were classified using multiple R/Bioconductor *MLInterfaces* machine learning algorithms and the results were compared. We utilized Monte–Carlo cross validation, where *k* hold-out (HO)

TABLE I
SEPARABLE 1-D WAVELET THRESHOLDING GENE CLASSIFICATION FOR GCM

GCM	Decomp. Level	Thresholding Method	Correct Normal Prediction	Incorrect Tumor Prediction	Incorrect Normal Prediction	Correct Tumor Precdiction	Original Dataset Error	Denoised Dataset Error	Monte Carlo
KNN	None	None	46	21	41	148	0.1867187		82%
KNN	5	1-D MAD/STD	66	1	4	185	0.18671	0.032031	97%
KNN	3	1-D Nonparametric	50	17	21	168	0.18671	0.133238	87%
KNN	6	1-D MAD/STD	54	13	11	178	0.18671	.0.04019	96%
KNN	6	1-D Nonparametric	52	15	19	170	0.18671	0.13710	87%
NNets	None	None	36	31	11	187	0.1524004		85%
NNets	5	1-D MAD/STD	61	6	3	186	0.18671	0.035156	97%
NNets	3	1-D Nonparametric	49	18	14	175	0.18671	0.115234	89%
NNets	6	1-D MAD/STD	64	3	6	183	0.18671	0.032812	97%
NNets	6	1-D Nonparametric	62	5	10	179	0.18671	0.057031	95%
SVM	None	None	31	36	14	175	0.1996094		81%
SVM	5	1-D MAD/STD	60	7	6	182	0.18671	0.045735	96%
SVM	3	1-D Nonparametric	54	13	8	181	0.18671	0.101608	90%
SVM	6	1-D MAD/STD	53	14	6	183	0.18671	.0.076562	93%
SVM	6	1-D Nonparametric	40	26	2	187	0.18671	0.063714	94%

a. Gene classification statistics for the separable 1-D wavelet thresholding technique.

TABLE II
TRADITIONAL 2-D WAVELET THRESHOLDING GENE CLASSIFICATION FOR GCM

GCM	Decomp. Level	Thresholding Method	Correct Normal Prediction	Incorrect Tumor Prediction	Incorrect Normal Prediction	Correct Tumor Precdiction	Original Dataset Error	Denoised Dataset	Monte Carlo
KNN	None	None	46	21	41	148	0.1867187		82%
KNN	2	2-D MAD/STD	55	12	11	178	0.18671	0.129336	88%
KNN	6	2-D Nonparametric	47	20	16	173	0.18671	0.164453	84%
KNN	6	2-D MAD/STD	54	13	23	166	0.18671	0.183282	82%
KNN	8	2-D Nonparametric	45	22	36	153	0.18671	0.207812	80%
NNets	None	None	36	31	11	187	0.1524004		85%
NNets	2	2-D MAD/STD	46	21	16	173	0.18671	0.148375	86%
NNets	6	2-D Nonparametric	41	26	16	173	0.18671	0.166796	84%
NNets	6	2-D MAD/STD	30	37	14	175	0.18671	0.199609	81%
NNets	8	2-D Nonparametric	30	37	15	174	0.18671	0.201562	80%
	None	None	31	36	14	175	0.1996094		81%
SVM	2	2-D MAD/STD	52	15	4	185	0.18671	0.113281	89%
SVM	6	2-D Nonparametric	40	27	8	181	0.18671	0.128175	88%
SVM	6	2-D MAD/STD	3	64	2	187	0.18671	.0.190625	81%
SVM	8	2-D Nonparametric	35	32	8	181	0.18671	0.187236	82%

b. Gene classification statistics for the traditional 2-D wavelet thresholding technique.

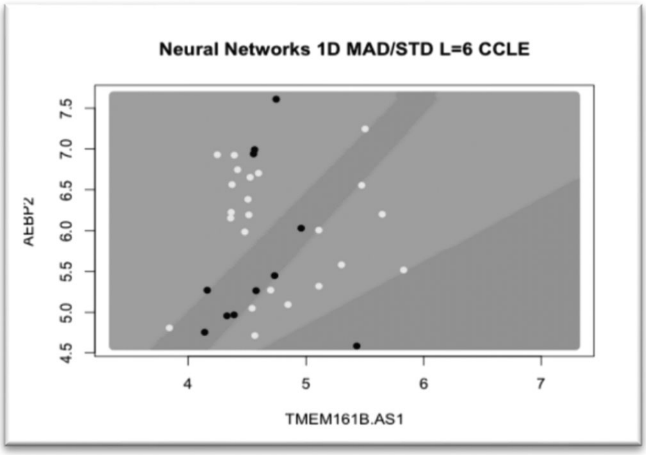


Fig. 13. Classification of CCLE data MAD thresholded at level 6. Expression set features produced by separable 1-D wavelet-based thresholding. Classification boundaries presented by neural nets on the plane dictated by two genes in the expression set.

validation sets were randomly drawn. This included 100 simulations of random realizations to make statistically meaningful conclusions.

For the classification accuracies associated with the original image without wavelet transformation and thresholding, we received upwards of around ∼85% accuracy. When the traditional 2-D wavelet thresholding methods were applied in Table I, we were only able to attain upwards of 89% accuracy using Monte–Carlo k-HO cross validation for each of the three classification methods. When the separable 1-D method was applied in Table II, this enabled the achievement of upwards of 97% accuracy using Monte–Carlo k-HO cross validation for the three classification methods. MAD thresholding had similar accuracies compared to the nonparametric t-test thresholding methodology because of the normality of the data. The separable nonparametric thresholding method was able to attain classification accuracies upwards of 96% accuracy for the 1-D separable method compared to 88% for thresholding when using the traditional 2-D method.

TABLE III
SEPARABLE 1-D WAVELET THRESHOLDING GENE CLASSIFICATION FOR CCLE

CCLE	Decomposition Level	Thresholding Method	Correct Normal	Incorrect Tumor	Incorrect Normal	Correct Tumor	Original Dataset Error	Denoised Dataset Error	Monte Carlo 10-k Cross Validation
Neural Nets	None	None—Original Data	48	8	3	26	0.1294118		88%
Neural Nets	6	1-D MAD/STD	47	0	3	20	0.0714285	0.0428571	96%
Neural Nets	6	1-D MAD/STD	47	0	3	20	0.0714285	0.0428571	96%
Neural Nets	6	1-D Nonparametric	47	0	3	20	0.0714285	0.0714285	93%
Neural Nets	6	1-D Nonparametric	46	1	3	20	0.0714285	0.0571428	95%
Neural Nets	4	1-D Int. Nonparametric	46	1	4	19	0.0714285	0.0714285	93%
Neural Nets	4	1-D Int. Nonparametric	47	0	3	20	0.0714285	0.0714285	93%

TABLE IV
TRADITIONAL 2-D WAVELET THRESHOLDING GENE CLASSIFICATION FOR CCLE

CCLE	Decomposition Level	Thresholding Method	Correct Normal Prediction	Incorrect Tumor Prediction	Incorrect Normal Prediction	Correct Tumor Prediction	Original Dataset Error	Denoised Dataset Error	Monte Carlo 10-k Cross Validation
Neural Nets	6	2-D MAD/STD	47	0	5	18	0.18671	0.0714285	93%
Neural Nets	6	2-D MAD/STD	47	0	5	18	0.18671	0.115234	93%
Neural Nets	6	2-D Nonparametric	46	1	4	19	0.18671	0.0714285	93%
Neural Nets	6	2-D Nonparametric	46	1	4	19	0.18671	0.035156	93%
Neural Nets	4	2-D Int. Nonparametric	46	4	1	19	0.18671	0.115234	93%
Neural Nets	4	2-D Int. Nonparametric	46	4	1	19	0.18671	0.115234	93%

TABLE V
SEPARABLE 1-D WAVELET THRESHOLDING GENE CLASSIFICATION FOR TCGA

TCGA	Decomposition Level	Thresholding Method	Normal Tumor Error	Tumor Tissue Error	Recurrent Tumor Error	Mean Error	Monte Carlo 10-k Cross Validation
SVM	None	None	70%	0.6%	96%	0.0933854	90%
SVM	6	1-D MAD/STD	87%	0.6%	96%	0.0559895	94%
SVM	6	1-D MAD/STD	41%	0.8%	95%	0.0254524	98%
SVM	6	1-D Parametric	37%	0.6%	98%	0.0481770	96%
SVM	6	1-D Parametric	53%	0.2%	98%	0.0273755	98%
SVM	6	1-D Nonparametric	62%	0.4%	98%	0.0440356	96%
SVM	6	1-D Nonparametric	53%	0.2%	98%	0.04302413	96%
SVM	4	1-D Int. Nonparametric	98%	0.5%	47%	0.04296875	96%
SVM	4	1-D Int. Nonparametric	37%	0.6%	98%	0.0331825	97%

TABLE VI
TRADITIONAL 2-D WAVELET THRESHOLDING GENE CLASSIFICATION FOR TCGA

TCGA	Decomposition Level	Thresholding Method	Normal Tumor Error	Tumor Tissue Error	Recurrent Tumor Error	Machine Learning Algorithm	Monte Carlo 10-k Cross Validation
SVM	6	2-D MAD/STD	70%	0.6%	96%	0.05338542	94%
SVM	6	2-D MAD/STD	37%	0.6%	98%	0.04294872	96%
SVM	6	2-D Parametric	37%	0.6%	98%	0.04817708	96%
SVM	6	2-D Parametric	62%	0.4%	98%	0.04498492	96%
SVM	6	2-D Nonparametric	41%	0.8%	95%	0.01260377	97%
SVM	6	2-D Nonparametric	53%	0.2%	98%	0.04886878	96%
SVM	4	2-D Int. Nonparametric	98%	0.7%	78%	0.05175781	94%
SVM	4	2-D Int. Nonparametric	37%	0.6%	98%	0.03706637	97%

Tables III and IV show the classification analysis of CCLE by neural networks, and Tables V and VI show the analysis of TCGA datasets using SVM package of *MLInterfaces* [4]. For the classification accuracies associated with the original CCLE microarray image without denoising, we achieved performance around 88% (>88%) accuracy. When the traditional 2-D wavelet thresholding methods were applied in Table III, we were only able to attain performance of 93% accuracy using Monte–Carlo k-HO cross validation at decomposition level 6 and level 4. When the separable 1-D method was applied in Table IV, performance rose to around 96% (>96%) accuracy using Monte–Carlo k-HO cross validation and the *MLInterfaces* neural network classification methods. For the classification accuracies associated with the original TCGA microarray image without denoising, we attained performance of ∼90% accuracy. When the traditional 2-D wavelet thresholding and denoising methods were applied to the TCGA data in Table V, we were able to attain performance of 97% (>97%) accuracy using Monte–Carlo *k*-HO cross validation at decomposition level 4. When the separable 1-D method was applied in Table VI, this enabled the performance of 97% (>97%) accuracy as well. We similarly utilized Monte–Carlo k-HO cross validation using the *MLInterfaces* neural network classification methods for the TCGA data.

VIII. Conclusion

For the wavelet-based thresholding of the GCM data, the separable 1-D technique took about 400.29 s to be computed in the cloud environment and 1000.53 s in the CSDP environment. In the analysis of cloud-scale wavelet decomposition and thresholding being applied to the GCM, the traditional 2-D denoised datasets classified by machine learning never achieved over 88% accuracy using SVM, NN, and *k-NN* classification methods. The results also showed that the novel separable 1-D denoising method achieved the highest overall classification accuracy using *k-NN* and neural nets for STD/MAD thresholded data at 97%. In the analysis of the cloud-scale wavelet decomposition and thresholding method applied to the CCLE dataset for classification by neural networks, the original datasets accuracy never increased over 88%. When separable 1-D technique and MAD/STD thresholding was applied, the accuracy increased to 96%. The highest accuracy achieved for the traditional 2-D method was less superior at around 93%. Overall, the analysis of the GCM and CCLE dataset generated by denoising allowed genes to keep their significance levels through the denoising process so that classification accuracies would remain consistent after denoising. In the analysis of cloud-scale wavelet decomposition and thresholding applied to the TCGA dataset, original datasets accuracy never increased over 90%. When separable 1-D technique and MAD/STD thresholding was applied, the accuracy increased to 97%. The highest accuracy achieved for the traditional 2-D method was similar to that of the separable technique at around 97%. The inter- and intragenomic sustainment testing with CCA correlation/error calculations and gene differential expression comparison proved that the datasets generated by the denoising method enabled genes to keep their significance levels through the denoising process so that classification accuracies would be optimal. The proposed novel separable 1-D method is more robust than traditional 2-D wavelet-based thresholding and can be potentially used to identify genes, which have the same patterns or biological processes in similar datasets. The 1-D wavelet method is more computationally expensive than other GSP methods, but was optimized through the CSDP environment. The novel methodology within this research utilized wavelet-based thresholding for denoising of gene expression on a gene-by-gene basis in a cloud environment allowing for higher accuracies and sustainment of genomic relationships, furthermore retaining important gene patterns, which inform us of biological processes. Ultimately, this study will help to face the present and forthcoming challenges that may arise in the large-scale analysis of genetics and genomics data.

References

[1] B. Harvey and S.-Y. Ji, “Cloud-scale genomic signals processing classification analysis for gene expression microarray data,” in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2014, pp. 1843–1846.

[2] R. S. Istepanian, A. Sungoor, and J.-C. Nebel, “Comparative analysis of genomic signal processing for microarray data clustering,” *IEEE Trans. NanoBiosci.*, vol. 10, no. 4, pp. 225–238, Dec. 2011.

[3] S. Ramaswamy *et al.*, “Multiclass cancer diagnosis using tumor gene expression signatures,” *Proc. Nat. Acad. Sci.*, vol. 98, pp. 15149–15154, 2001.

[4] R. C. Gentleman *et al.*, “Bioconductor: Open software development for computational biology and bioinformatics,” *Genome Biol.*, vol. 5, p. R80, 2004.

[5] J. Barretina *et al.*, “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity,” *Nature*, vol. 483, pp. 603–607, 2012.

[6] J. N. Weinstein *et al.*, “The Cancer Genome Atlas Pan-Cancer analysis project,” *Nature Genetics*, vol. 45, pp. 1113–1120, 2013.

[7] V. Prajapati, *Big Data Analytics with R Hadoop*. Birmingham, U.K.: Packt Publishing, 2013.

[8] D. L. Donoho, “De-noising by soft-thresholding,” *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

Authors’ photographs and biographies not available at the time of publication.