Microarray Image Compression: SLOCO and the Effect of Information Loss

R. Jornsten [a,1] W. Wang [b,2] B. Yu [c,1] K. Ramchandran [b,3]

[a] *Department of Statistics, Rutgers University*

[b] *Department of Electrical Engineering, University of California, Berkeley*

[c] *Department of Statistics, University of California, Berkeley*

Abstract

Microarray image technology is a powerful tool for monitoring the expression of thousands of genes simultaneously. Each microarray experiment produces immense amounts of image data, and efficient storage and transmission require compression that takes advantage of microarray image structure. In this paper we develop a compression scheme for microarray images which can be either lossless or lossy with successive refinements. Existing measures of distortion such as mean squared pixel-wise error and visual fidelity are *not* appropriate for microarray images. We introduce a new measure of distortion for lossy compression: the sensitivity of microarray information extraction to compression loss. Furthermore, our scheme has a coded data structure that allows fast decoding and reprocessing of image subblocks, and includes summary statistics and image segmentation information.

The average lossless compression ratio is 1.83:1 for our cDNA test images and 2.43:1 for our inkjet test images, comparable or better than state-of-the-art lossless schemas, yet with additional structure and information. At an average lossy compression ratio of 8:1 for cDNA microarrays, we find that our scheme minimizes the effects of compression loss compared to other algorithms. We show that the variability in differential gene expression levels extracted from *lossily* versus *losslessly* compressed microarray images is less than both the variability between different arrays and the variability between different extraction algorithms. In fact, lossy compression can improve the estimation of gene expression levels for cDNA images.

Email addresses: rebecka@stat.rutgers.edu (R. Jornsten), wangwei@eecs.berkeley.edu (W. Wang), binyu@stat.berkeley.edu (B. Yu), kannanr@eecs.berkeley.edu (K. Ramchandran).

[1] NSF FD98-02314, NSF DMS-9803063, ARO DAAG55-98-1-0341, NSF FD01-12731

[2] NSF Graduate Fellowship, Lucent GRPW Grant

[3] NSF Career Award

1 Introduction

Microarrays have become an important tool for developing understanding of gene function, regulation and interaction through the simultaneous study of thousands of genes. The raw data produced by a microarray experiment is an image (or more accurately, a two-dimensional matrix of intensity values) of a grid of thousands of spots, one for each gene.[4] The proven and potential scientific value of microarray image technology is enormous, but it is widely acknowledged that the quality of microarray data is highly variable. Without standardization of methods and microarray processing tools, results from different labs are rarely comparable. We need to distinguish between the inherent noise of the data, the source of which lies in experimental variation, and the noise introduced by different image processing methods used for genetic information extraction. For this reason, it is often advisable to store the raw image data from an experiment rather than just the summary statistics from a particular extraction algorithm. Moreover, microarray experiments are costly and the image data files are large (>40 MB). Various organizations are establishing publicly accessible databases for sharing microarray data. Though the cost of data storage is rapidly decreasing, compression is an essential tool for efficient transmission and data distribution, and for storage of large quantities of image data.

We propose a compression scheme which can be either lossless or lossy with successive refinements. Thus, given the prohibitive size of microarray images and transmission bandwidth limitations, it may be useful, for example, to quickly download low rate representations of the data for the user to decide whether to obtain the full image. We add necessary structure and flexibility to traditional lossless coding that facilitates downstream genetic information extraction analysis. Thus researchers may examine individual "spots" or genes without decoding the entire file, or simply decode the statistics needed for analysis of gene expression levels. We preserve the genetic information in the lossily compressed microarray images, such that the effect of compression loss is within the experimental noise level. We take into consideration that the appropriate measure of performance for compression of microarray images is not visual fidelity or MSE, as is generally the case in image compression. Microarray images are processed using a multi-step procedure (image segmentation, information extraction, normalization), and thus we do not have a simple distortion criterion over which to optimize. One variant of our lossy scheme has a locally varying bound on the maximum pixel-wise error, i.e. the error bound is a function of the local signal-to-noise ratio. With this approach, at low bit-rates (compression ratio 8:1 for cDNA microarrays), the effect of lossy compression on the extracted genetic information is smaller than the array-to-

[4] Section 2 discusses the mechanics of microarray experiments.

array variability in replicated experiments. Moreover, the impact of compression is smaller than that of changing the method of information extraction. Thus we define a new criterion of acceptable distortion for lossy compression of microarray images. In fact, for large (absolute) gene expressions, compression acts as *denoising* of the estimates extracted from a single array, towards the mean over replicated arrays.

Microarray imaging is an emerging technology and several experimental procedures have been developed producing different image characterizations, including cDNA, oligonucleotide, and inkjet technologies (section 2). We will outline our compression algorithm, SLOCO, and describe how it differs from established compression schemes (section 3). We will give compression results using SLOCO and other state-of-the-art algorithms on cDNA and inkjet microarrays (section 5). We will then show empirical evidence that our lossy compression scheme minimally affects statistical analysis of gene expressions.

2 cDNA and Inkjet Microarray Images

A gene is a segment of DNA that encodes the amino acid sequence comprising a specific protein. The expression of a gene is achieved in two stages: *transcription*, where the DNA segment is transcribed into a messenger (m)RNA, and *translation*, where mRNA is used as a blueprint for building the protein. One microarray experiment simultaneously measures the level of activity of several thousand genes in the transcription stage, which is then related to the amounts of the corresponding proteins being produced. Microarray analysis studies the difference in gene expression levels in two samples, e.g. a control sample and a treatment sample.

In the microarray experimental procedure, fluorescently labeled cDNA targets, which have been synthesized from cellular RNA, are hybridized to DNA probes immobilized on a solid support (spots on a glass slide). The DNA probes can be cloned cDNA gene fragments, or oligonucleotides corresponding to known genes. Two cDNA targets are prepared independently. One target is prepared from control mRNA (e.g. from a normal tissue sample), while the second target is prepared from mRNA extracted from the treatment cells under investigation (e.g. diseased tissue). The cDNA targets are labeled with green (Cy3) and red (Cy5) fluorescent dye, respectively. In a process called *competitive hybridization*, the two targets are hybridized on the same slide. A laser scan of the array produces two fluorescent intensity images. The intensity ratio for each spot is proportional to the amounts of specific transcripts in the two samples. Thus microarray experiments measure *differential gene expression*.

Several variants of microarrays differ in the types of probes used and the method by which the probes are immobilized on the glass slide. In the cDNA microarray procedure, cDNA probes are deposited onto a glass slide by a robotic *arrayer* [1]. Affymetrix developed high-density oligonucleotide microarrays where the oligonucleotide probes are synthesized on a glass surface in situ using a photolithographic mask [2]. Inkjet microarrays deposit pre-synthesized oligonucleotide probes onto a glass slide with a non-contact inkjet printing device [3].

The microarray images are structured, with high intensity spots located on a grid (see figure 1). The spots are submerged in a noisy and non-stationary background. The spots have a roughly circular shape, though some show significant deviation from this shape due to the experimental variation of the spotting procedure.

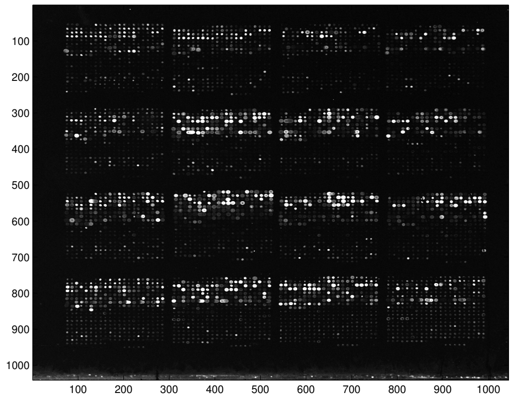


Fig. 1. test cDNA microarray image: 4×4 grid set-up, 19×21 spots per grid.

2.1 Genetic Information Extraction

To estimate the differential gene expression, we have to identify the high intensity region in the image corresponding to each probe (image segmentation), and estimate and remove the local background noise (background correction).

Automatic registration of the image is used to determine the approximate centers, or the grid locations, of the spots. [4][5]. We use a seeded region growing algorithm for initial segmentation, followed by a two-component Gaussian mixture model, to further refine the boundaries of the spots, or regions of interest (ROI) [6]. Various approaches to background correction are described in [5], [7], [4].

An example of a spot, or gene, summary statistic for cDNA is the ratio of background corrected mean intensities. We denote by R_i the red fluorescent scan pixels, and by G_i the green fluorescent scan pixels. The spot differential expression level, R/G, is then calculated as the ratio of the background corrected mean ROI intensities:

$$\frac{R}{G} = \frac{\frac{1}{S}\sum_{R_i \in ROI} R_i - BgR}{\frac{1}{S}\sum_{G_i \in ROI} G_i - BgG},$$

where Bg refers to the estimates of the local background, and S is the number of ROI pixels. The quantity used in downstream statistical analyses, such as clustering and classification, is the vector $M = \log_2 \frac{R}{G}$ of log-ratios of differential gene expressions. Other quantities of interest are measures of quality, such as spot variances, spot shapes, and product intensities $A = \log_2 \sqrt{RG}$. The product intensities, A, are indicative of how reliable the measurements of the gene expression are.

Normalization is also necessary prior to statistical modeling and comparisons between microarrays. Without normalization, systematic variation in the data may dominate over chance variation. We apply the normalization scheme of Dudoit et al [8] when analyzing the effect of compression loss using replicates.

3 Lossy and Lossless Compression of Microarray Images

State-of-the-art lossy compression schemes for natural and medical images are typically wavelet based. However, this is not appropriate for the compression of microarray images. In fact, the many and small high intensity regions create large wavelet coefficients over almost the entire image subbands. At low bitrates, algorithms such as SPIHT, or wavelet and zero-tree coding [9] will be dominated by the edges around the high intensity spots. Compression schemes that are not wavelet based, but based on causal predictions in the spatial domain, create a "smearing" bias in the image reconstruction.

We take a segmented spatial prediction approach. To avoid the "smearing" bias we encode the spot and background pixels separately. We first transmit an *overhead* containing a segmentation map, along with summary statistics necessary for the encoding (section 3.2). We refer to this approach as segmented LOCO, or SLOCO (section 3.1).

We want the output of the compression scheme to have a data structure that would enable us to transmit and reconstruct image subblocks at different precisions. SLOCO encodes a microarray image in blocks, each containing an individual or a group of spot regions, according to the grids established by the segmentation (sections 3.3 3.4).

In compression of natural images and medical images, acceptable loss is defined as that which makes the lossy and lossless reconstructions *visually* indistinguishable. Microarray image data analysis, however, is not a visual task. Instead, our goal is to minimize the effect of compression loss on downstream genetic information extraction. Specifically, our aim is to ensure that the variability between differential gene expression levels extracted from lossless and lossy microarray image reconstructions is smaller than the variability between replications of the same experiment, or the variability between different methods of information extraction. We define this as *acceptable loss* for microarray image compression.

Background correction and normalization involve segmenting foreground from background, and microarray backgrounds are invariably low-intensity. To facilitate this segmentation, a compression scheme must keep more precision for low intensity spot regions, but can use a coarse image reconstruction for high intensity spot regions. Note that the commonly used MSE criterion does not reflect this. We find that local SNR is a good indicator for background-corrected spot intensity. Thus we vary the pixel-wise error bound using local SNR thresholds (section 3.3).

3.1 Segmented LOCO (SLOCO)

Our scheme builds on the JPEG 2000 lossless standard, LOCO (LOw COmplexity), [10]. SLOCO differs from standard LOCO in three primary aspects. First, the spots and backgrounds are encoded separately. Second, a UQ-adjust quantizer is used instead of the UQ quantizer. Third, we allow for varying maximum pixel-wise error bounds δ. We use a runlength code that takes the segmentation map and the varying δ into account.

Outline of Encoding

The LOCO algorithm [11] uses a robust pixel domain prediction scheme, e.g. the causal neighbors of pixel x is used to predict its value. To further improve on the fixed prediction scheme, a context-based adaptive predictor is also used, where the *contexts* are defined by the vector of local gradients. The standard LOCO algorithm has been extended to near-lossless compression [10]. There, the prediction errors are quantized with a uniform quantizer (UQ) with bin-widths $2\delta + 1$, and reconstruction at the center of each bin.

We use a UQ-adjust quantizer [12] for large δ. Thus, the bins near the center of the error distribution, where most of the probability mass is located, have adjusted reconstruction levels closer to the MSE distortion optimal, as well as smaller bin-widths $2\delta' + 1$, such that the maximum error is still bounded by δ. The outer bins have bin-widths $2\delta + 1$, and reconstruction levels at the center of the bins.

We assume that the distribution of the quantized errors follows a two-sided geometric distribution. We map the quantizer error distribution into a smaller alphabet without loss of information. Using Rice mapping, the tails of the distribution can be mapped to the center of the distribution, but with opposite sign [11]. The reduced-range error distribution is then mapped to the positive integers. The distribution of the mapped and quantized prediction errors is close to a one-sided geometric distribution.

Golomb codes [11] are optimal for such distributions. A quantizer bin index y is then encoded using Golomb codes. The Golomb parameter k is estimated adaptively for each pixel according to $k = argmin_{k'}\{2^{k'} \geq \bar{A}\}$, where $\bar{A}$ is the accumulated mean absolute quantized prediction error for the context class of that pixel.

Efficient encoding of the smooth regions of an image is achieved by means of a run-length code [11].

3.2 Overhead

The overhead of the SLOCO algorithm contains the spot means and standard deviations, as well as the local background intensity estimates and estimates of the local background standard deviation. The overhead also contains the estimated grid structure of the microarray images, and the segmentation map.

The spot and background means are encoded using adaptive Lempel-Ziv. The segmentation map is efficiently encoded using the chain code of Lu and Dunham [13]. The average cost of the overhead is ~ 0.376 bpp for our cDNA test images.

If no re-processing of the images is needed, the overhead contains all relevant information for down-stream analysis. In addition, it contains spot quality measurements such as spot shapes, variances and the local background variance.

3.3 Coding the Spot Regions

Given the overhead we can compute the signal-to-noise ratio of each spot. Based on the SNR we can pick a bound on δ for each spot. We threshold the SNR distribution at the $10th$, $50th$, and $90th$ quantiles, and use increasing error bounds accordingly. Each spot is encoded in a raster scan manner. Missing context pixels for the predictor are filled in with the spot mean from the overhead. For cDNA, the size of each spot is too small to allow for adaptive prediction or adaptive estimation of the Golomb parameter. We therefore use a fixed Golomb code and only the fixed predictor for cDNA. Inkjet images have larger spot sizes than cDNA and using the adaptive Golomb parameter greatly improves compression.

3.4 Coding the Background

Sublocks of the background image pixels are encoded in a raster scan fashion. Missing context pixels, i.e. the spot pixels, are filled-in using local background intensity estimates. It is marginally more efficient to encode background pixels for the entire image as one block. However, we find that encoding sub-blocks (in cDNA, corresponding to the 4×4 print-tip configuration) gives similar bit-rate but allows for image subset reconstruction.

The runlength coding of SLOCO differs from regular LOCO in that we do not allow runs to cross from a region with higher maximum error bound δ, into one with smaller δ. If a spot is encountered during a run, we skip ahead to the next background pixel.

4 Successive Refinement

Our lossy compression scheme can be extended to a fully lossless reconstruction of the microarray images. Given the initial lossy reconstruction, the image reconstruction can be refined, spot by spot, background region by background region, or even pixel by pixel, to any bit-rate above the minimum decodable bit-rate. Our scheme is thus progressive.

The overhead provides us with the maximum error δ in each region of the lossy reconstruction of the images. We encode bit planes of the residual image. If the maximum error in a region chosen is δ, then the quantization errors are approximately uniformly distributed on $[-\delta, \delta]$. This holds if δ and δ' are reasonably close, and small. A UQ with two reconstruction levels corresponds to the first bit plane, and equals the sign of the quantization errors, which can be sent at rate 1 bpp. The decoder then reconstructs a refined error as $sign \times \delta/2$ and adds this to the previous lossy reconstruction. If we want to refine this region of the image further, we compute a new residual image, again send the sign of the residual to the decoder, and reconstruct at levels $sign \times \delta/4$. This achieves the $\log_2(2\delta+1)$ bit-rate. We choose this simple coding scheme since it gives us total freedom to encode any part of the residual image at any rate we desire, independently of what we choose to do in other regions of the image.

If δ and δ' are very dissimilar, the quantization errors will not be uniformly distributed. We then compute the quantization bin centroids at the encoder, and send as overhead to the decoder, with negligible increase in coding cost for the large microarray images.

5 Compression Results and Effect of Compression Loss

5.1 Image Data Sets

We used cDNA images from three different labs: 16 images from Lawrence Berkeley National Lab, courtesy of Matthew Callows, Genome Sciences; 2 images from P. Brown's lab, Department of Biochemistry, Stanford University; and 2 images from Lawrence Livermore National Lab. The 16 images are from 8 replicate experiments (8 images from the control group and 8 images from the treatment group) from LBNL allow us to evaluate the effects of compression on the statistical analysis of extracted genetic information. A discussion on the statistical analysis of this data set can be found in [14]. We compare two estimates of normalized gene expression levels, $\tilde{M}$, one using the full image data, and the other using lossy reconstructions of the images. We also look at the extracted product intensities A.

We used 10 inkjet microarray images from Rosetta Inpharmatics, courtesy of E. Schadt. However, we do not have image data sets from replicate experiments for inkjet microarrays.

5.2 Compression Results

The average LOCO lossless compression ratio for our 16 replicate cDNA images is 1.85:1. The SLOCO ratio is 1.83:1 (including the overhead information). In comparison, Lempel-Ziv (gzip) gives a compression ratio of 1.48:1, SPIHT 1.65:1, and wavelet zero-tree coding + entropy coding of residual (WT+ZT+EC) 1.72:1. The 8 least significant bits of cDNA images are close to random, i.e. have marginal entropy 8 bpp, and are unpredictable. This puts a ceiling of 2:1 on the lossless compression ratio. Inkjet images are much less noisy than cDNA images. The average LOCO lossless compression ratio for our 10 inkjet images is 2.44:1, while the SLOCO ratio is 2.43:1.

Note that the SLOCO encoding contains easily extractable summary statistics and segmentation information not available from LOCO. Moreover, SLOCO allows fast reconstruction of subsets of the image, so researchers can access individual spots out of hundreds of thousands per image.

The lossy compression ratio, using variable pixel-wise distortion bound, is 8:1 for the replicate cDNA microarray images. We find that the variability introduced by the lossy compression in the extraction of the differential gene expression levels, is smaller than the array-to-array variability. We also find that the difference in extracted gene expression levels between different methods of genetic information extraction is much greater than the difference between lossless and lossy reconstructions of the images (Section 5.3, [12]). For lack of replicate inkjet images, we do not present empirically the effect of compression loss on downstream analysis of inkjet microarrays. The lossy compression ratio for inkjet, with fixed maximum pixel-wise error $\delta = 8$ (3 bpp), is 6.51:1 using LOCO and 6.44:1 using SLOCO.

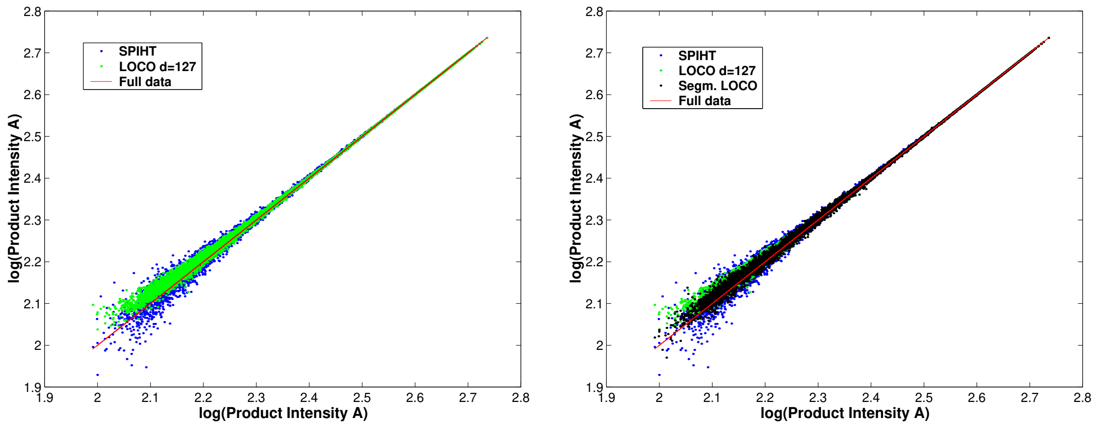
5.3 Effects of Lossy Compression on cDNA Microarray Analysis

We compare SPIHT, LOCO, and SLOCO in terms of how well the spot product intensity, A, and the differential gene expression levels, $\tilde{M}$, are preserved in the lossy image reconstructions. We compare the extracted information from *lossy* microarray image reconstructions at a given bit-rate, to the extracted information from the full (lossless) image data. The (WT+ZT+EC) compression scheme shows very poor results, and is dropped from the comparison. The wavelet transform has a wide support, such that the spots are blended with each other and the extracted information is far from accurate.

Figure 2 shows the spot product intensity A on a log scale (section 2.1) obtained from the SPIHT reconstruction of a cDNA image at 4.1 bpp, the LOCO reconstruction at 4.4 bpp, and the SLOCO reconstruction at 4.1 bpp. The variance of A is high using the SPIHT algorithm, whereas LOCO shows significant positive bias in A for low A values. Segmented LOCO controls both the bias and variance better than LOCO and SPIHT respectively.

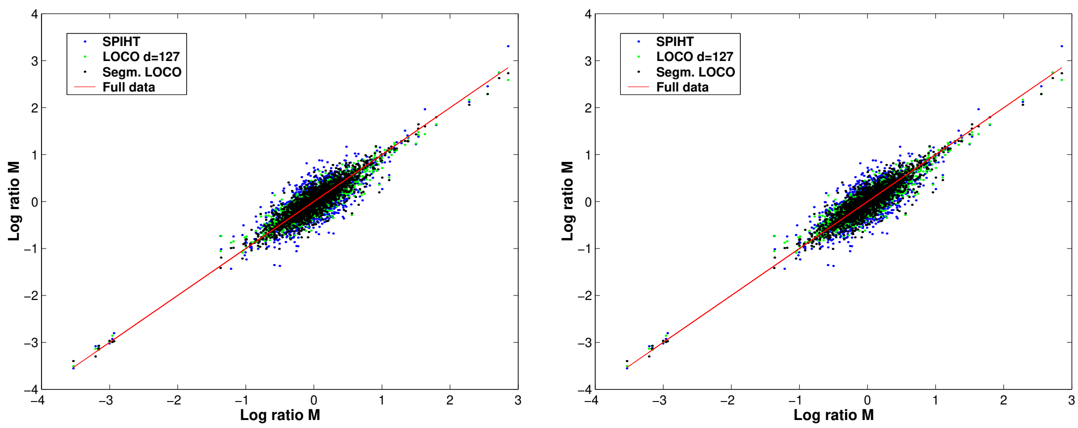
Figure 3 shows the normalized differential gene expression levels $\tilde{M}$ (section 2.1) obtained using SPIHT, LOCO and segmented LOCO at the above bit-rates. We again see that the wavelet based SPIHT algorithm performs very poorly. The segmented LOCO scheme fairs slightly better than standard LOCO. Note that all schemes are able to preserve the large absolute $\tilde{M}$ values with great accuracy. Only for small $\tilde{M}$ values does the compression seem to have an impact.

Our scheme is progressive: we can opt to add more precision to the regions of the images corresponding to small $\tilde{M}$, where the effect of compression is more noticeable. However, we find that the variability introduced by the lossy compression, in the extraction of the differential gene expression levels, is smaller



(a) (b)

Fig. 2. Effect of compression loss on spot product intensities A: A extracted from lossy image reconstructions are plotted against A extracted from lossless image representations. (a) SPIHT (blue) at rate 4.1 bpp, LOCO (green) at rate 4.4 bpp, and (b) SLOCO (black) at 4.1 bpp superimposed. Variance and Bias both under control with Segmented LOCO (black) at 4.1 bpp

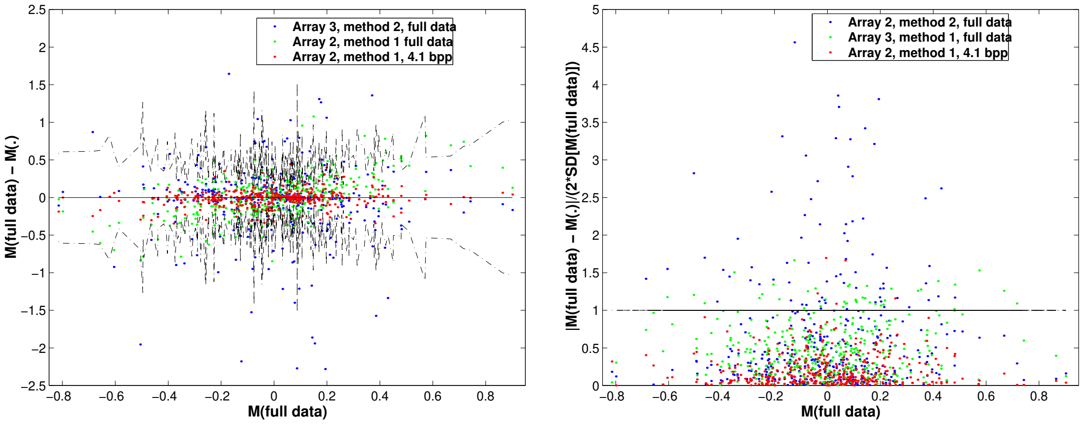


(a) (b)

Fig. 3. (a) Differential gene expression $\tilde{M}$ for control group (b) treatment group. SPIHT and Segmented LOCO at 4.1 bpp (blue, black) and LOCO (green) at 4.4 bpp.

than the array-to-array variability, calculated from replicates. We also find that the difference in extracted gene expression levels between different methods of image processing is much greater than the difference between lossless and lossy reconstructions of the images.

We compare two methods of image processing: Method 1 employs seeded region growing + robust filtering background correction; Method 2 employs adaptive circle segmentation + valley-between-the-peaks background correction (section 2.1). In figure 4 (a) the baseline for comparison is the gene expression levels extracted from Array 2 using Method 1. We compare the results we get using Method 2 on the same array, and Method 1 on Array 3 (blue and green markers), to the full data. We also compare the results we get using Method 1 on a 4.1 bpp lossy reconstruction of Array 2 (red markers). We add 2*SD bands to the plot, where the SD has been computed from the 8 replicated arrays of the same type as Array 2. The array-to-array variability



(a) (b)

Fig. 4. (a) 2*SD bands centered at M using method 1 on array 2. The marker display differences in M from array 2 using i) method 2 on full data, ii) method 1 on array 3, iii) method 1 on 4.1 bpp reconstruction of array 2. (b) 'Standardized plot', i.e. z-statistic/2.

is much greater than the lossy-lossless difference. The method-to-method difference is far greater than the array-to-array variability, indicating that great care has to be taken in which method is used for information extraction. Since the SD bands are hard to distinguish in the plot, we add a standardized plot of z-statistics (divided by 2) in figure 4 (b). We obtain similar results for all other arrays, and when the roles of Method 1 and Method 2 are reversed.

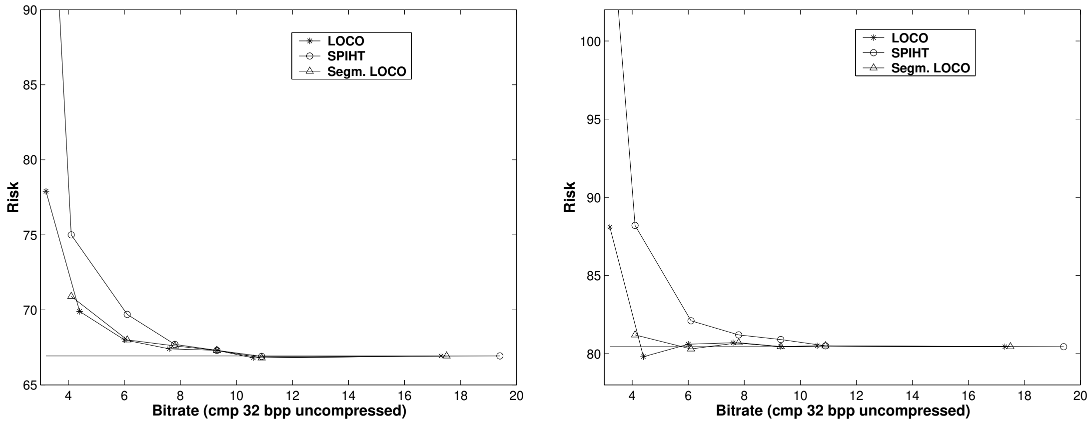
Risk, Denoising, and Shrinkage

Using our cDNA replicate image set (section 5.1), we compute the mean gene expression vector from the 8 replicates using Method 1, and denote this by M^0. We construct a measure of risk for the compressed data at various bit-rates as follows. Let $\widehat{M}^j$ be the vector of extracted gene expressions from a lossy reconstruction of array j. If there are p genes on each array the risk R is computed as

$$R = \sum_{k=1}^{p} \frac{1}{8} |\widehat{M}_k^j - M_k^0|.$$

We use the L_1 norm to avoid having the large gene expressions dominate the calculated risk.

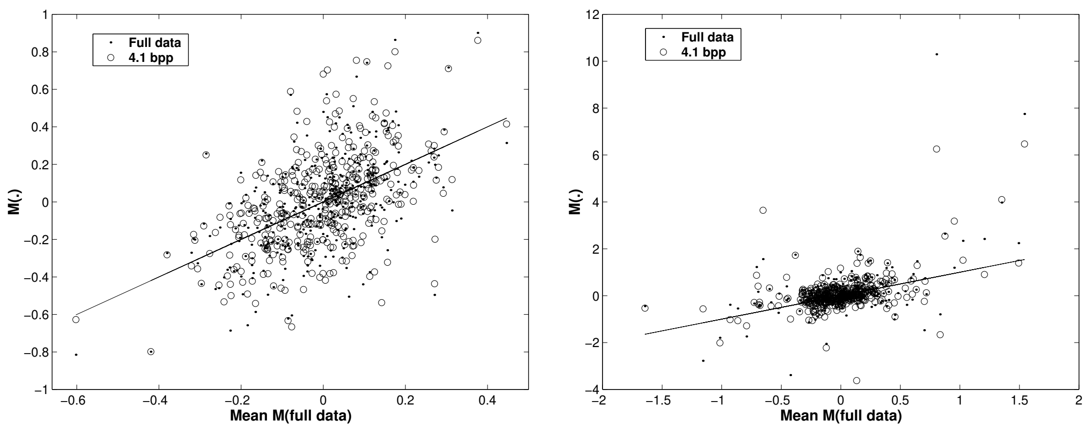
In figure 5, the risk, using Method 1, is plotted as a function of bit-rate, using the SPIHT, LOCO and segmented LOCO algorithms. SPIHT performs poorly



(a) (b)

Fig. 5. Method 1: (a) Risk (summed over all genes on array) vs bit-rate. Control group. (b) Treatment group.

at all bit-rates compared to the other algorithms. The segmented LOCO and standard LOCO show comparable risk results. However, there is a caveat in this comparison. We know that standard LOCO leads to a considerable bias in the estimate of A for low bit-rates, which figure 5 does not show. The overall risk for the control group is smaller than the treatment groups (figure 5 (a) vs (b)), since the (absolute) gene expressions levels extracted from the treatment arrays are overall larger. Both standard LOCO and segmented LOCO improves on the risk of the full data for bit-rates $\sim$ 6-8 bpp for the treatment arrays (b). We can think of the compression at these bit-rates as *denoising* of the microarray images. For both groups of arrays, we see only a marginal increase in risk for rates $>$ 4 bpp.



(a) (b)

Fig. 6. M(Array 2, full data) vs Mean M, (.), and M(Array 2, 4.1 bpp) vs Mean M , (o). Shrinkage toward Mean M for large absolute M values. (a) Method 1, (b) Method 2.

The equivalence of compression and denoising has been discussed in the literature previously (e.g. [15]). If the noise level is lower than the "signal", i.e. the extracted gene expression, we can get a better estimate of the gene expressions using compressed data, than we get using the full data. This is what figure 5 illustrates. Moreover, the effect is more apparent when the method of extraction is susceptible to noise (Method 2). In figure 6 the shrinkage, or denoising, effect of the compression on the extracted gene expressions are shown. The solid line is the mean gene expression vector M^0. The black dots correspond to the estimate of the gene expressions using a lossless reconstruction of Array 2. The black rings are obtained from a lossy reconstruction. We see that for large absolute values of M^0, the lossy reconstruction $\widehat{M}$ is closer to M^0 than the lossless reconstruction M. For the small and highly variable M^0, compression does neither good, nor harm. If there is no significant signal present, we have no chance of retrieving it, no matter which method we use. Of course, with only 8 replicates in the experiment, our ground truth may not be a very reliable estimate of the true gene expression. A more comprehensive study on a larger set of replicate images would allow more conclusive results about the effect of compression on the extraction of information from microarray images.

6 Conclusion and Future Work

We have presented a lossless and lossy compression scheme with successive refinements for microarray images. The flexible structure of our scheme allows for reconstruction of any subset of the images. At a compression ratio of 8:1 for cDNA images, we find that the tasks of genetic information extraction are only marginally affected by the compression. The effect of compression is smaller than the array-to-array variability. The effect is also smaller than the difference between alternative methods for information extraction. In fact, compression can have a denoising effect on the estimation of gene expression levels for cDNA images. We found that compression acts as a form of shrinkage for large absolute gene expression levels toward the mean of replicated arrays (taken as "truth").

As future work, we plan to formalize our bit allocation scheme in lossy SLOCO into a rigorous framework. Under this framework, we require a distortion function that measures the sensitivity of downstream information extraction to compression. Using this sensitivity function, we allocate bits among the different gene spots. To encode m independent random sources $X_1, \ldots, X_m$ (e.g. the spots) with R bits, the classic problem is how to allocate these bits to the different components to minimize the total distortion or sensitivity. For example, using local SNR as an indicator of this sensitivity, an analogy between our SNR-thresholded variable pixel-wise error bound and reverse water-filling seems natural. We spend more bits (by lowering the pixel-wise error) on spots with 1/SNR values greater than a threshold.

References

[1] P. Brown, `http://www-genome.stanford.edu` .

[2] `http://www.affymetrix.com` .

[3] `http://www.rii.com` .

[4] Y. Yang, M. Buckley, S. Dudoit, T. Speed, Comparisons of methods for image analysis on cdna microarray data, Tech. rep., Statistics, UC Berkeley, Berkeley, California (2000).

[5] M. Eisen, Scanalyze, `http://rana.stanford.edu/software` .

[6] R. Adams, L. Bischof, Seeded region growing, IEEE Trans Pattern Analysis and Machine Intelligence 16 (1994) 641–647.

[7] G.-A. Instruments, Genepix 4000a, user's guide .

[8] Y. Yang, S. Dudoit, P. Luu, T. Speed, Normalization for cdna microarray data, Tech. rep., SPIE BiOS, San Jose, California (Jan 2001).

[9] A. Said, W. Perlman, Reversible image compression via multiresolution representation and predictive coding, Proc SPIE 2094 (1993) 664–674.

[10] M. Nerhav, G. Seroussi, M. Weinberger, Modeling and low-complexity adaptive coding for image prediction residuals, Int'l Conference on Image Processing, Lausanne .

[11] M. Weinberger, G. Seroussi, G. Sapiro, Loco-i: A low complexity, context-based, lossless image compression algorithm, in: Proc. IEEE Data Compression Conf., 1996.

[12] R. Jornsten, Data compression and its statistical implications, Ph.D. thesis, University of California, Berkeley (2001).

[13] C. Lu, J. Dunham, Highly efficient coding schemes for contour lines based on chain code representation, IEEE trans Comm 39 (10) (1997) 1511–1514.

[14] Y. Yang, S. Dudoit, M. Callow, T. Speed, Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments, Tech. rep., Statistics, UC Berkeley, Berkeley, California (Aug 2000).

[15] M. Hansen, B. Yu, Wavelet thresholding via mdl for natural images, IEEE trans Inform. Theory, special issue on information theoretic imaging 46 (2000) 1778–1788.