Integrated Analysis of Gene Expression and Copy Number Data using Sparse Representation Based Clustering Model.

Conference Paper · January 2011

Source: DBLP

2 authors, including:

Hongbao Cao
National Institutes of Health
44 PUBLICATIONS **163** CITATIONS

SEE PROFILE

BICoB(2011): 172-177

INTEGRATED ANALYSIS OF GENE EXPRESSION AND COPY NUMBER DATA USING SPARSE REPRESENTATION BASED CLUSTERING MODEL

Hongbao Cao
Department of Biomedical Engineering
Tulane University
New Orleans, LA, 70118 USA
caohon2010@gmail.com

Yu-ping Wang
Department of Biomedical Engineering
Tulane University
New Orleans, LA, 70118 USA
wyp@tulane.edu

ABSTRACT

Among biological measurements, DNA microarray gene expression and array comparative genomic hybridization (aCGH) have been widely used. Due to the vast information of the biological data, various clustering techniques have been developed to identify subsets of genes with specific gene expression patterns and large variations across samples. Since integrated analysis of genomic data from different sources can further increase the reliability of biological analysis results, methods of integrating and analyzing different types of genomic measurements have emerged. In this work, we jointly examine gene expression and copy number data and iteratively project the data on different clusters through the sparse representation based clustering (SRC) model. Our method has been tested on a breast cancer cell lines data set and a breast tumors data set. In addition, simulated data sets were used to test the robustness of the method to noise. Experiments showed that our proposed method can effectively identify genes with large variations in gene expression and copy number, and locate genes that are statistically significant in both measurements. The proposed method can be applicable to a wide variety of biological problems where joint analysis of biological measurements is a common challenge.

Index Terms— Sparse representations, DNA microarray, gene expression, gene copy number

1 INTRODUCTION

To our present knowledge, there are approximately 20,000 to 25,000 protein-coding genes for human genome [1]. Techniques for analyzing those molecular level data including: serial analysis of gene expression (SAGE) [2], oligonucleotide arrays [3], cDNA microarrays [4], and array comparative genomic hybridization (aCGH)[5]. Among those techniques, cDNA microarrays can provide rapid, parallel surveys of gene-expression patterns for thousands of genes in a single assay. During the progress of some complex diseases, such as cancers, many aberrations at the molecular level may occur in the form of gene over/under expression [6], which has been studied extensively with microarrays [7]. Gene expression study in cancers may provide new insights in cancer development and facilitate directing therapeutic interventions [8]. It was reported that, for breast cancers, new molecular sub-class can be identified through gene expression profiles analysis [9].

However, aberrant gene expression patterns may be caused by other process independent of tumorigenesis. Therefore, integrating data from other sources, such as DNA, gene, and protein level data, may be necessary for increasing the reliability of the analysis results. Previous study showed that gene copy number variations (CNV) play a prominent role in cancer pathogenesis [10]. Currently, aCGH with submegabase levels resolution [5] can be used to localize CNVs associated with breast cancers [11], [12]. It is reported that breast cancer sub-classification can be done utilizing CNVs [13].

Since the cause of the disease is very complex [14][15], utilizing one type of data alone may cause inaccuracy in analysis results. Furthermore, for several cancers, the copy number and gene expression data are highly correlated [16]. Thus by joint study of gene expression and copy number data, the false positives can be reduced, which will in turn increase the analysis accuracy [16]. Currently, there are only a few studies that have performed genome-wide analysis by combining both data types for breast cancer [14][17][18]. In [18], Berger et. al. proposed a generalized singular value decomposition (GSVD) method that is able to analyze both gene expression and copy number data to locate genes showing both high variation across the samples and correlation across the genes. The GSVD method they proposed assumes that both inputs have similar distributions, and data distribution transformation has to be made before the analysis.

In recent years, sparse representations of signals/images have received a great deal of attentions in applied mathematics and signal processing community [19]~[22]. Although the sparse representations have been used in many fields, to our knowledge, little work exists on their use for analyzing genomic data. In this work, we applied the sparse representation clustering (SRC) model to jointly analyze gene expression and copy number data. The SRC algorithm was obtained by L1-minimization using Homotopy method

[23]. The Homotopy method was originally proposed by Osborne et al. for solving noisy over-determined L1-penalized least square problem [24]. Donoho et al. [23] applied it to solve the noiseless underdetermined L1-minimization problem

$$\text{(P1)}\ \hat{x} = \text{argmin}\|x\|_1 \text{ subject to } Ax = y, \quad (1)$$

and showed that Homotopy runs much more rapidly than general-purpose linear programming (LP) solvers when sufficient sparsity is present.

In our proposed method, there is no requirement for data distributions. The only thing needed for our proposed method is, which is also required by the GSVD method, a simple data scale transformation so that the dynamic ranges of feature vectors extracted from both input data types have similar intervals. Moreover, the SRC approach can process two types of data either separately or jointly, and select the genes with significant variances. In addition, our methods is able to locate genes with more characteristics to satisfy different requirements, such as finding genes with high L-1 norm across genes, high correlation across the genes, high/low variation across the samples, and finding genes with specific value distributions (for example, genes with at least one third big samples), etc. With those different features, more thorough analysis of the data is available. After introducing the SRC based algorithm, we examine its robustness to noise and its convergence properties using simulated data. Then, we tested our method on the breast cancer cell line data [25] and breast tumors data [26]. Results showed that our proposed method can effectively locate genes with significant variances when analyzing gene expression data and copy number data separately. Moreover, our proposed method can identify those highly correlated genes with significant variance in both gene expression and copy number data, which is essential for the joint analysis of different genomic data.

2. METHODS

2.1 Feature selection and scale transformation

Feature selection addresses the characteristic of genes to be located. For example, if we try to locate genes with high correlation across the genes and high variation across the samples, as was performed by [14][17][18], we extract, from each gene samples, the variation and the correlation coefficients between the two data sets. On the other hand, we may also interest in genes with big L1-norm across genes and genes with at least n big samples across the data set. In this case we need to extract, from each gene samples, L1-norm and the first n biggest absolute sample values.

Once the protocol of feature selection is set, each given gene will be expected to generate one feature vector for the following analysis. For all the genes in both data set, we will have a feature matrix $V = [V_1, V_2, \dots, V_{nv}]^T \in R^{nv \times p}$, where p is the number of genes/variables, and nv is the number of features for each gene. We use Eq. (2) to transfer the scale of each feature vector $V_i = [v_{1i}, v_{2i}, \dots, v_{pi}]^T \in R^{p \times 1}$ to the rang of [0, 1], where $i = 1, \dots, nv$.

$$V_i = (V_i - min(V_i))/(\max(V_i) - \min(V_i)) \in R^{p \times 1} \quad (2)$$

In this work, we selected different features to perform jointly/separately analysis of different data sets. For signal data set analysis, we selected $n + 2$ features for each gene: the first n biggest absolute sample values, their L-1norm and variances. Scale transformation by Eq. (2) was performed on the extracted feature vectors. In order to fit the requirement of SRC clustering algorithm, we inversed the scale-transformed L1-norms and variances and get feature vectors as is given in equation (3):

$$v_k = [v_{k1}, v_{k2}, \dots, v_{kn}, 1 - \bar{v}_k, 1 - var(v_k)]^T \in R^{(n+2) \times 1} \quad (3)$$

Where $k = 1{:}p$; v_i are i-th biggest absolute values of a given gene's samples (scale transformation was performed so that $v_i \in [0,1]$), and $\bar{v}$ and $var(v)$ are the mean and variance of v_i respectively, $i = 1{:}n$.

For the joint data analysis, we selected $2n + 5$ features for each gene: the first n biggest absolute sample values from each data set, their L1-norms and variances, and Pearson correlation coefficients between samples from the two data sets, which gives feature vectors as following:

$$Jv_k = [v_{1k}, v_{2k}, cor_k]^T \in R^{(2n+5) \times 1} \quad (4)$$

Where Jv_k is the joint feature vector for the k-th gene, $k = 1{:}p$; v_{1k} and v_{2k} are the feature vectors from the first and second data sets respectively, which are extracted in the way stated by Eq. (3); cor_k is the Pearson correlation coefficients between samples of the k-th gene from the two data sets.

2.2 Cluster design and gene shaving

Following the feature selection protocol, a feature vector was generated from each gene. According to the different values of feature vectors, genes can be classified into different clusters and those that do not fall into clusters with interests can be shaved off. This is the sparse representation based clustering (SRC) based gene selection process, or so called gene shaving. To better understand the SRC based gene shaving method, we give an example as following:

Suppose we extract a feature vector from each gene as $v_k = [v_{k1}, v_{k2}, v_{k3}, 1 - \bar{v}_k]^T$, where v_{ki} are the i-th biggest absolute values of a given gene's samples (copy number data/gene expression data), and scale transformation was performed so that $v_{ki} \in [0,1]$ for $i = 1,2,3$; $\bar{v}_k \in [0,1]$ is the L1-norm of v_{ki} with scale transformation by Eq. (2) . Then

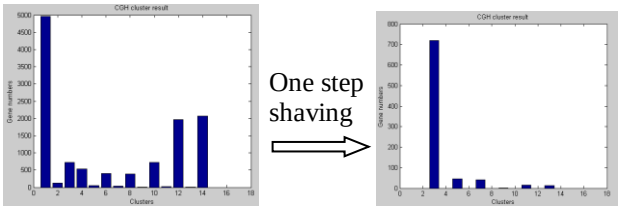
we can design clusters with center vectors $v_c = [v_{c1}, v_{c2}, v_{c3}, 1-\bar{v}_c]^T$ as is shown in Table 1.

Table 1 Example of designed clusters and their center vectors

#of clusters / Feature entries	1	2	3	4	5	6
v_{c1}	1	1	1	1	1	1
v_{c2}	1	1	1	1	0	0
v_{c3}	1	1	0	0	0	0
$1-\bar{v}_c$	1	0	1	0	1	0

With the feature selection protocol, we can easily understand the meanings of each cluster. For example, if a gene data was classified into cluster 2, it means this gene has 3 relative big valued samples with a relative high mean. While if it is classified into cluster 4, it means this gene only has 2 relative big values, and yet the mean is still relatively big, which means those 2 values are even bigger than those in cluster 2. If a gene data is clustered into cluster 6, it means there is only one very strong valued sample in this gene data that is more likely to be an outlier noise. The meaning of other clusters can be explained accordingly. Thus if we want to locate genes with at least 2 big samples with high L1-norm, we should select the genes that are clustered into cluster 2 for the further analysis.

In this work, we perform the gene shaving by using our proposed SRC model for separately and jointly analysis of gene expression and copy number data. Fig. 1 gives an illustration of the gene shaving process using SPC based gene shaving method. As is shown in Fig.1, all genes were firstly projected onto different clusters (Fig.1 (a)). Then the genes that fall out of the clusters of interest will be removed from further analysis (Fig.1 (b)). The process will continue unless the number of genes left meets the requirement.



(a) (b)

Fig. 1 Diagram of gene shaving by SRC

2.3 Character matrix design

In this section, we describe the design of the character matrix A to fit into the sparse linear system $Ax = y$ according to the feature extraction protocol and clusters design protocol we discussed above.

Suppose there are nc clusters, and the feature vector of each gene $v_k \in R^{nv\times 1}$, then we can expand the center vector of the j-th cluster $v_{jc} \in R^{nv\times 1}$ to the character matrix of the cluster as $A_j = [v_{jq}] \in R^{nv\times nv}$ in the following way:

$$\|v_{jq} - v_{jc}\|_\infty < \Delta v \tag{5}$$

where $v_{jq} \in R^{nv\times 1}$, $j = 1{:}\,nc$, $q = 1{:}\,nv$, and $\|*\|_\infty$ is the infinite norm. Δv is a parameter with small positive value to guarantee that there is no overlap among each cluster. In this work, we set Δv=0.3. To guarantee all genes from the j-th cluster be represented by its character matrix A_j, it requires that:

$$|A_j| > 0, \tag{6}$$

where j=1:nc, and $|*|$ is the determinant of a square matrix. In addition, v_{jc} and v_{ic} should be linearly independent for $i \neq j$ so that they are differentiable by the system.

Once the character matrix for each cluster was designed, we denote $A = [A_1, A_2, \ldots, A_{nv}] \in R^{nv\times(nc*nv)}$ as the character matrix for the total nv clusters. Since data vector from j-th cluster can be expressed by its character matrix A_j, then any given data vector $y \in R^{nv\times 1}$ should be represented by A through a linear system

$$Ax = y. \tag{7}$$

In the linear system given by Eq. (7), since y can only belong to one cluster and should be represented by the character matrix of that cluster, the solution $x \in R^{(nc*nv)\times 1}$ is expected to be sparse with none-zero entries focusing on one cluster only, which makes the linear system given by Eq. (7) a sparse system that can be solved in the way solving the L-1 norm minimization problem (P1) give by Eq. (1). In this work, we employed the Homotopy algorithm [23] to solve the system which is disrobed in the following section 2.4.

2.4 Homotopy algorithm for solving (P1)

For the L1-minimization problem (P1) given by (1), it is convenient to consider the unconstrained optimization problem instead:

$$\text{(P2)}\ \hat{x} = argmin\|Ax - y\|_2^2/2 + \lambda\|x\|_1, \tag{8}$$

where λ is a non-negative coefficient. Homotopy method tries to find a pathway, which starts at large λ and $x_\lambda = 0$, and terminates when λ=0 and x_λ converge to the solution of (P1).

Let $f_\lambda(x)$ denote the objective function of (P2). By classical ideas in convex analysis, a necessary condition for x_λ to be a minimizer of $f_\lambda(x)$ is that $0 \in \partial_x f_\lambda(x_\lambda)$, i.e., the zero vector is an element of the subdifferential of f_λ at x_λ. We calculate

$$\partial_x f_\lambda(x_\lambda) = -A^T(y - Ax_\lambda) + \lambda\partial\|x_\lambda\|_1, \quad (9)$$

where $\partial\|x_\lambda\|_1$ is the subgradient:

$$\partial\|x_\lambda\|_1 = \left\{u \in R^n \middle| \begin{matrix} u_i = sgn(x_{\lambda,i}), & x_{\lambda,i} \neq 0 \\ u_i \in [-1,1], & x_{\lambda,i} = 0 \end{matrix}\right\} \quad (10)$$

Let $I = \{i: x_\lambda(i) \neq 0\}$ denote the support of x_λ, and call $c = A^T(y - Ax_\lambda)$ the vector of residual correlations. Then the condition on the gradient expressed in (9) being zeros can be written equivalently as the two conditions:

$$c(I) = \lambda sgn(x_\lambda(I)), \quad (11)$$

and

$$|c(I^c)| \leq \lambda, \quad (12)$$

In other words, residual correlations on the support of I must all have magnitude equal to λ, and signs that match the corresponding elements of x_λ, whereas residual correlations off the support must have magnitude less than or equal to λ. The Homotopy algorithm now follows from these two conditions, by tracing the optimal path x_λ that maintains (11) and (12) for all $\lambda \geq 0$. The key to the successful implementation is that the path x_λ is a piecewise linear path, with a discrete number of vertices .

Homotopy algorithm:

1) Initialize $x_0 = 0$.
2) For the l-th stage ($l = 1,2, \ldots$), compute an update direction d_l by solving

$$A_I^T A_I d_l(I) = sgn(c_l(I)), \quad (13)$$

with d_l set to zero in coordinates not in I, where

$$I = \{j: |c_l(j)| = \|c_l\|_\infty = \lambda\} \quad (14)$$

3) Calculate the residual γ_l^+

$$\gamma_l^+ = min_{i \in I^c}\left\{\frac{\lambda - c_l(i)}{1 - a_i^T v_l}, \frac{\lambda + c_l(i)}{1 + a_i^T v_l}\right\} \quad (15)$$

where $v_l = A_I d_l(I)$, and the minimum is taken only over positive arguments. Record the corresponding index as i^+.

4) Calculate the residual γ_l^-

$$\gamma_l^- = min_{i \in I}\{-x_l(i)/d_l(i)\}, \quad (16)$$

Again the minimum is taken only over positive arguments. Record the corresponding index as i^-.

5) Calculate the residual r_l,

$$r_l = min\{\gamma_l^+, \gamma_l^-\}, \quad (17)$$

6) Update the active set I by either appending I with i^+ or removing i^- from I.

7) Update x_l

$$x_l = x_{l-1} + r_l d_l \quad (18)$$

8) If $\|c_l\|_\infty = 0$, terminate and x_l is the solution of (P1); Otherwise, go back to step 2).

2.5 SRC clustering model

For a given data vector $y \in R^{nv \times 1}$, we decide to which cluster this gene belongs to by the sparse representation-based clustering (SRC) algorithm described as following [27]:

Sparse Representation-based clustering (SRC) algorithm:

1. Inputs: character matrix $A \in R^{nv \times (nc * nv)}$ for nc clusters; and a test sample $y_i \in R^{nv \times 1}$.
2. Normalize the columns of A to have unit L2-norm.
3. Solve the L1-norm minimization problem (P1) defined by Equation (1).
4. Calculate the residuals $r_k(y_i) = \|y_i - A\delta_k(x)\|_2$, $k \in \{1:nc\}$;
5. $Identity\ (y_i) = \arg\min_k r_k(y_i)$

In the algorithm described above, $r_k(y_i)$ is the residual between y_i and $\hat{y}_i$, $k \in \{1:nc\}$ and $\|*\|_2$ represents the L2-norm.

2.6 **Character matrix training**

Before fitting A to the linear system given by Eq. (7) to cluster an un-classified data vector y, columns of A should be tested using the SRC clustering model to make sure that they are mathematically valid vectors for the classification task, which is the process of *character matrix training*.

For the sparse representation based classifier, a valid training vector should have a sparse representation whose nonzero entries concentrate mostly on one subject, whereas an invalid vector has sparse coefficients spread widely among multiple subjects. To quantify this observation, we use the sparsity concentration index (SCI) that was proposed in [27] to measure how concentrated the feature vectors are on a single class in the dataset:

$$\mathrm{SCI(x)} = \frac{ck * \max_i \frac{\|\delta_i(x)\|_1}{\|x\|_1} - 1}{nc - 1} \quad \in [0,1] \quad (3)$$

where nc is the number of classes. For a solution $\hat{x}$ found by the SRC algorithm, if $\mathrm{SCI}(\hat{x}) = 1$, the feature vector y is represented using only vectors from a single class; if $\mathrm{SCI}(\hat{x}) = 0$, the sparse coefficients are spread evenly over all classes. We choose a threshold $\tau \in [0,1]$ and accept a test vector as valid if $\mathrm{SCI}(\hat{x}) > \tau$; otherwise, reject it as invalid.

If a column vector of A is mis-classified, or it is reject as invalid for having low SCI, we have the character matrix of the cluster matrix that the column vector belongs to re-designed following the process give by section 2.3.

In this work, all column vectors of the character matrix A that was generated following the process given by section 2.3 always pass the training process with $\mathrm{SCI} = 1$ and 100% classification ratio. To further verify the validity of character matrixA, SRC based clustering using A to classify the column vectors of 10 other character matrixes was also performed. Those 10 character matrixes were generated with same cluster centers used to generate A. The overall classification ratio>99.5% is set as the pass threshold, otherwise the generated character matrix A will be rejected as invalid and will be redesigned according to the process given by section 2.3.

2.7 Separately/jointly analysis of gene expression and copy number data

Through the analysis of sections 2.1~2.6, we summarize the process to identify subsets of genes through the separately/jointly analysis of genomic data as following:

SRC based gene shaving algorithm:

1. Specify the characteristics of interesting genes.
2. Extract features from genomic data according to the characteristics specified.
3. Perform scale transfer using Eq. (2) so that the dynamic ranges of all feature vectors extracted $\in [0,1]$.
4. Design clusters according to the feature vectors. It should include both clusters of interesting and clusters to be shaved off.
5. Design character matrix A_i for each cluster, and character matrix $A = [A_i]$ for all the clusters, where $i = 1{:}nc$, and nc is the number of clusters. The character matrix A can be saved for the use of other loop.
6. Using SRC method with A to classify y_j into different clusters. Where y_j is the feature vector extracted for the j-th gene data, where $j = 1{:}p$ and p is the number of genes.
7. Select genes that fall into the interesting clusters.
8. If the genes selected are more than needed, go to step 2, otherwise break.

The difference of jointly and separately analysis of genomic data is that, instead of using feature vectors from one genomic data set, we use features from both data sets. As a matter of fact, the proposed SRC genomic data analysis method can be easily generalized to N types of data analysis by simply feeding joint features from N types of data as input of the algorithm.

3. RESULTS

3.1 Test on simulation data

In order to evaluate the robustness of the method to noise, the gene list percentage similarity (PS) was computed by counting the number of genes obtained from noisy data (ND) intersecting with that obtained from the original data (OD) [18].

$$PS = \frac{\#ND \cap \#OD}{\#Tot} \times 100\% \qquad (4)$$

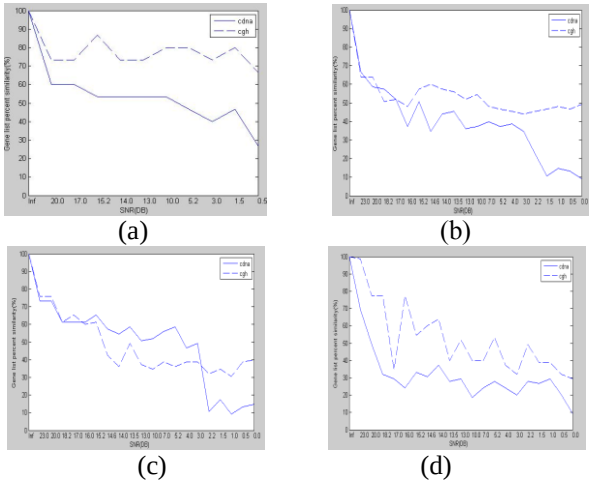
where Tot is the number of total genes in the list.

Fig. 2 gives PS of SRC based gene shaving method on simulated data sets with different signal to noise ratio (SNR). The SNR is defined as following:

$$SNR = 20 * log_{10}(\frac{Var(signal)}{Var(noise)}) \qquad (5)$$

where Var(*) is the variance; noises used in the model are white noises.

For the data sets of Fig.2 (a) and (b), copy number data were generated using the model proposed by Wang et al [28]; Gene expression data were generated based on the model of Attoor et al [29]. For the data sets of Fig.2 (c) and (d), the breast cancer cell lines [25] and breast tumors [26] were used after adding noise with different variance.



(a) (b)

(c) (d)

Fig. 2 *PS* of SRC based gene-shaving method on simulated data sets with different SNR. (a) and (b) are the results from simulated data sets with 100 and 1000 genes respectively. (c) and (d) used real data after adding noise (1000 genes).

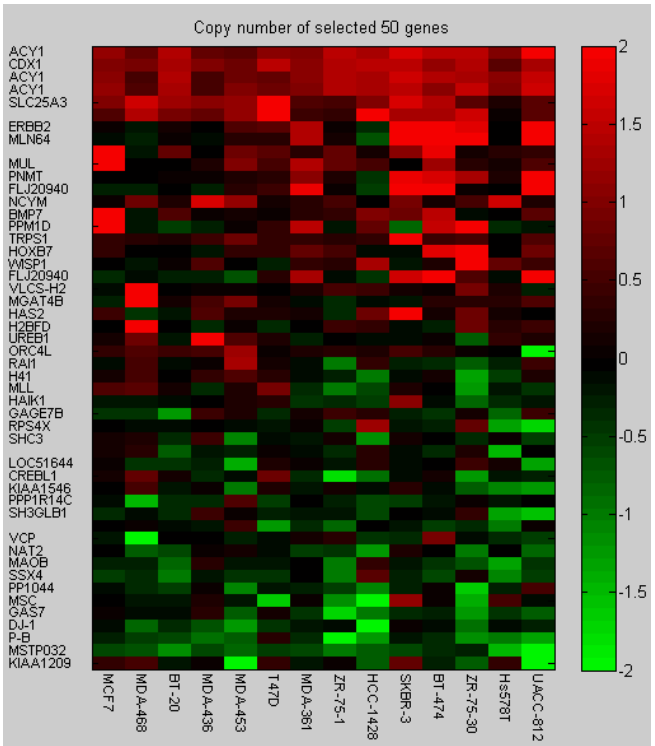
3.2 Breast cancer cell lines data study

In this section, we analyzed a set of breast cancer cell lines data [17], which have 14 cell lines (BT-20, BT-474, HCC-1428, Hs578T, MCF7, MDA-361, MDA-436, MDA-453, MDA-468, SKBR-3, T47D, UACC-812, ZR-75-1, and ZR-75-30), with 11994 genes. It includes both copy number data and gene expression data, and the expression and copy number ratios were log2 -transformed prior to analysis

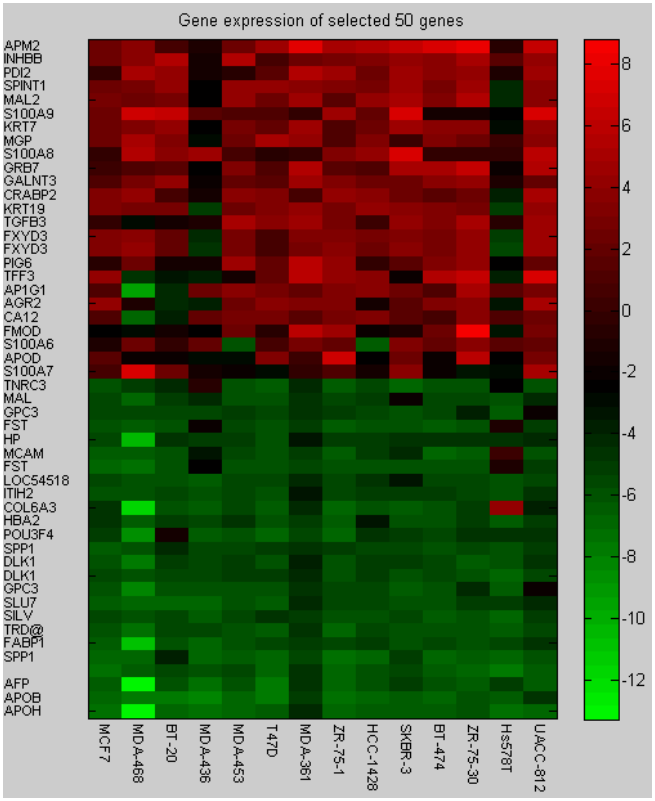
Separately and jointly analysis of copy number data and gene expression data were studies. The goal of this study is to locate 'abnormal' genes that may contribute to the breast cancer.

For separately, we address the 'abnormalities' of interesting genes with high percentage big-valued samples, big L1-norm values for both copy number ratios and gene expression ratios. Genes with those characters have big variance across the other genes and thus should contribute big to the genomic disease as breast cancer.

One of the tasks of the analysis is to select genes with top highest variances, which may be responsible for the disease. In this section, the top highest variant genes selected by using the SPC method proposed in this work were given in Fig. 3 and Fig. 4.

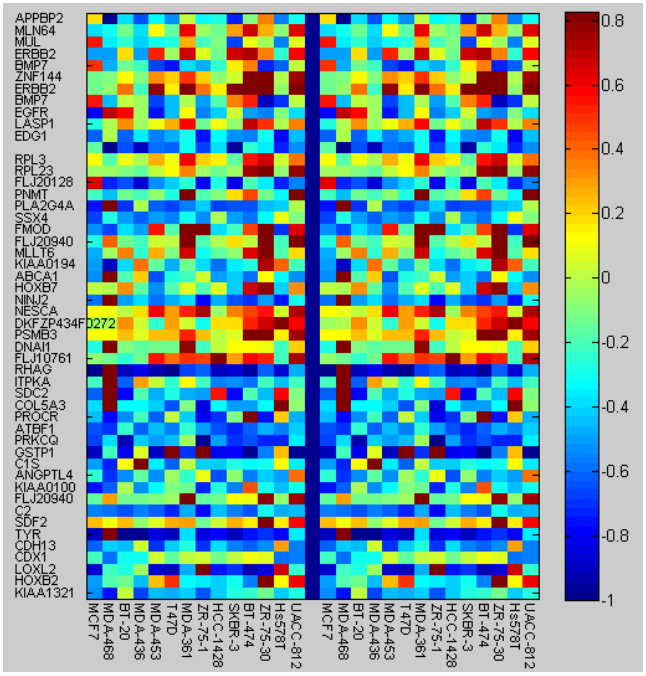


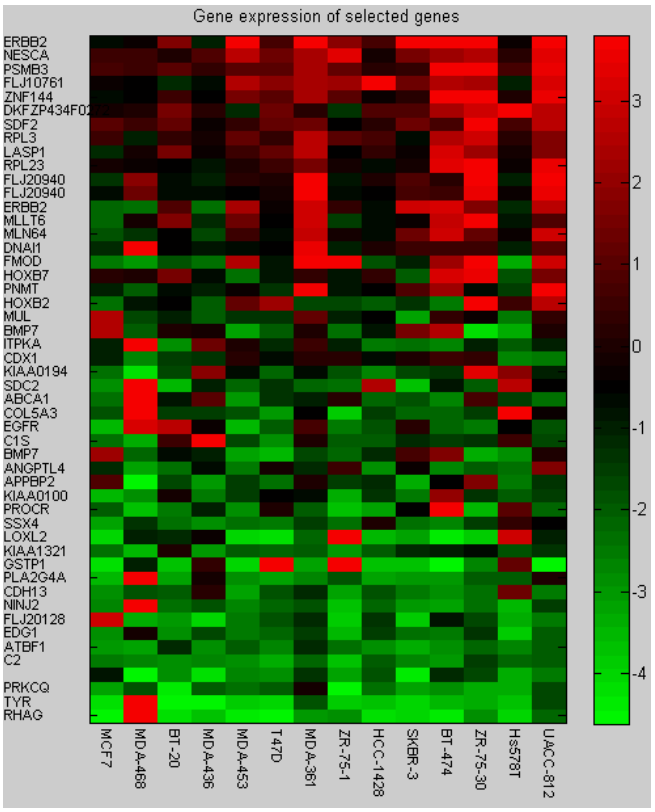
(a)

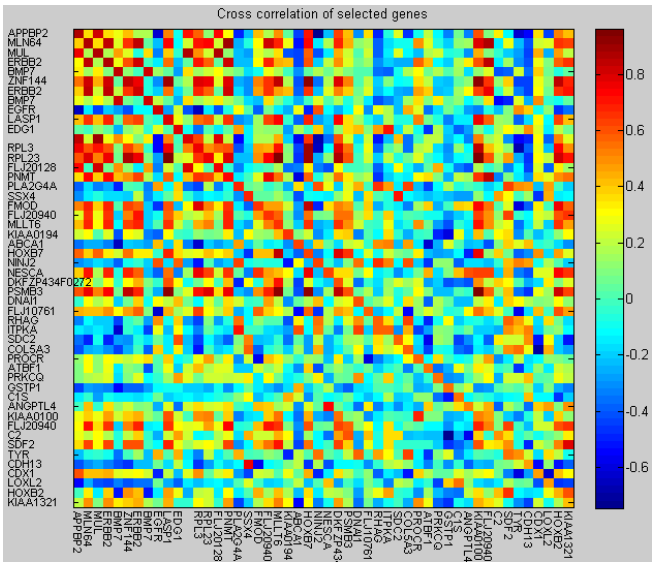


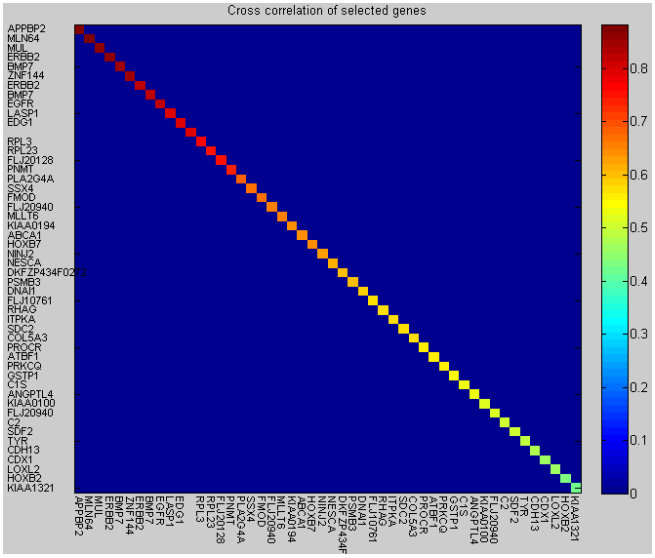
(b)

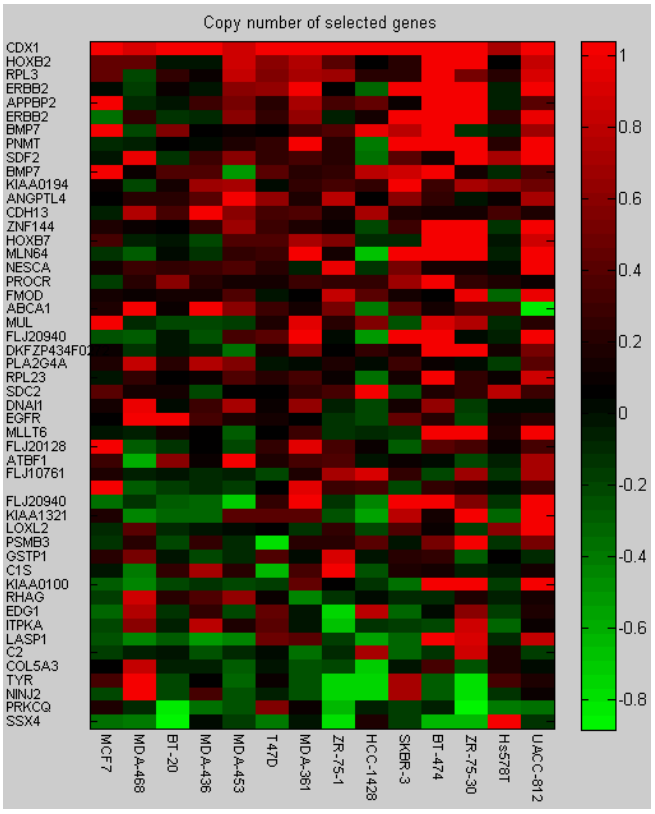
Fig. 3 The top highest variant genes using (a) copy number and (b) gene expression analysis in 14 samples from breast cancer cell lines data

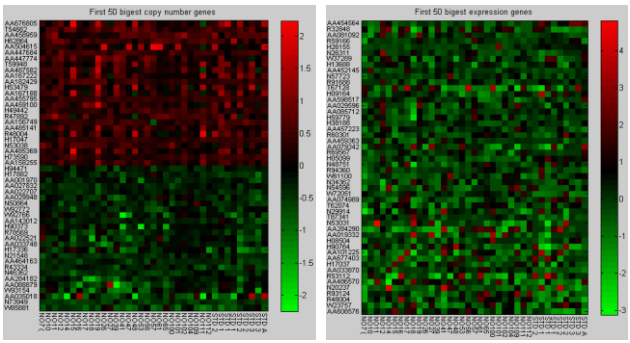



Gene expression of selected genes
ERBB2
NESCA
PSMB3
FLJ10761
ZNF144
DKFZP434F02
SDF2
RPL3
LASP1
RPL23
FLJ20940
FLJ20940
ERBB2
MLLT6
MLN64
DNAI1
FMOD
HOXB7
PNMT
HOXB2
MUL
BMP7
ITPKA
CDX1
KIAA0194
SDC2
ABCA1
COL5A3
EGFR
C1S
BMP7
ANGPTL4
APPBP2
KIAA0100
PROCR
SSX4
LOXL2
KIAA1321
GSTP1
PLA2G4A
CDH13
NINJ2
FLJ20128
EDG1
ATBF1
C2
PRKCQ
TYR
RHAG
3
2
1
0
-1
-2
-3
-4
MCF7
MDA-468
BT-20
MDA-436
MDA-453
T47D
MDA-361
ZR-75-1
HCC-1428
SKBR-3
BT-474
ZR-75-30
Hs578T
UACC-812


Cross correlation of selected genes
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6


Cross correlation of selected genes
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0


Copy number of selected genes
CDX1
HOXB2
RPL3
ERBB2
APPBP2
ERBB2
BMP7
PNMT
SDF2
BMP7
KIAA0194
ANGPTL4
CDH13
ZNF144
HOXB7
MLN64
NESCA
PROCR
FMOD
ABCA1
MUL
FLJ20940
DKFZP434F02
PLA2G4A
RPL23
SDC2
DNAI1
EGFR
MLLT6
FLJ20128
ATBF1
FLJ10761
FLJ20940
KIAA1321
LOXL2
PSMB3
GSTP1
C1S
KIAA0100
RHAG
EDG1
ITPKA
LASP1
C2
COL5A3
TYR
NINJ2
PRKCQ
SSX4
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
MCF7
MDA-468
BT-20
MDA-436
MDA-453
T47D
MDA-361
ZR-75-1
HCC-1428
SKBR-3
BT-474
ZR-75-30
Hs578T
UACC-812



(a) (b)

Fig. 4 The top highest variant genes using (a) copy number and (b) gene expression analysis in 37 samples from breast tumor data

In order to compare the ability of identifying genes with significant variance, a simple comparison was performed between our proposed SRC method and the GSVD method [18] on the absolute value of the first top 50 copy number genes selected from breast cancer cell lines data. Mean ±stds are given for GSVD method and our proposed SRC method as 0.5188±0.2778 and 0. 6949±0.2674 respectively. Fig. 5 gives the box plot of the absolute mean copy number values of the top highest variant genes selected by the two methods: GSVD, and SPC. From Fig. 5 it can be seen that the top highest variant genes selected by using SRC method are better than that of GSVD with higher variant values with p value<10e-5.

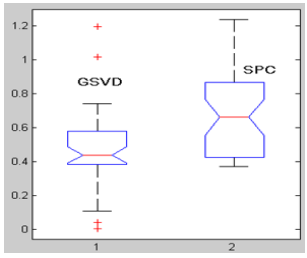
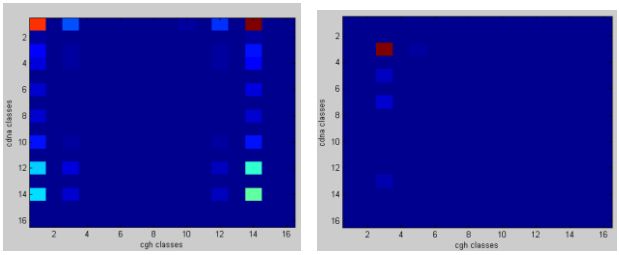


Fig. 5 Box plot of the top highest variant genes selected by two methods

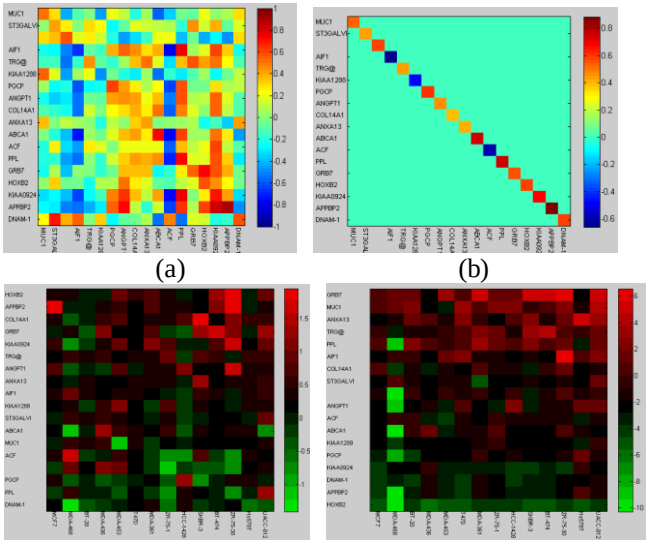
3.3 Selection of genes with correlated high variant aCGH and cDNA data

Another task of the work is to find the genes with correlated high variant aCGH and cDNA data. Those genes have the characteristics of: 1. they have relatively strong variant data value for both CGH and CDNA data. 2. They are highly correlated. Fig. 6 gives the cross matrix of the aCGH data and cDNA data clustering results. Fig. 7 gives selected genes from breast cancer cell line data when the threshold of correlation coefficient is set as 0.4. Fig. 8 gives selected genes from breast tumor data when the threshold of correlation coefficient is set as 0.65.



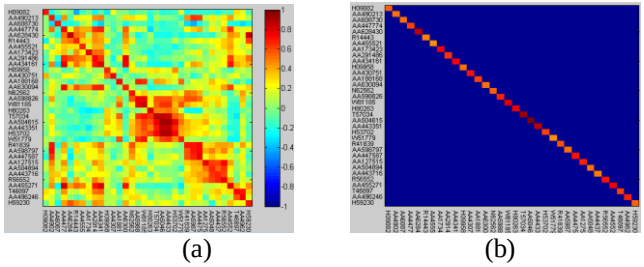
(a) (b)

Fig. 6 Cross matrix of aCGH data and cDNA data clustering results (a) cross matrix for all genes (b) cross matrix for genes with high variant.

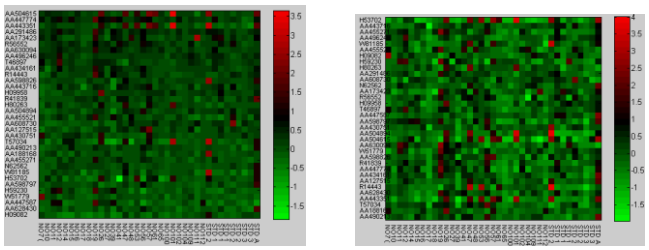


(a) (b) (c) (d)

Fig. 7 Selected genes from breast cancer cell line data with linear correlation coefficient >0.4 or <-0.4. (a) is the plot of linear correlation coefficient matrix of the selected genes (b) is the plot of the linear correlation coefficient of each selected gene's copy number and gene expression data. (c) is the copy number data of the selected genes and (d) is the gene expression data of the selected genes.



(a) (b)



(c) (d)

Fig. 8 Selected genes from breast tumor data with linear correlation coefficient >0.65 or <-0.65. (a) is the plot of linear correlation coefficient matrix of the selected genes (b) is the plot of the linear correlation coefficient of each selected gene's copy number and gene expression data. (c) is the copy number data of the selected genes and (d) is the gene expression data of the selected genes.

4. DISCUSSION AND CONCLUSION

From the experimental results of this work, it can be seen that once a cluster with a specified feature is set, SRC based clustering methods can identify the genes with the same features. In this way, it will become easier to find those genes with significant changes.

By employing normalized features, different datasets can be analyzed in the same way using SRC based clustering model, which is an effective way for joint analysis of different types of data.

From the simulated data experiments results (Fig. 2), we can see that the proposed SRC method is stable for data with different size and different pattern. From the analysis on the selection of top variant genes from each data set, we can see that the top highest variant genes selected by using SRC method are better than that of GSVD with higher variant values with p value<10e-5, which proved our model's ability in identifying genes with significant variance over samples. From the joint analysis results we can see that our proposed SRC method is effective in identifying data vectors with large variants when data sets are highly correlated.

In addition to gene shaving to extract significantly abnormal genes for certain diseases, the proposed SRC model can be used for classifications by simply picking those clusters with opposite features for two measurements. Currently, we are testing more data from different cancer and tumor models to prove the effectiveness of jointly data analysis.

5. REFERENCES

[1] Int'l Human Genome Sequencing Consortium, "Finishing the Euchromatic Sequence for the Human Genome," Nature, vol. 431, pp. 931-945, Oct. 2004.

[2] V.E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler, "Serial Analysis of Gene Expression," Science, vol. 270, pp. 484-487, Oct. 1995.

[3] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Norton, and E.L. Brown, "Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays," Nature Biotechnology, vol. 14, pp. 1675-1680, Dec. 1996.

[4] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," Science, vol. 270, pp. 467-470, Oct. 1995.

[5] A.S. Ishkanian, C.A. Malloff, S.K. Watson, R.J. de Leeuw, B. Chi, B.P. Coe, A. Snijders, D.G. Albertson, D. Pinkel, M.A. Marra, V. Ling, C. MacAulay, and W.L. Lam, "A Tiling Resolution DNA Microarray with Complete Coverage of the Human Genome," Nature Genetics, vol. 36, pp. 299-303, Mar. 2004.

[6] D. Hanahan and R.A. Weinberg, "The Hallmarks of Cancer," Cell, vol. 100, pp. 57-70, Jan. 2000.

[7] D.J. Lockhart and E.A. Winzeler, "Genomics, Gene Expression and DNA Arrays," Nature, vol. 405, pp. 827-836, June 2000.

[8] S.S. Jeffrey, M.J. Fero, A.-L. Børresen-Dale, and D. Botstein, "Expression Array Technology in the Diagnosis and Treatment of Breast Cancer," Molecular Interventions, vol. 2, pp. 101-109, Apr. 2002.

[9] C.M. Perou, T. Sørlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen, Ø. Fluge, A. Pergamenschikov, C. Williams, S.X. Zhu, P.E. Lønning, A.-L. Børresen-Dale, P.O. Brown, and D. Botstein, "Molecular Portraits of Human Breast Tumours," Nature, vol. 406, pp. 747-752, Aug. 2000.

[10] F. Forozan, R. Karhu, J. Kononen, A. Kallioniemi, and O.-P. Kallioniemi, "Genome Screening by Comparative Genomic Hybridization," Trends in Genetics, vol. 13, pp. 405-409, Oct. 1997.

[11] A.M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A.K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J.P. Yue, J.W. Gray, A.N. Jain, D. Pinkel, and D.G. Albertson, "Assembly of Microarrays for Genome-Wide Measurement of DNA Copy Number," Nature Genetics, vol. 29, pp. 263-264, Nov. 2001.

[12] J.R. Pollack, C.M. Perou, A.A. Alizadeh, M.B. Eisen, A. Pergamenschikov, C.F. Williams, S.S. Jeffrey, D. Botstein, and P.O. Brown, "Genome-Wide Analysis of DNA Copy-Number Changes Using cDNA Microarrays," Nature Genetics, vol. 23, pp. 41-46, Sept. 1999.

[13] L.W.M. Loo, D.I. Grove, E.M. Williams, C.L. Neal, L.A. Cousens, E.L. Schubert, I.N. Holcomb, H.F. Massa, J. Glogovac, C.I. Li, K.E. Malone, J.R. Daling, J.J. Delrow, B.J. Trask, L. Hsu, and P.L. Porter, "Array Comparative Genomic Hybridization Analysis of Genomic Alterations in Breast Cancer Subtypes," Cancer Research, vol. 64, pp. 8541-8549, Dec. 2004.

[14] J.R. Pollack, T. Sørlie, C.M. Perou, C.A. Rees, S.S. Jeffrey, P.E. Lonning, R. Tibshirani, D. Botstein, A.-L. Børresen-Dale, and P.O. Brown, "Microarray Analysis Reveals a Major Direct Role of DNA Copy Number Alteration in the Transcriptional Program of Human Breast Tumors," Proc.

Nat'l Academy of Science USA, vol. 99, pp. 12 963-12 968, Oct. 2002.

[15] S. Hautaniemi, M. Ringne ŕ, P. Kauraniemi, R. Autio, H. Edgren, O. Yli-Harja, J. Astola, A. Kallioniemi, and O.-P. Kallioniemi, "A Strategy for Identifying Putative Causes of Gene Expression Variation in Human Cancers," J. Franklin Inst., vol. 341, pp. 77-88, Mar. 2004.

[16] O. Monni, M. Ba ŕlund, S. Mousses, J. Kononen, G. Sauter, M. Heiskanen, P. Paavola, K. Avela, Y. Chen, M.L. Bittner, and A. Kallioniemi, "Comprehensive Copy Number and Gene Expression Profiling of the 17q23 Amplicon in Human Breast Cancer," Proc. Nat'l Academy of Science USA, vol. 98, pp. 5711-5716, May 2001.

[17] E. Hyman, P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringne ŕ, G. Sauter, O. Monni, A. Elkahloun, O.-P. Kallioniemi, and A. Kallioniemi, "Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer," Cancer Research, vol. 62, pp. 6240-6245, Nov. 2002.

[18] J. A. Berger, S. Hautaniemi, S. K. Mitra and J. Astola, "Jointly Analyzing Genes Expression and Copy Number Data in Breast Cancer using Data Reduction models", IEEE Transactions on Computational Biology and Bioinformatics, vol. 3, no.1, pp.2-16 2006.

[19] D. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution," Comm. on Pure and Applied Math, vol. 59, no. 6, pp. 797–829, 2006.

[20] E. Cand`es, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," Comm. on Pure and Applied Math, vol. 59, no. 8, pp. 1207–1223, 2006.

[21] E. Cand`es and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" IEEE Trans. Information Theory, vol. 52, no. 12, pp. 5406–5425, 2006.

[22] P. Zhao and B. Yu, "On model selection consistency of lasso," Journal of Machine Learning Research, no. 7, pp. 2541–2567, 2006.

[23] D. Donoho and Y. Tsaig, "Fast solution of `1-norm minimization problems when the solution may be sparse," preprint, http://www.stanford.edu/ tsaig/research.html, 2006

[24] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. A new approach to variable selection in least squares problems. IMA J. Numerical Analysis, 20:389–403, 2000.

[25] E. Hyman, P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringne ŕ, G. Sauter, O. Monni, A. Elkahloun, O. P. Kallioniemi, and A. Kallioniemi, "Impact of DNA amplification on gene expression patterns in breast cancer", Cancer Research, vol. 62, pp.6240-6245, 2002.

[26] J. R. Pollack, T. Sørlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A. L. Børresen-Dale, and P. O. Brown, "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors", The National Academy of Sciences, USA, vol. 99, pp.12963-12968, 2002.

[27] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust Face Recognition via Sparse Representation, IEEE TRANS. PAMI, Feb. 2009, vol. 31 no. 2, pp. 210-227

[28] P. Wang, Y. Kim, J. Pollack, B. Narasimhan, and R. Tibshirani, "A Method for Calling Gains and Losses in Array CGH Data," Biostatistics, vol. 6, pp. 45-58, Jan. 2005.

[29] S. Attoor, E. R. Dougherty, Y. Chen, M. L. Bittner, and J. M. Trent, "Which Is Better for cDNA-Microarray-Based Classification: Ratios or Direct Intensities," Bioinformatics, vol. 20, pp. 2513-2520, Nov. 2004.