Accurate Genomic Signal Recovery using Compressed Sensing

Bakhtiyar Uddin
Department of Computer Science, University of Texas, Austin, TX, USA

M. Emre Celebi*
Department of Computer Science, Louisiana State University, Shreveport, LA, USA

Hassan Kingravi
School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Gerald Schaefer
Department of Computer Science, Loughborough University, Loughborough, U.K.

Abstract

Microarrays are massively parallel biosensors that can simultaneously detect and quantify a large number of different genomic particles. A DNA microarray is a nucleic acid-based microarray that contains probe spots testing a multitude of targets in one experiment. Ideas from compressive sensing have been utilized in different ways in the analysis of DNA microarrays. One of the proposed methods is compressed microarrays, where each spot contains copies of several probes and the total number of spots is lower, resulting in significantly reduced costs due to cheaper array manufacturing. In this paper, we perform compressed microarray experiments with real aCGH data and demonstrate the accuracy of various recovery methods. Our experimental results suggest that the measurements that can be captured by compressed microarrays can be recovered accurately using the proposed norm-minimization methods.

1 Introduction

Sensing in DNA microarrays is based on the process of hybridization in which DNA strands complementary to each other bind and create structures in lower energy states [8, 9, 10, 4, 3, 2, 7, 5]. The surface of a DNA microarray consists of an array of spots, where each spot contains a large number of identical single-stranded DNA sequences, called the probes. Probes are designed to capture copies of a single DNA molecule of interest called the target. Microarrays determine the gene expression levels, which essentially determine the process of transcription of DNA information into messenger RNA. The transcribed information is then translated to proteins that perform most of the functions in the cells. Measuring gene expression levels may allow to extract critical information about the functionality of the cells, study diseases and the effects of drugs on them. Usually, DNA microarrays are used to compare the gene expression levels of a given gene sample with a reference gene sample.

Typically, only a fraction of the total number of genes is differentially expressed. Shmulevic *et al.* [11] proposed composite microarrays, in which each spot consists of several probes. A signal measured at each spot is potentially a combination of many targets. Using a composite microarray one could acquire multiple data points for each of the targets being tested. However, the signal recovery does not exploit the inherent sparseness of the signal. Parvaresh *et al.* [6] borrowed ideas from compressive sampling and proposed compressed microarrays. Compressive sampling is closely related to the problem of solving the following undetermined system of linear equations with a sparseness constraint:

$$\begin{aligned}&\text{minimize } ||\mathbf{x}||_1\\&\text{subject to } \mathbf{A}\mathbf{x} = \mathbf{y}\end{aligned} \tag{1}$$

where $\mathbf{A}$ is a binary sparse matrix that represents the probe distribution in the microarray, $\mathbf{x}$ is the raw signal that we are trying to estimate and $\mathbf{y}$ is the measured signal. In this paper we use the notation $||\mathbf{x}||_p$ to denote the p-th norm of $\mathbf{x}$, i.e. $||\mathbf{x}||_p = ({x_1}^p + {x_2}^p + \ldots + {x_n}^p)^{1/p}$.

*This work was supported by grants from the Louisiana Board of Regents (LEQSF2008-11-RD-A-12) and US National Science Foundation (0959583, 1117457)

978-4-9906441-1-6 ©2012 IAPR

Another method to exploit the inherent sparseness is to apply a block-sparse signal reconstruction method as proposed by Stojnic *et al.* [12]. A signal $\mathbf{x}$ (for example, a series of DNA copy numbers) is d-block-sparse if it consists of n blocks, each of size d where each block is either a zero or a non-zero vector. A convex relaxation for the recovery of $\mathbf{x}$

$$\begin{array}{l} \text{minimize } \|X_1\|_2 + \|X_2\|_2 + \ldots + \|X_n\|_2 \\ \text{subject to } \mathbf{Ax} = \mathbf{y} \end{array} \tag{2}$$

was proposed, where $X_i = (x_{(i-1)d+1}, x_{(i-1)d+2}, \ldots, x_{id})$ for $i = 1, 2, \ldots, n$.

Compressed sensing with norm-minimization recovery methods has been investigated in several studies [6, 12]. However, to the best of our knowledge, its performance was not demonstrated on real microarray data. In this paper, we apply various recovery techniques on real genomic data. We compare the accuracy of the methods that use block-sparse recovery, inspired by the fact that alterations typically affect contiguous segments of a genome [6], with those that use L_1 optimization recovery which takes advantage of the inherent sparseness of the aCGH data.

2 Related Work

DNA microarrays are used to compare the gene expression levels of a test sample with that of a reference sample. In practice, only a fraction of the total number of genes is differentially expressed, that is the difference of the signals produced by the two samples is sparse. Linear combinations of the signal components may be acquired by the composite probe spots that are comprised of a mixture of several probe sequences. However, the sparseness constraint suggests possible recovery of the signal from potentially far fewer probe spots than the total number of probe sequences.

A compressed microarray with m spots containing probes designed to quantify n different targets can be modelled as

$$\mathbf{y} = A\mathbf{x} + \mathbf{w} + \mathbf{v} \tag{3}$$

where $\mathbf{x}$ denotes the n-dimensional data vector representing the gene expression levels, $\mathbf{y}$ denotes the m-dimensional measurement, $\mathbf{w}$ is the shot noise, $\mathbf{v}$ is the m-dimensional zero-mean i.i.d. Gaussian additive noise due to instrumentation and other biochemistry-independent noise sources, and $\mathbf{A}$ is an $m \times n$ binary matrix containing information about probe mixing. Each row in $\mathbf{A}$ corresponds to a probe spot. The composition of the i-th probe is determined by the positions of ones in the i-th row of $\mathbf{A}$. A_{ij} is non-zero if and only if the j-th target can bind to some probes in the i-th spot. $\mathbf{A}$ is limited to binary 1/0 for the sake of manufacturing simplicity.

In a two color microarray experiment, we compare two samples characterized by data vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, and we are interested in finding differentially expressed genes. Ideally, this means $\mathbf{x} = \mathbf{x}_1 - \mathbf{x}_2$, $\mathbf{y} = \mathbf{y}_1 - \mathbf{y}_2$, $\mathbf{v} = \mathbf{v}_1 - \mathbf{v}_2$. We can write, $\mathbf{y} = A\mathbf{x} + \mathbf{w} + \mathbf{v}$, where $\mathbf{x}$ is sparse, i.e. it has a small number of entries that are non-zero (or significantly larger than zero). This means that one may be able to recover $\mathbf{x}$ using L_1 minimization $\min_{\mathbf{x},\mathbf{Ax}=\mathbf{y}} ||\mathbf{x}||_1$.

The recovery accuracy of this method is supported by the study of sparse signal recovery using sparse matrices by Berinde and Indyk [1]. They discuss recovering a high dimensional vector $\mathbf{x}$ from its lower dimensional measurement $\mathbf{Ax}$, where $\mathbf{A}$ is binary and sparse, i.e. it has only a fixed small number of ones in each column and all other entries are zero. It is shown that such matrices satisfy a weaker form of the RIP-p property. Use of these matrices is advantageous because it fits well in many applications such as ours. Furthermore, it has an efficient update time, which is equal to the sparsity parameter d. Another advantage is that such matrices can be constructed using expander graphs [1]. Berinde and Indyk focused on the recovery method that computes a solution $\mathbf{x}$ to the linear system given in (1) and proved that the recovery can be very accurate.

Parvaresh *et al.* [6] proposed the application of this method to the recovery of compressed DNA signals after compressed sensing. They also proposed the inclusion of a differentiation operator to reduce the number of spikes in the recovered signal. The purpose of this operator is to generate a piecewise constant signal by minimizing the number of jumps. This approach can be formulated as follows:

$$\begin{array}{l} \text{minimize } \|\mathbf{x}\|_1 + \gamma \|D\mathbf{x}\|_1 \\ \text{subject to } A\mathbf{x} = \mathbf{y} \end{array} \tag{4}$$

where $\mathbf{A}$ is a binary sparse matrix that represents the probe distribution in the microarray and D is the differentiation operator given by

$$D = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & . & . & . \\ 0 & 1 & -1 & 0 & 0 & . & . & . \\ 0 & 0 & 1 & -1 & 0 & . & . & . \\ & & . & . & . & . & . & \\ & & . & . & . & . & . & \\ 0 & 0 & . & . & . & 0 & 1 & -1 \end{pmatrix}$$

Stojnic *et al.* [12] investigated the application of block-sparse signal reconstruction methods to DNA microarrays. A signal of dimension N consists of n blocks of size $d = N/n$. Such a signal is k-block-sparse if only k blocks of the signal out of n are nonzero. Instead of

L_1 norm relaxation, they considered L_p/L_1 relaxation, which solves the convex problem given in (2). Their main result is the following theorem:

Theorem 1 *Let* $\mathbf{A}$ *be an* $md \times nd$ *matrix. Further, let* $\mathbf{A}$ *be an instance of the random Gaussian ensemble. Assume that* ϵ *is a small positive number, i.e.* $0 < \epsilon << 1$, $d = \Omega(log(1/\epsilon)/\epsilon)$, $\alpha > 1 - 1/d$, $\beta = 1/2 - O(\epsilon)$. *Also, assume that* n *tends to infinity,* $m = \alpha n$, *and the block-sparsity of* $\mathbf{x}$ *is smaller than* βn. *Then with overwhelming probability, any* d*-block-sparse signal* $\mathbf{x}$ *can be reconstructed efficiently from* $\mathbf{y} = \mathbf{A}\mathbf{x}$ *by solving the optimization problem given in* (2).

Stojnic *et al.* proposed the application of this method to DNA microarrays because over/under expressions of genes usually occur in contiguous segments.

3 Recovery of Genomic Data

All of the recovery methods described above can be represented by the following general expression:

$$\begin{aligned}&\text{minimize } \|X_1\|_p + \|X_2\|_p + \ldots + \|X_n\|_p + \gamma\,\|D\mathbf{x}\|_1\\&\text{subject to } \mathbf{A}\mathbf{x} = \mathbf{y}\end{aligned} \tag{5}$$

For genomic data recovery using (5) with $p = 2$ minimization appears to be promising. This recovery method tries to maximize the number of zero elements because only a small fraction of the total number of genes is differentially expressed. It also tries to provide a solution in which non-zero elements occur in blocks, unlike the solutions given by (1) with $p = 1$ that usually contain spikes which do not occur in real aCGH data. We experimented with various p and γ values in (5) and demonstrated their performance for the detection of over and under expressions of genes. From a data set that contains about $10,000$ signals, we selected a subset of approximately $1,000$ for our experiments.

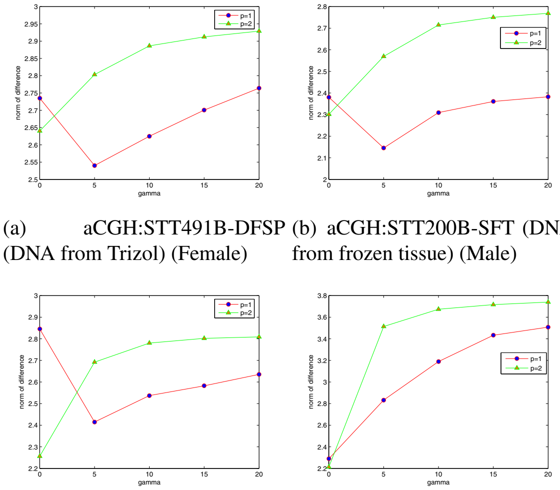
4 Experimental Results and Discussion

We represented the probe mixing of the microarray using using a binary sparse matrix $\mathbf{A}_{m\times n}$, which is generated randomly using the method described in [1]. More specifically, for each column, we generated δ random integers between 1 and m, and assigned 1s to the corresponding rows. The resulting matrix would be an adjacency matrix of an expander graph of degree δ with a high probability and therefore would satisfy the RIP-p property. In the experiments we set $\delta = m/16$, $m = 400$, and $n = 1,000$, so that the corresponding microarray would have 400 spots and would be able to detect $1,000$ unique targets. Each spot would have exactly 25 probes. These values were chosen to make the experiments realistic.

We obtained the raw signal $\mathbf{x}$ from the aCGH tumor tissue subcategory of the Stanford Microarray Database[1]. We then calculated the measurement signal $\mathbf{y} = \mathbf{A}\mathbf{x}$ and determined the recovered signal, $\mathbf{x}'$ using various p and γ values. Finally, we analyzed how close the recovered signal $\mathbf{x}'$ was to the original signal $\mathbf{x}$.

It is difficult to perform analysis on the raw data because of noise. It can be difficult to identify aberrant regions by examining the raw data. It is much easier to do so when we analyze the smoothed data. That is why, when we compare the performance of a particular recovery method, we analyze the recovered data and the raw data after smoothing both of them.

Fig. 1 demonstrates the performance of the methods on four data sets. The y-axis corresponds to the norm of difference given by $||\text{smooth}(\mathbf{x}) - \text{smooth}(\mathbf{x}')||_2$, where smooth$(\mathbf{x})$ and smooth$(\mathbf{x}')$ denote the vectors after smoothing the signals $\mathbf{x}$ and $\mathbf{x}'$, respectively.



(a) aCGH:STT491B-DFSP (DNA from Trizol) (Female)

(b) aCGH:STT200B-SFT (DNA from frozen tissue) (Male)

(c) aCGH:STT491-DFSP (DNA from PAR) (Female)

(d) CGH:STT200B-SFT (DNA from frozen tissue) (Male)

Figure 1. γ vs. norm of difference

In our experiments, when $\gamma = 0$, block-sparse recovery with $p = 2$ provided consistently superior results. Since L_1 minimization recovery ($p = 1$) tries to sparsify the recovered vector, its results would lead to the conclusion that most genes are not differentially expressed except a very small subset, even though this is not that case in the original signal. This recovery method is not suitable for a signal which has a sparsity as high as the aCGH data. (5) with $p = 2$, on the

[1] `http://smd.stanford.edu/`

other hand, can handle higher sparsity values [12]. It tries to find a solution which is block-sparse by trying to maximize the number of non-zero blocks. We included a penalty term ($\gamma > 1$) to the differentiation matrix to penalize jumps. This helped us finding solutions with a minimal number of spikes in the data. However, if an excessively large penalty value is used, this approach provides solutions that are approximately piecewise, but with too many neighborhoods that are over- or under-expressed, i.e. the recovered signal becomes too serrated. However, for a small enough γ, e.g. $\gamma = 5$, we consistently achieved near-optimal recovery. Fig. 2 demonstrates that, after smoothing, the signal recovered by (5) with $p = 1, \gamma = 5$, maintains most of the peaks and valleys with approximately the same expression levels as those in the smoothed raw signal.

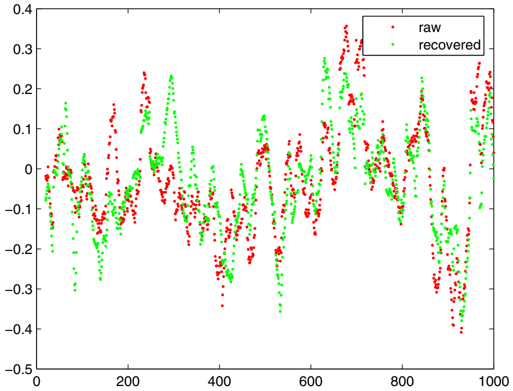


Figure 2. Recovery performance of (5) **with** $p = 1, \gamma = 5$ **on a subset of data aCGH:STT491B-DFSP (DNA from Trizol) (Female)**

5 Conclusions

In this paper, we investigated the recovery of microarray data from the measurement signals of compressed microarrays using methods that take advantage of the inherent sparseness of the raw microarray data. We demonstrated that the method given in (5) with $p = 2, \gamma = 0$ yields consistently near-optimal results. The parameter combination $p = 1, \gamma \approx 5$ performs similarly well. The latter configuration is in fact a better choice because with $p = 2$ we need to define another variable, namely the block size. If the block size is too small, our result would be almost as bad as the one we obtain with $p = 1, \gamma = 0$. On the other hand, if it is too big, then it decreases our resolution by a factor of the block size when we try to analyze whether a gene is over- or under-expressed. If in a particular block, there are genes with irregular expression levels, i.e. some genes are over/under-expressed and some are not, then it is very unlikely that the block-sparse recovery method will recover the expression levels of the individual genes.

In this study, we have not taken into account the practical limitations that would result from biological and measurement noise. In future work, we hence plan to factor in those limitations.

References

[1] R. Berinde and P. Indyk. Sparse Recovery Using Sparse Random Matrices. Technical report, Massachusetts Institute of Technology, 2008.

[2] A. P. Blanchard and L. E. Hood. Sequence to Array: Probing the Genome's Secrets. *Nature Biotechnology*, 14(13):1649, 1996.

[3] A. P. Blanchard, R. J. Kaiser, and L. E. Hood. High-Density Oligonucleotide Arrays. *Biosensors and Bioelectronics*, 11(6/7):687–690, 1996.

[4] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent. Use of a cDNA Microarray to Analyse Gene Expression Patterns in Human Cancer. *Nature Genetics*, 14(4):457–460, 1996.

[5] U. R. Müller and D. V. Nicolau. *Microarray Technology and Its Applications*. Springer, 2004.

[6] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi. Recovering Sparse Signals Using Sparse Measurement Matrices in Compressed DNA Microarrays. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):275–285, 2008.

[7] M. Schena. *Microarray Analysis*. Wiley-Liss, 2003.

[8] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235):467–470, 1995.

[9] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis. Parallel Human Genome Analysis: Microarray-Based Expression Monitoring of 1000 Genes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(20):10614–10619, 1996.

[10] D. Shalon, S. J. Smith, and P. O. Brown. A DNA Microarray System for Analyzing Complex DNA Samples Using Two-Color Fluorescent Probe Hybridization. *Genome Research*, 6(7):639–645, 1996.

[11] I. Shmulevich, J. Astola, D. Cogdell, S. R. Hamilton, and W. Zhang. Data Extraction from Composite Oligonucleotide Microarrays. *Nucleic Acids Research*, 31(7):e36, 2003.

[12] M. Stojnic, F. Parvaresh, and B. Hassibi. On the Reconstruction of Block-Sparse Signals with an Optimal Number of Measurements. *IEEE Transactions on Signal Processing*, 57(8):3075–3085, 2009.