Analysis-Driven Lossy Compression of DNA Microarray Images

Miguel Hernández-Cabronero*, Ian Blanes, *Member, IEEE*, Armando J. Pinho, *Member, IEEE*,
Michael W. Marcellin, *Fellow, IEEE* and Joan Serra-Sagristà, *Senior Member, IEEE*

***Abstract*—DNA microarrays are one of the fastest-growing new technologies in the field of genetic research, and DNA microarray images continue to grow in number and size. Since analysis techniques are under active and ongoing development, storage, transmission and sharing of DNA microarray images need be addressed, with compression playing a significant role. However, existing lossless coding algorithms yield only limited compression performance (compression ratios below 2:1), whereas lossy coding methods may introduce unacceptable distortions in the analysis process. This work introduces a novel Relative Quantizer (RQ), which employs non-uniform quantization intervals designed for improved compression while bounding the impact on the DNA microarray analysis. This quantizer constrains the maximum relative error introduced into quantized imagery, devoting higher precision to pixels critical to the analysis process. For suitable parameter choices, the resulting variations in the DNA microarray analysis are less than half of those inherent to the experimental variability. Experimental results reveal that appropriate analysis can still be performed for average compression ratios exceeding 4.5:1.**

***Index Terms*—DNA microarray images, Image compression, Quantization**

I. Introduction

The lossy compression of DNA microarray images can attain almost arbitrary compression ratios at the cost of distorting the results of subsequent analysis algorithms performed on them. Nevertheless, if the introduced distortion is smaller than the experimental variability that is inherent to DNA microarrays, the lossy compression can be considered acceptable [1]–[3]. Several generic image compression methods have been adapted or directly applied to DNA microarray images [1], [2], [4]–[6]. However, to the best of the authors' knowledge, no lossy compression technique specifically designed for microarray images has been published. This work aims to introduce such a technique with the goal of significantly outperforming existing lossy compressors.

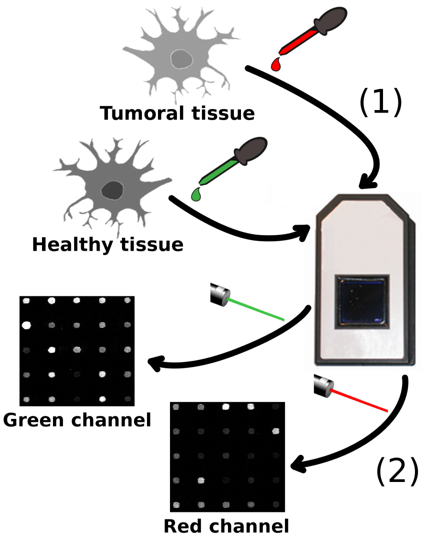


Fig. 1: Outline of an example DNA microarray image acquisition procedure. Samples from healthy and tumoral tissue are dyed with fluorescent pigments and put on a DNA microarray chip (1), which is optically scanned using two different wavelength lasers to produce two microarray images (2).

A. DNA Microarrays

DNA microarrays are widespread tools in biological and medical research. They are useful to analyze the function and regulation of individual genes from many organisms, including humans. The fight against Cancer, HIV and Malaria are among their most important applications.

In a typical DNA microarray experiment, two biological samples are compared. One sample corresponds to control (e.g., healthy) cells, and the other sample corresponds to experimental (e.g., tumoral) cells. A given gene can have different *expression intensities* –i.e., different amounts of activity– in the two biological samples. By studying the expression intensity differences between these two biological samples, it is possible to analyze the function of each gene in an illness or in other biological processes.

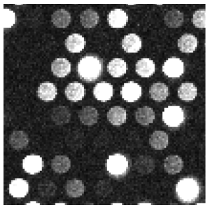
Samples coming from the healthy and tumoral tissues are first dyed with, respectively, green and red fluorescent markers (step (1) in Fig. 1). After that, the biological samples are left to react on the surface of the DNA microarray chip, which contains microscopic holes or *spots* arranged in a regular grid, as shown in Fig. 2a. Each of the spots is related to a single gene of the organism, and the quantity of each dyed biological

This work has been partially funded by FEDER, the Spanish Ministry of Economy and Competitiveness (MINECO) and the Catalan Government under projects TIN2012-38102-C03-03, FPU AP2010-0172 and 2014SGR-691.

*M. Hernández-Cabronero, I. Blanes and J. Serra-Sagristà are with the Department of Information and Communications Engineering, Universitat Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain (e-mail: mhernandez@deic.uab.cat).

A. J. Pinho is with the Signal Processing Lab, DETI/IEETA, University of Aveiro, 3810-193 Aveiro, Portugal.

M. W. Marcellin is with the Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ 85721-0104, USA.



(a)

22831	**21698**	**21631**	**11488**
20030	**23157**	**13673**	**4524**
17021	**10024**	**3330**	869
3779	**1408**	242	1
1223	41	314	1055
871	179	284	46

(b)

Fig. 2: Example DNA microarray images. (a) 100×100 crop of *slide_1-red* image from the Arizona corpus with hexagonal grid spot layout. Gamma levels have been adjusted for visualization purposes; (b) Pixel intensity values of a 6×4 crop of *134044022_Cy5* from the IBB corpus. Pixels belonging to a spot are highlighted in bold font.

sample that remains in it is proportional to the activity of that gene in the corresponding biological sample. The chip is then optically scanned while exciting the fluorescent marker used to dye one of the biological samples (step (2) in Fig. 1). This results in an unsigned 16 bit per pixel (bpp) grayscale image. The chip is scanned again while exciting the other fluorescent marker in order to produce a second 16 bpp grayscale image. In each of these images –usually referred to as the green and red channels because of the associated dye color– the brightness of each spot is related to the activity of the gene related to that spot in the corresponding biological sample.

Once DNA microarray images have been obtained, microarray-specific image analysis software is employed to quantify the genetic expression intensities in each of the biological samples. Finally, the extracted data are processed to detect relevant genetic expression differences between the control and experimental tissue samples, which enables the study of the function of individual genes.

DNA microarray image analysis is an active research field [7]–[13]. As new analysis techniques are developed, it will be possible to re-analyze existing images to obtain more accurate genetic data. Since it is not practical to preserve the biological samples indefinitely nor share them among laboratories around the world, replicating the whole DNA microarray experiment is usually not feasible or convenient. A preferable alternative is to store the DNA microarray images. Image coding techniques can help alleviate the costs associated with the storage and management of this data, and can also accelerate their transmission to other researchers wishing to perform analysis (or re-analysis with new techniques).

B. Compression of DNA Microarray Images

DNA microarray images present several features that render their compression a very challenging task. In each of the grayscale images, thousands of round spots of varying intensities are displayed on a dark background following a regular pattern. A crop of an example DNA microarray image with hexagonal grid is shown in Fig. 2a. As a consequence of the abrupt pixel intensity variations induced by the spots, as shown in Fig. 2b, DNA microarray images contain high frequencies which are hard to code efficiently. Furthermore, the original image data are represented with 16 bpp, and typically 7 or more of the least significant bitplanes exhibit binary entropy values close to 1 bpp [14].

A complete review of the state of the art in both lossless and lossy compression of DNA microarray images can be found in [14]. When lossless compression is employed, perfect pixel fidelity is guaranteed. However, the best reported lossless compression ratios (summarized later in Table V) are smaller than 2:1 for most corpora. This is believed to be a practical bound to lossless compression methods [1], [15].

Lossy compression, on the other hand, can provide essentially any desired compression ratio, but at the expense of introducing changes (distortion) in the image data. Depending on this distortion, the results for current and future image analysis methods may be severely affected, which may render images unusable. For this reason, it is necessary to assess the impact of lossy compression on the analysis of DNA microarray images. Previous work has indicated that lossy compression can produce acceptable results when the distortion introduced is smaller than the variability observed in replicated experiments [1]–[3].

To the best of the authors' knowledge, no existing compression technique in the literature has been designed to directly take into account the DNA microarray image analysis process (e.g., [1], [2], [4]–[6]). The main novelty of this work consists in the design of a lossy compression method specifically conceived to minimize its impact on subsequent analysis processes applied on DNA microarray images. In particular, it aims at providing significant reductions in errors introduced into the CRM, compared to existing lossy coding techniques.

C. Paper Structure

A Relative Quantizer (RQ) designed for DNA microarray images is proposed in Section II and its impact on the genetic data extraction process is addressed in Section III. The effectiveness of (further) lossless compression on images that have been quantized using the RQ is discussed in Section IV. Some conclusions are drawn in Section V.

II. The Relative Quantizer

A. Motivation

In a DNA microarray image, the brightness of each spot is related to the expression intensity of the gene (in the biological sample) associated with that spot. In order to quantify the expression intensities for the different genes under test, microarray image analysis techniques *segment* the red and green channel to detect the position of the spots and differentiate spot pixels from background pixels. A recent review of the state of the art on microarray image segmentation can be found in [12]. The positions and shapes of the spots are not perfectly regular, so that segmentation is a challenging task.

If the red and green images are subjected to lossy compression prior to analysis, the resulting distortion may cause the segmentation process to fail to detect a spot in at least one of the two images and no genetic information will be subsequently extracted from it. A spot that is correctly detected in both images is hereinafter referred to as *positively detected*.

Even if a spot is positively detected, pixels belonging to the spot may be incorrectly tagged as background and *vice versa*. Thus, the segmentation step is crucial in the analysis process. Since a large fraction of the spots have low intensities [14], the absolute distortion introduced in low-intensity pixels should be limited, so that spots can be accurately separated from the dark background.

After the spots are segmented, the pixel values from the co-located spots (i.e., the spot at the same location of the red and green channel images) are compared to assess whether the gene corresponding to that spot is expressed differently in the two biological samples. Even though spots can exhibit different sizes in the two images, approximate co-location in both images is necessary for a correct detection and comparison of these spots. To this end, professionals working with DNA microarrays usually employ the *corrected ratio of means* (CRM) of each positively detected spot [12], defined as

$$\mathrm{CRM} \quad = \quad \frac{\mu_{\mathrm{spot}}^{\mathrm{red}} - \mu_{\mathrm{localBG}}^{\mathrm{red}}}{\mu_{\mathrm{spot}}^{\mathrm{green}} - \mu_{\mathrm{localBG}}^{\mathrm{green}}}. \tag{1}$$

Here, μ_{spot} and μ_{localBG} are the average pixel intensity within a spot and its *local background*, while the red and green superscripts refer to the channel being analyzed, respectively. The μ_{localBG} is subtracted from μ_{spot} to compensate for background noise and unavoidable inaccuracies in the segmentation process. The local background of each spot is calculated as a region of background pixels surrounding the spot. The exact shape and size of the local background is determined by the segmentation algorithm. Several local background definitions have been proposed in the literature [16]. Three of these are illustrated in Fig. 3: a circular crown which surrounds but does not include the spot pixels (depicted in pink), a square crown (depicted in yellow), and the union of four diamond-shaped regions (depicted in orange). In this figure, the spot is depicted by the gray circle.

For more than 20 years, statistical analyses of DNA microarray data have consistently relied on segmenting images into spots and then extracting a CRM value for each pair of spatially corresponding spots [17]. Hence, it is reasonable to expect future image analysis techniques to also be based on this principle. Therefore, lossy compression methods applied to DNA microarray images should aim to minimize the impact on this type of analysis. As mentioned previously, certain aspects of the analysis have been the subject of intensive recent research [7]–[13]. The proposed Relative Quantizer is designed specifically to minimize the impact to segmentation and to the ratios of spot intensities (CRM), regardless of the specific algorithmic details used within the analysis process.

In what follows, the error introduced in the CRM is taken as a measure of distortion introduced by lossy compression within positively detected spots. Because the CRM is defined as a quotient, the absolute error introduced in the image intensities is not enough to characterize the impact on the CRM. For instance, an absolute error of $\varepsilon_{\mathrm{abs}}$ in the numerator of (1) will induce different absolute errors in the CRM depending on the value of the denominator of (1). For example, the absolute error in the CRM will be 2 times larger for a denominator of value d than for a denominator of value $2d$. On the other hand, if a relative error of $\varepsilon_{\mathrm{rel}}$ is introduced in the numerator, the same relative error is introduced in the CRM, regardless of the value of the denominator. In this paper, the relative error is defined as $\varepsilon_{\mathrm{rel}} = \varepsilon_{\mathrm{abs}}/v_{\mathrm{original}}$, where v_{original} is the original value without errors. Therefore, it is arguably more useful to limit the relative error than to limit the absolute error. That is, the error introduced in each pixel should be bounded by a certain percentage of the original pixel value. This is in stark contrast to traditional lossy compression algorithms, which attempt to limit the squared (absolute) error.

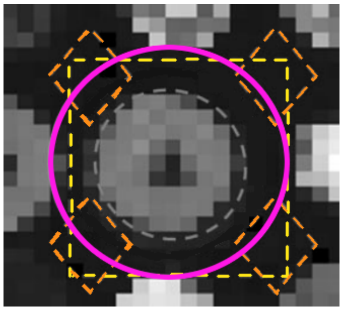


Fig. 3: Different local background definitions.

B. Definition and Properties

In what follows, we assume that DNA microarray images are analyzed subsequent to lossy compression. Motivated by the discussion above, we propose a Relative Quantizer (RQ) designed to provide superior compression performance for DNA microarray images while limiting the impact on the analysis of these images. Specifically, the quantizer is designed to have minimal impact on segmentation, as well as on CRM values. The impact on segmentation is controlled by limiting errors in the pixels having small values, while errors in the CRM are controlled by limiting the pixel-wise relative error, thence the name “Relative”.

The fixed-rate scalar quantizer that minimizes relative error for continuous-amplitude sources has been described in the literature [18]. For sources with probability density functions equal to $f(x) = a/x$, $a \in \mathbb{R}$, the optimal solution is a logarithmic quantizer DNA microarray image pixel distributions, in which low values are much more probable than high values [14], can be approximated by such density functions. Therefore, the design of the proposed RQ is based on the logarithmic quantizer to minimize the pixel-wise relative error and, in turn, control the impact on the CRM values extracted from the quantized images. On the other hand, the proposed RQ is designed for discrete-amplitude (integer pixel) sources, rather than continuous-amplitude (real number) sources. Additionally, in order to minimize the impact on the spot segmentation, the RQ further prioritizes low-intensity pixels. As explained above, low-intensity pixels are critical to detect spots and separate them from the dark background of DNA microarray images. Specifically, as described in detail below, low-intensity pixels that fall within a prescribed range are guaranteed to be preserved perfectly.

The RQ is applied independently to each pixel of the original image. Each such pixel is assumed to be an unsigned

TABLE I: Original pixel values (Orig.), quantization indices (QI) and reconstructed values (Rec.) for the RQ using $B = 4$ and $k = 2$. Bits preserved in each value are highlighted in bold font. The interval midpoint rounded up is employed for the reconstruction.

Orig.	0 **0000**	1 **0001**	2 **0010**	3 **0011**	4 **010**0	5 **010**1	6 **011**0	7 **011**1	8 **10**00	9 **10**01	10 **10**10	11 **10**11	12 **11**00	13 **11**01	14 **11**10	15 **11**11
QI	0	1	2	3	4	4	5	5	6	6	6	6	7	7	7	7
Rec.	**0000** 0	**0001** 1	**0010** 2	**0011** 3	**010**1 5	**010**1 5	**011**1 7	**011**1 7	**10**10 10	**10**10 10	**10**10 10	**10**10 10	**11**10 14	**11**10 14	**11**10 14	**11**10 14

integer of bitdepth $B \geq 1$. The RQ is parameterized by an integer k in $\{1, \dots, B\}$. As explained in detail below, the parameter k determines the number of quantization intervals and their size, which in turn controls the fidelity with which each individual pixel is preserved. Small values of k yield low fidelity, while increasing k yields higher fidelity. Indeed, when $k = B$, the size of all quantization intervals is exactly 1, so no quantization is applied, and all pixels are preserved perfectly. In order to describe the quantization intervals, it is useful to consider pixel values in their binary representation. For a given pixel, let N be the position of its most significant bit having value equal to 1, where $B - 1$ and 0 are the most and least significant positions, respectively. For example, let $B = 4$. Then pixels having values $v_1 = 0001_2$, $v_2 = 0100_2$ and $v_3 = 0101_2$ have $N_1 = 0$, $N_2 = 2$ and $N_3 = 2$, respectively.

The main idea of the RQ is to then preserve only the bits in positions $B - 1, \dots, N - k + 1$. Note that, by definition, only the k bits in positions $N, ..., N - k + 1$, can be different from 0. Thus, the parameter k effectively determines the maximum number of non-null bits that are preserved for each pixel. From this observations, it follows that if $N < k$, then all bits of the pixels are preserved. That is, all pixels having values in $\{0, 1, \dots, 2^k - 1\}$ are preserved losslessly.

Table I shows the operation of the RQ for $B = 4$ and $k = 2$. The first two rows in the table show the decimal and binary representations for each possible pixel value. The bits to be preserved are highlighted in bold font. Pixel values that are identical in the preserved positions are assigned to the same quantization interval, and hence, have the same quantization index, as given in the third row. The fourth and fifth rows show the binary and decimal representations of the reconstructed pixel values at the output of the dequantizer. The interval mid-point rounded up to the next integer has been used for reconstruction. As an example, two pixels taking values $\mathbf{010}0_2$ and $\mathbf{010}1_2$ belong to the same quantization interval. They share a common quantization index of 4, and are both reconstructed as 5. As expected, pixels having values less than $2^k = 4$ are preserved perfectly.

As seen in Table I, when $B = 4$ and $k = 2$, there are 8 distinct quantizer indices. To calculate the number of quantizer indices for arbitrary B and k, it is illustrative to view the RQ as a quantizer having non-uniform intervals. For any choice of B, the first 2^k intervals correspond to preserving all bits of any pixel having value $0 \leq p < 2^k$. Each interval thus contains only one value. That is, each interval is of size $2^0 = 1$. The next 2^{k-1} intervals correspond to preserving all but the least significant bit of any pixel with value $2^k \leq p < 2^{k+1}$. Hence, two values are assigned to each interval. That is, each interval is of size 2^1. The next 2^{k-1} intervals correspond to preserving all but the two least significant bits of any pixel with value $2^{k+1} \leq p < 2^{k+2}$. Each such interval is of size 2^2. Each subsequent group of 2^{k-1} intervals has size 2^3, 2^4, etc. Finally, the 2^{k-1} intervals of the last group each have size 2^{B-k}. This last group of intervals preserves the k most significant bits of any pixel having value $2^{B-1} \leq p < 2^B$.

For any values of B and k, the exact number of quantization intervals I_k can then be easily calculated. Since there are 2^k intervals of size 1 and 2^{k-1} intervals of each size s with $s \in \{2^1, 2^2, \dots, 2^{B-k}\}$, there are exactly $2^k + (B - k)2^{k-1} = (B - k + 2)2^{k-1}$ quantization intervals. Table II provides I_k for several values of B and k.

TABLE II: Number of quantization intervals I_k for the RQ using $B = 4$ and $B = 16$.

$\boldsymbol{k}$	**1**	**2**	**3**	**4**	**5**	**6**	**7**
$\boldsymbol{I_k}$ $(B = 4)$	5	8	12	16	N/A	N/A	N/A
$\boldsymbol{I_k}$ $(B = 16)$	17	32	60	112	208	384	704

In summary, no error is incurred in the 2^k lowest pixel values. Additionally, several of the next quantization interval groups have small lengths: 2, 4, 8, 16, etc., implying small maximum errors. As discussed in Section II-A, low intensity pixels are crucial for the spot segmentation. Thus, the small maximum error introduced by the RQ in this intensity range attenuates the impact on the segmentation process. Moreover, the maximum relative error in each pixel is bounded. Specifically, pixels having values $2^{k+j} \leq p < 2^{k+j+1}$ are quantized using an interval of size exactly 2^{j+1}, $j = 0, 1, \dots, B-(k+1)$. It follows that the absolute error introduced in a pixel by quantization/dequantization is at most $\varepsilon_{\text{abs}} = 2^j$, so that the maximum relative error is bounded by $\varepsilon_{\text{rel}} = 2^j / 2^{k+j} = 2^{-k}$. As explained in Section II-A, limiting the pixel-wise relative error helps control the distortion in the extracted CRM values. Note that this approach is not specifically designed for any concrete DNA microarray image analysis algorithm. Instead, the Relative Quantizer relies on preserving the intensity ratios, and not on the specific way in which a particular algorithm segments the images. Hence, using the proposed approach is reasonable for any existing or future analysis algorithm based on the CRM.

III. Impact of the Relative Quantizer on Genetic Data Extraction

A. Distortion Metrics

The main drawback of lossy coding methods applied to DNA microarray images is the possibility of distorting the results of any subsequent genetic data extraction process. As explained in Section II-A, the segmentation step may fail to detect one or more spots. Also, the *corrected ratio of means* (CRM) values extracted for positively detected spots may be distorted.

CRM values are usually classified into one of three categories: *a)* CRM $< \alpha$, *b)* CRM $> \beta$ or *c)* CRM $\in [\alpha, \beta]$. Typically, $\alpha = 0.5$ and $\beta = 2$ [3], so that if CRM $< 1/2$ for a given spot (Category *a)*), then the gene associated with that spot is declared to be more strongly expressed in the "green sample" (that is, the biological sample associated with the green florescent marker) than in the red sample. Indeed, CRM $< 1/2$ implies that the denominator of Eq (1) is more than twice as large as the numerator. Similarly, if CRM > 2 (Category *b)*), then the gene is declared to be more strongly expressed in the red sample. Finally, Category *c)* indicates that the gene is expressed roughly equally in both samples. This classification is usually the only output considered, and experts from the Genomics and Bioinformatics Service of the Biology and Biomedicine Institute (IBB) at the Universitat Autònoma de Barcelona (UAB) agree that any lossy process for which no detection errors occur and the extracted CRM values remain unmodified is equivalent to a numerically lossless process.

In practice, most CRM values are close to 1, and the fraction of spots whose CRM lies outside $[\alpha, \beta]$ is relatively small and depends on the type of biological samples being compared. A histogram of the CRM values extracted from the positively detected spots of all 22 image pairs of the IBB corpus is shown in Fig. 4. As reported later in Table IV, the average number of spots in the IBB set is about 14,000. For visualization purposes, all values smaller than -1 are considered to be exactly -1, and all values larger than 5 are considered to be exactly 5. Note that according to the CRM definition of (1), negative CRM values can occur when the average intensity of the local background is larger than the average intensity of the spot pixels.

Based on this, two full-reference distortion metrics are defined below to assess the acceptability of the changes introduced in the images by lossy processes, including the proposed RQ. The first one is the *average relative error in the CRM* ($\mathrm{ARE}_{\mathrm{CRM}}$). Given the analysis results of an original and a distorted (e.g., quantized) image pair, it is defined as

$$\mathrm{ARE}_{\mathrm{CRM}} = \frac{1}{n} \sum_{i=1}^{n} \frac{|\mathrm{CRM}_i - \widehat{\mathrm{CRM}}_i|}{\delta + |\mathrm{CRM}_i|}. \tag{2}$$

Here, n is the number of spots positively detected in both the original and the distorted image pairs. The CRM extracted from the i-th such spot in the original and distorted image pairs are denoted as CRM_i and $\widehat{\mathrm{CRM}}_i$, respectively. The parameter δ is set to 0.001 to stabilize the case $\mathrm{CRM}_i = 0$. As an example, a value of $\mathrm{ARE}_{\mathrm{CRM}} = 0.5$ would indicate that, on average, the distorted CRM values differ by 50% of their original values. This metric provides insight on the global distortion in the analysis process. Similar analysis distortion metrics have been employed in the literature [1], [2].

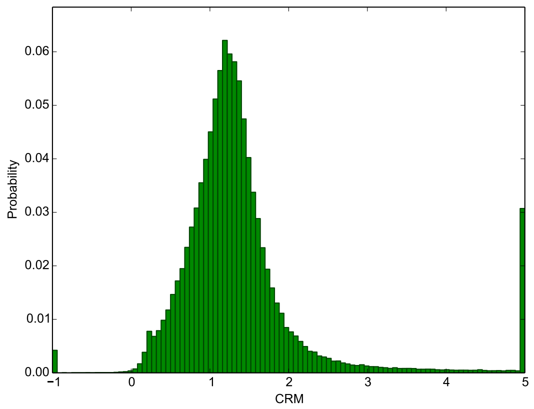


Fig. 4: Distribution of the CRM values extracted from all 22 original image pairs of the IBB corpus.

The second metric is the *fraction of spots wrongly detected or classified* (FWDOC). It is defined as

$$\mathrm{FWDOC} = (d + c)/m, \tag{3}$$

where d is the number of spots that are detected differently in the original and quantized image pairs, c is the number of spots that are positively detected in both the original and quantized image pairs but are classified differently, and m is the total number of spots. Here, a spot is considered to be detected differently if it is positively detected in the original image pairs but not in the quantized image pairs, or *vice versa*. Note that m includes both positively detected and not positively detected spots and, hence, $m \geq n$. Similar approaches have been used in [2] and [3]. This metric quantifies the probability of a spot becoming unusable because of the introduced distortion. As suggested by the IBB experts, the interval $R = [\alpha, \beta] = [0.5, 2]$ is employed in this work to perform all classification operations.

B. Distortion Results

A number of tests have been carried out to evaluate the performance of the RQ with respect to microarray images. The first such test was to evaluate the distortion resulting from the RQ for various values of k. This test was performed using a corpus of 44 images (22 red/green image pairs), obtained from real experiments at the IBB, hereinafter referred to as the IBB corpus. Specifically, all images from the corpus were quantized by the proposed RQ using $k \in \{1, \ldots, 7\}$. The images were then reconstructed from quantization indices by employing interval mid-points rounded up to the next integer. The original IBB corpus and the 7 reconstructed versions were analyzed with the GenePix software at the IBB [19]. The results for the original and the quantized versions were compared using the two metrics described in Section III-A.

In the IBB corpus, each spot is replicated, i.e., there are 2 spots devoted to each gene. Ideally, identical segmentation and CRM results should be obtained for both spots. However, in real experiments, they differ *even in the original images before quantization*. Thus, two more metrics have been derived from (2) and (3) to calculate the variability present between pairs of replicated spots in the original images. Given the analysis results of an original image pair, the *replicate ARE*$_{\text{CRM}}$ (Rep-ARE$_{\text{CRM}}$) is defined as

$$\text{Rep-ARE}_{\text{CRM}} = \frac{1}{p}\sum_{i=1}^{p}\frac{|\text{CRM}_i^{\text{f}} - \text{CRM}_i^{\text{s}}|}{\delta + |\text{CRM}_i^{\text{f}} + \text{CRM}_i^{\text{s}}|/2}, \qquad (4)$$

where CRM_i^{f} and CRM_i^{s} are, respectively, the CRM of the first and second spots of the i-th replicated spot pair, p is the number of such spot pairs where both spots are positively detected, and $\delta = 0.001$. Similarly, the *replicate fraction of spots wrongly detected or classified* (Rep-FWDOC) is defined as

$$\text{Rep-FWDOC} = (d_{\text{pair}} + c_{\text{pair}})/q, \qquad (5)$$

where d_{pair} and c_{pair} are the number of pairs whose spots are differently detected or classified, respectively, and q is the total number of pairs in the image. Since not all spots are necessarily detected, $q \geq p$.

Results for the quantized images and for the replicated spots in the original images are provided in Table III. In the most aggressive case ($k = 1$), large errors are apparent, especially in the ARE$_{\text{CRM}}$. Nevertheless, rapid improvement is observed as the parameter k is increased. For $k \geq 4$, the ARE$_{\text{CRM}}$ and the FWDOC are below 8.0% and 4.5%, respectively. Significantly, for all $k > 1$, the ARE$_{\text{CRM}}$ and the FWDOC metrics show a better behavior than the Rep-ARE$_{\text{CRM}}$ and the Rep-FWDOC for the replicated spots in the original images. In the literature on lossy compression of DNA microarray images, distortions smaller than the experimental variability are considered acceptable [1]–[3]. The distortion among replicated spots can be understood as a measure of this variability. In this light, the results of Table III suggest that the proposed RQ yields acceptable distortions for all $k > 1$.

Arguably, the selection of a suitable value for k might be specific to the scanner and analysis software employed. Given a set of images and analysis software appropriate for the scanner from which the images were acquired, a conservative approach might be to select a value of k for which the average distortion measured by the metrics proposed in (2) and (3) are between one half and one third of the replicate variability as defined in (4) and (5), respectively. For the IBB corpus and the GenePix analysis software, this leads to the choice of $k = 3$ or $k = 4$.

TABLE III: Average relative error in the CRM (*ARE*$_{CRM}$) and fraction of spots wrongly detected or classified (*FWDOC*) after the RQ. Results have been averaged over all 22 image pairs of the IBB corpus. Average data for the pairs of replicated spots in the original images (the *Rep-ARE*$_{CRM}$ and *Rep-FWDOC* metrics) are provided at the bottom.

Images	**ARE$_{\text{CRM}}$**	**FWDOC**
Original vs. RQ $k = 1$	0.562	0.148
Original vs. RQ $k = 2$	0.124	0.100
Original vs. RQ $k = 3$	0.121	0.064
Original vs. RQ $k = 4$	0.078	0.044
Original vs. RQ $k = 5$	0.064	0.030
Original vs. RQ $k = 6$	0.039	0.019
Original vs. RQ $k = 7$	0.028	0.014
Images	**Rep-ARE$_{\text{CRM}}$**	**Rep-FWDOC**
Original	0.254	0.212

Results for the test described in this section have not been obtained for DNA microarray images from other corpora. Other corpora employed in the literature either do not consist of green/red channel pairs from the same DNA microarray experiment, or no compatible analysis software is publicly available. Thus, an exhaustive study on the impact of k on the analysis of such corpora is beyond the scope of this work. Nevertheless, the properties of the RQ described in Section II-B (bounded relative error for all pixels and small absolute error for low-intensity pixels) do not depend on the source of the image being quantized. Moreover, since the *maximum* relative error of 2^{-k} quickly decreases as k is increased, it is reasonable to expect the analysis distortion to be a monotonically decreasing function of k for any image set, and that a very small analysis distortion should be obtained for any image whenever $k > 5$.

Additional tests that employ the IBB corpus, as well as other corpora from the literature, are discussed in the next section.

IV. Lossless Coding of RQ Indices

The previous section characterized the distortion introduced to DNA microarray images by the proposed RQ as a function of the parameter k. In this section, compression performance is discussed. To this end, several techniques for storing the RQ indices of quantized images are investigated. These techniques can be considered as lossless coding (or compression). For each technique, the RQ indices can be recovered exactly, and thus the distortion introduced into the images is entirely independent of the lossless coding technique employed. Indeed, the resulting decompressed images are identical for each and every technique. The only difference between the differing lossless compression techniques is the resulting compression performance. In what follows, compression performance is reported in terms of the average rate R in bits per pixel (bpp) required to store the indices of an image. This quantity can be converted into compression ratio for a 16-bit image as $CR = 16/R$.

A. DNA Microarray Image Corpora

A total of 228 DNA microarray images in 7 corpora produced by different types of scanners have been gathered to evaluate the lossless compression of indices produced by the proposed RQ. All images most often used for benchmarking in the DNA microarray image compression literature –the Yeast, the ApoA1, the ISREC, the Stanford and the MicroZip corpora– have been included. Additionally, the Arizona and IBB corpora, which contain images representative of the output of more modern DNA microarray scanners, have been included. Table IV summarizes some of the most important image characteristics. In particular, the total number of *grayscale* images in each corpus is provided in the *Images*

TABLE IV: Image corpora used for benchmarking in this work. Original image pixels are unsigned 16-bit integers.

Property	**Yeast [20]**	**ApoA1 [21]**	**ISREC [22]**	**Stanford [23]**	**MicroZip [4]**	**Arizona [24]**	**IBB [25]**
Year	1998	2001	2001	2001	2004	2011	2013
Images	109	32	14	20	3	6	44
Size	1024×1024	1044×1041	1000×1000	$>$ 2000×2000	$>$ 1800×1900	4400×13800	2019×6235
Spot count	$\sim 9 \cdot 10^3$	$\sim 6 \cdot 10^3$	$\sim 2 \cdot 10^2$	$\sim 4 \cdot 10^3$	$\sim 9 \cdot 10^3$	$\sim 2 \cdot 10^5$	$\sim 1.4 \cdot 10^4$
Avg. intensities	5.39%	39.51%	33.34%	28.83%	37.71%	82.82%	54.07%
Avg. entropy (bpp)	6.628	11.033	10.435	8.293	9.831	9.306	8.503
Best rate (bpp) [26]	5.521	10.223	10.199	7.335	8.667	8.275	8.039

row. All images are 16 bpp. Some of the corpora do not contain green/red channel pairs, which yields an odd number of grayscale images in some cases. Along with the spatial size for each corpus, the average number of spots in each image of a given corpus is reported in the *Spot count* row. The percentage of the 2^{16} possible pixel intensity values that are actually present in each image has been computed, and the average percentage for each corpus is reported in the *Avg. intensities* row. The average first-order entropy of each corpus is reported in the row labeled *Avg. entropy*. Results for the best method known for lossless DNA microarray compression [26] are expressed in terms of bpp in the last row. These results were obtained for the original unquantized images using an implementation provided by the authors of [26]. For each corpus, the reported results are better than the first-order entropy, due to the fact that pixel dependencies are effectively exploited by the coding method employed.

B. Compression Experiments

All images from the described corpora were quantized by the proposed Relative Quantizer (RQ) and the quantization indices were stored as an image. Each resulting index image was then subjected to lossless compression using several lossless algorithms. As mentioned previously, each of these algorithms may yield a different value for compression performance (in bpp). However, the resulting image distortion (and indeed, the images themselves) will be identical for each algorithm. This experiment was performed for each value of $k \in \{1, 2, \ldots, 7\}$. Since $k = 7$ already yields analysis distortions 10 to 20 times smaller than the experimental variability (see Section III-B), larger values of k have not been considered here.

The tested lossless coding algorithms include generic data compressors (bzip2), image and video compressors not specifically designed for DNA microarrays (JPEG-LS [27], CALIC [28], lossless JPEG2000 [29] and lossless HEVC/H.265 [30]), and the best lossless microarray-specific image compressor (Neves and Pinho's method [26]). Unless stated otherwise, all codecs were invoked with default parameters. The HEVC coder was invoked using coding unit size equal to 64×64, and bypassing the lossy stages (which would be enabled by default). For the HEVC experiments, all images were stored using YUV 4:2:2 format, with the U and V components being exactly zero. The exact configuration parameters employed for HEVC are available at http://deic.uab.es/~mhernandez/media/software/hevc_lossless.cfg. Note that publicly available CALIC codecs support only images up to 8 bpp, i.e., 256 different pixel values. As can be seen from Table II, the proposed RQ yields index images with 384 or more intensities whenever $k \geq 6$. Therefore, CALIC has only been applied for $k < 6$. The H.264 standard has not been included in this study since it does not support image sizes large enough for many DNA microarray images [31].

In addition to the variable-rate methods listed in the previous paragraph, fixed-rate coding is also considered. The simplest such strategy is to assign to each index a fixed length codeword of length $\lceil \log_2 I_k \rceil$ bits. For example, when $k = 4$ (and $B = 16$), a codeword length of $\lceil \log_2 112 \rceil = 7$ bits will suffice. If a block of L indices are coded together, a codeword of length $\lceil \log_2 I_k^L \rceil$ will suffice. The rate of the resulting code is then $\frac{1}{L} \lceil \log_2 I_k^L \rceil < \log_2 I_k + \frac{1}{L}$. Thus, fixed length coding can approach $\log_2 I_k$ bits per pixel as closely as desired.

Table V presents the results obtained for each value of k by the different coding techniques. Data for the original images before quantization is also provided. Results are given in bits per pixel, calculated as the combined size in bits of all compressed images divided by the total number of pixels in a corpus. The fixed-rate results do not depend on the corpus and are reported once at the top of the table for block size $L = 2000$. The *Lossless JPEG2000* row contains results for Kakadu v7.4 using the best choice of parameters for each column[1].

As expected, the bitrate decreases (compression ratio increases) as k is decreased for all tested coders. For example, bitrate reductions of over 20% are observed for $k = 7$, as compared to the original images. As another example, 60% reductions are observed for $k = 3$. For a specific corpus and specific value of k, the rates resulting from different coders are typically within 0.5 bits per pixel, with several notable exceptions. For example, the state-of-the-art video coder HEVC/H.265 produces very poor results for the original images, but provides more competitive performance for RQ index images. For every value of k, the best-performing coder for all image corpora is that of Neves and Pinho [26]. This should be expected, at least for the case of the original DNA microarray images, for which it was designed. Therefore, Neves' method is hereinafter used to losslessly code the quantization indices produced by the proposed Relative Quantizer.

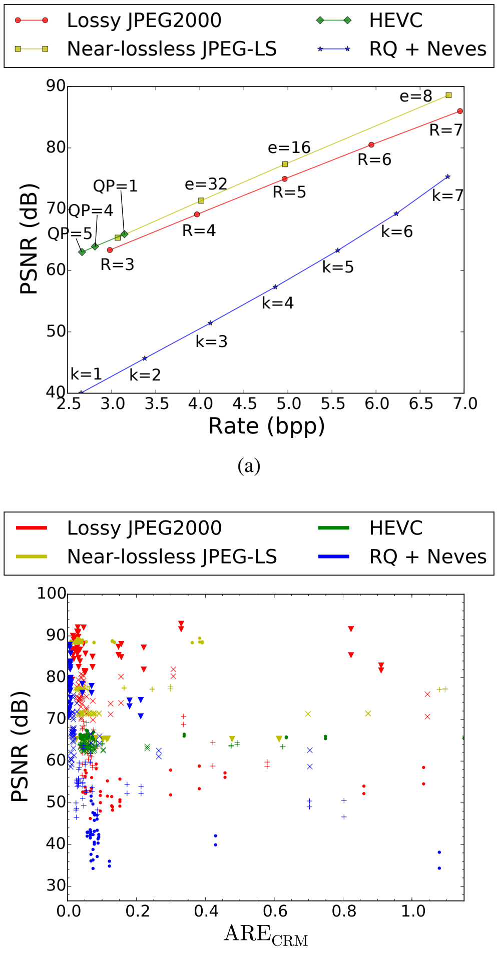
C. Rate-distortion Analysis

The previous sections have demonstrated that compression systems based on the proposed RQ can provide significant

[1] Ssigned=no and 0 wavelet decomposition levels for $1 \leq k \leq 4$; Ssigned=no and 3 DWT decomposition levels for $5 \leq k \leq 6$; Ssigned=yes and 0 DWT decomposition levels for $k = 7$ and for the original images.

TABLE V: Compression results in bpp for RQ followed by different lossless coding algorithms. Lossless compression results for the original images are also provided.

Corpus	**Algorithm**	**RQ index images**							**Original images**
		$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	
	Fixed-length coder	4.087	5.000	5.907	6.807	7.700	8.585	9.459	16.000
Yeast	Average entropy	1.854	2.474	3.272	4.156	5.074	5.945	6.294	6.628
	bzip2	1.028	1.614	2.462	3.399	4.355	5.250	5.655	6.075
	JPEG-LS	1.007	1.497	2.231	3.082	3.986	4.989	5.892	8.580
	CALIC	0.977	1.503	2.268	3.075	3.940	–	–	–
	Lossless JPEG2000	1.355	1.473	2.282	3.417	4.339	5.308	6.219	5.903
	HEVC/H.265	1.241	1.844	2.632	3.532	4.495	5.532	6.650	10.660
	Neves & Pinho	0.900	1.339	2.017	2.921	3.887	4.769	5.056	5.511
ApoA1	Average entropy	1.704	2.504	3.442	4.423	5.417	6.415	7.414	11.033
	bzip2	1.357	2.121	3.062	4.052	5.063	6.090	7.106	11.064
	JPEG-LS	1.258	1.921	2.746	3.691	4.680	5.728	6.698	10.606
	CALIC	1.202	1.889	2.729	3.620	4.588	–	–	–
	Lossless JPEG2000	1.404	1.930	2.822	3.758	4.859	5.896	7.518	10.787
	HEVC/H.265	1.348	2.054	2.943	3.936	5.009	6.168	7.408	14.482
	Neves & Pinho	1.041	1.715	2.604	3.565	4.562	5.557	6.565	10.223
ISREC	Average entropy	2.674	3.617	4.597	5.585	6.543	7.442	8.277	10.435
	bzip2	2.681	3.621	4.604	5.599	6.561	7.476	8.373	10.921
	JPEG-LS	2.725	3.671	4.663	5.660	6.670	7.601	8.494	11.145
	CALIC	2.639	3.526	4.482	5.471	6.464	–	–	–
	Lossless JPEG2000	2.690	3.518	4.536	5.575	6.703	7.695	8.491	10.625
	HEVC/H.265	2.623	3.618	4.705	5.880	7.102	8.503	10.077	14.876
	Neves & Pinho	2.403	3.317	4.291	5.281	6.241	7.144	7.976	10.199
Stanford	Average entropy	2.021	2.863	3.801	4.785	5.777	6.662	7.268	8.293
	bzip2	1.415	2.205	3.107	4.090	5.098	5.982	6.553	7.887
	JPEG-LS	1.343	1.974	2.839	3.802	4.796	5.700	6.241	7.597
	CALIC	1.230	2.003	2.786	3.701	4.678	–	–	–
	Lossless JPEG2000	1.524	2.048	3.053	4.120	4.946	5.865	6.589	7.685
	HEVC/H.265	1.373	2.051	2.958	3.952	5.024	6.034	6.702	8.897
	Neves & Pinho	1.105	1.793	2.695	3.653	4.659	5.512	6.053	7.335
Microzip	Average entropy	1.859	2.729	3.679	4.665	5.662	6.661	7.639	9.831
	bzip2	1.574	2.435	3.380	4.370	5.381	6.408	7.379	9.394
	JPEG-LS	1.448	2.149	3.037	4.013	5.011	6.028	7.005	8.974
	CALIC	1.383	2.176	2.977	3.915	4.904	–	–	–
	Lossless JPEG2000	1.825	2.161	3.178	4.275	5.171	6.212	7.597	9.157
	HEVC/H.265	1.609	2.403	3.339	4.343	5.447	6.638	7.893	11.179
	Neves & Pinho	1.243	1.957	2.864	3.868	4.856	5.859	6.830	8.667
Arizona	Average entropy	2.094	2.959	3.902	4.887	5.881	6.877	7.781	9.306
	bzip2	1.577	2.398	3.321	4.304	5.309	6.331	7.234	8.944
	JPEG-LS	1.491	2.270	3.139	4.102	5.093	6.125	7.005	8.646
	CALIC	1.464	2.250	3.061	4.003	4.980	–	–	–
	Lossless JPEG2000	1.742	2.216	3.273	4.351	5.241	6.274	7.424	8.795
	HEVC/H.265	1.470	2.280	3.229	4.236	5.338	6.532	7.664	10.592
	Neves & Pinho	1.201	1.976	2.874	3.878	4.870	5.867	6.766	8.275
IBB	Average entropy	3.168	3.906	4.651	5.386	6.095	6.756	7.340	8.503
	bzip2	3.048	3.832	4.649	5.448	6.206	6.927	7.590	9.081
	JPEG-LS	3.571	4.490	5.373	6.227	7.029	7.733	8.429	9.904
	CALIC	3.366	4.235	5.091	5.936	6.740	–	–	–
	Lossless JPEG2000	3.179	3.880	4.788	5.646	7.271	8.076	7.261	8.392
	HEVC/H.265	3.654	4.671	5.685	6.716	7.717	8.863	9.991	12.262
	Neves & Pinho	2.653	3.363	4.105	4.844	5.556	6.214	6.800	8.039
Corpora averages	Average entropy	2.196	3.007	3.906	4.841	5.778	6.680	7.430	9.010
	bzip2	1.817	2.610	3.519	4.473	5.432	6.362	7.148	9.052
	JPEG-LS	1.835	2.567	3.433	4.368	5.324	6.272	7.109	9.350
	CALIC	1.752	2.512	3.342	4.246	5.185	–	–	–
	Lossless JPEG2000	2.006	2.745	3.596	4.532	5.511	6.483	7.392	9.759
	HEVC/H.265	1.903	2.703	3.642	4.656	5.733	6.896	8.055	11.850
	Neves & Pinho	1.507	2.209	3.064	4.001	4.947	5.846	6.578	8.321



(b)

Fig. 5: Evaluation of the PSNR metric. (a) PSNR as a function of the average compression bitrate; (b) PSNR as a function of the $\mathrm{ARE}_{\mathrm{CRM}}$ metric.

compression with negligible effect on the DNA microarray analysis. In what follows, we compare the performance of RQ-based compression with other lossy compression approaches, using the distortion metrics developed in Section III.

As discussed in Section II-A, metrics based on the quadratic pixel-wise error like the PSNR (or MSE) are not adequate in this regard. Similarly, metrics based on the human visual system such as SSIM [32] and HDR-VDP 2 [33] may not be useful, since microarray images are analyzed by algorithms and not by human observers. Nevertheless, it is of interest to examine the results for these metrics yielded by the algorithms of interest. To this end, images from the IBB corpus were compressed with lossy JPEG2000[2] compression for target bitrates $R \in 3, 4, \ldots, 7$, near-lossless JPEG-LS compression for maximum absolute error $e \in \{4, 16, 32, 64\}$[3], lossy HEVC with quality parameter $QP \in \{1, 4, 5\}$ and the proposed RQ for $k \in \{1, 2, \ldots, 7\}$. Note that current rate-control algorithms for HEVC are not designed for intra-coding of a single picture [34]. Their accuracy in yielding a target bitrate is usually low, hence only the QP is used here. Each parameter choice for each algorithm yields a different compression treatment. When applied to the entire corpus, each such treatment yields an average PSNR and average compression performance (rate in bpp). Each such PSNR/rate pair is plotted in Fig. 5a. Each algorithm is represented by a different color in the figure. Each parameter choice is labeled next to its corresponding PSNR/rate pair. It can be seen that the proposed method yields lower PSNR results than lossy JPEG2000, near-lossless JPEG-LS and lossy HEVC. This is as expected since those algorithms consider the absolute error in each pixel. This is exactly the error targeted by traditional metrics such as MSE/PSNR. On the other hand, the proposed Relative Quantizer preserves the smaller pixel values more carefully while tolerating larger errors in pixels having large values.

To emphasize that PSNR is not a good predictor of the errors incurred in the CRM, all PSNR/$\mathrm{ARE}_{\mathrm{CRM}}$ pairs are depicted in Fig. 5b. Each point of this graph corresponds to one image of the IBB corpus, one algorithm and one parameter value. The color of each marker identifies the algorithm employed, and different parameter values are denoted with different marker shapes. It can be observed that the PSNR is not well correlated with the relative error introduced in the extracted CRM. Similar results are obtained for the MSE, SSIM and HDR-VDP-2 metrics. Therefore, only the $\mathrm{ARE}_{\mathrm{CRM}}$ and FWDOC metrics defined in Section III-A are employed for the results that follow.

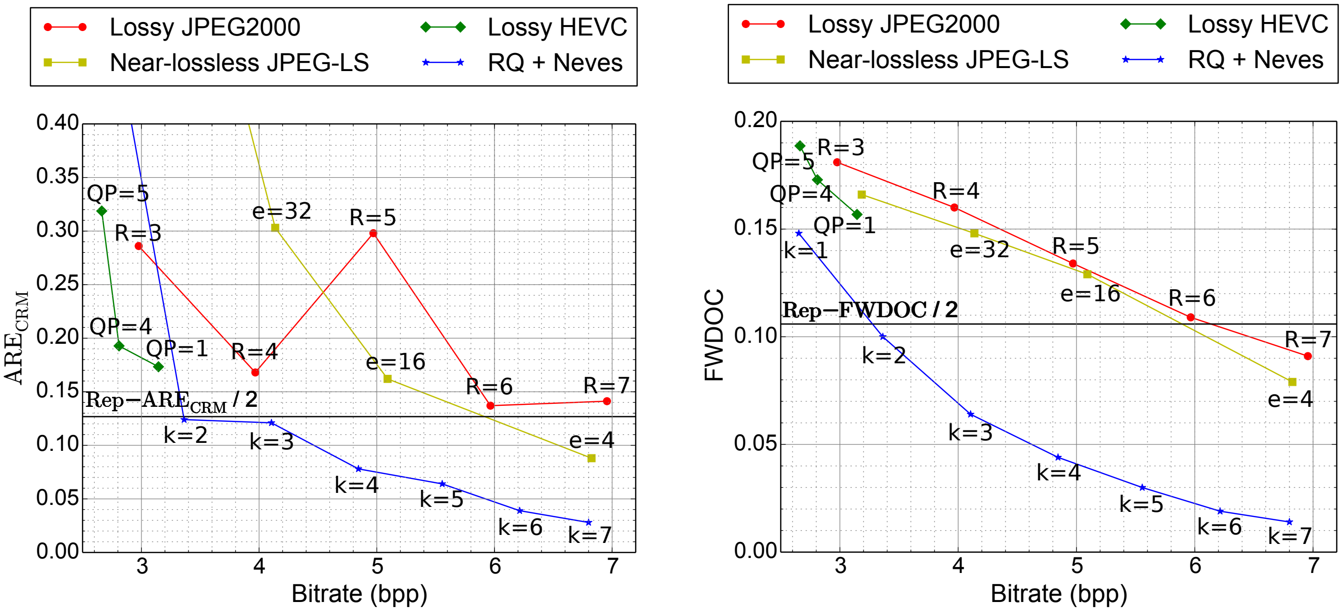
To the best of the authors' knowledge, all lossy algorithms published in 2004 or earlier [1], [2], [4] are based on either lossy JPEG2000 or near-lossless JPEG-LS. Two microarray-specific lossy compression methods [5], [6] have been published since then. The results reported in [5] are more than 5 dB worse than JPEG2000 in terms of PSNR. Using the authors implementation of the wavelet-fractal algorithm reported in [6], we obtained $\mathrm{ARE}_{\mathrm{CRM}}$ and FWDOC results of 1.030 and 0.286, respectively. The results of the proposed RQ method are significantly better for every value of k. In light of these results, only the standard lossy JPEG2000, lossy HEVC and near-lossless JPEG-LS algorithms are used to provide comparisons with the proposed RQ-based coder. In these comparisons, Neves and Pinho's algorithm is used for lossless compression of the RQ indices.

The resulting rate-distortion curves for the IBB corpus are shown in Fig. 6. Results for the lossy JPEG2000 algorithm have been obtained without applying the level offset and using 3 decomposition levels of the 9/7 irreversible DWT, the best choice for this corpus.

It can be observed that for $k > 1$, the proposed system consistently yields better results than lossy JPEG2000, near-lossless JPEG-LS and lossy HEVC for both the $\mathrm{ARE}_{\mathrm{CRM}}$ and FWDOC metrics. Notice that for $QP = 1$, lossy HEVC produces worse distortion results than the proposed RQ with $k = 2$, and that, as explained above, $QP = 1$ corresponds to the maximum quality for lossy HEVC and it is not possible to generate results for higher bitrates using this algorithm. At about only 3.4 bpp, the proposed algorithm produces less than Rep-$\mathrm{ARE}_{\mathrm{CRM}}$ / 2 and Rep-FWDOC / 2, i.e., half the

[2] Without applying the level offset and using 3 levels of the 9/7 irreversible DWT, the best choice for this corpus.

[3] Using parameters `-ls 0 -m` e.



(a)

(b)

Fig. 6: Distortion metrics versus bitrate: (a) average relative error in the CRM ($\mathrm{ARE_{CRM}}$); (b) fraction of spots incorrectly detected or classified (FWDOC). Half the average replicate CRM relative error ($\mathrm{Rep\text{-}ARE_{CRM}}/2$) and half the fraction of replicated spots wrongly detected or classified (Rep-FWDOC / 2) are shown as horizontal lines in (a) and (b), respectively.

acceptable experimental variability. This should be compared to an average of 8.039 bpp required to achieve strictly lossless compression of the original images without quantization (as can be seen in Table V).

The sudden increase in $\mathrm{ARE_{CRM}}$ for JPEG2000 in Fig. 6a for target bitrate $R = 5$ bpp is due to a small number of very large errors introduced into the CRM by JPEG2000 at that bit rate. In particular, 10 spots belonging to 3 different images have relative errors of more than 1000 after compression with JPEG2000. The sudden increase observed in Fig. 6a is due entirely to these 10 spots. The behavior of JPEG2000 can be explained by noting that when the denominator of Eq. (1) is small, even small changes in the numerator or the denominator can cause large errors in the CRM. JPEG2000 makes no attempt to account for this issue. On the other hand, this behavior is not present in the proposed RQ, since the error introduced in the CRM is strictly limited by construction.

V. Conclusions

DNA microarray images are usually stored so that they can be re-analyzed with future algorithms or in different laboratories. Due to the large amount of DNA microarray image information being currently generated, image compression is a useful tool to cope with the storage and transmission of these data. State-of-the-art lossless coding algorithms typically yield compression ratios of only 2:1 or less. Lossy coding methods can attain much higher compression ratios, however, some distortion is introduced in the decompressed images. Thus, it is necessary to assess the acceptability of this distortion in regards to subsequent image analysis.

In this paper, a Relative Quantizer (RQ)-based lossy compression method is proposed. The RQ is designed to limit two quantities that are crucial to the analysis process: the relative error of all pixels and the absolute error of low-intensity pixels. The distortion introduced by the proposed RQ results in errors in the analysis process that are smaller than those due to the experimental variability inherent to DNA microarrays. The proposed coding algorithm results in compression ratios exceeding 4.5:1 without introducing any additional analysis error. Furthermore, the k parameter of the RQ can be adjusted to trade off compression bitrate for analysis result precision. The rate-distortion results of the proposed coder significantly outperform those of state-of-the-art lossy coding algorithms.

Acknowledgments

The Arizona corpus was provided by Prof. D. Galbraith and Dr. M. Sweeney from the University of Arizona. The IBB corpus was provided by A. Barceló and A. Casamayor from the Genomics and Bioinformatics Service of the IBB at the UAB, whom the authors also thank for providing expertise on DNA microarrays.

References

[1] R. Jörnsten, W. Wang, B. Yu, and K. Ramchandran, “Microarray image compression: SLOCO and the effect of information loss,” *Signal Processing*, vol. 83, no. 4, pp. 859–869, April 2003.

[2] J. Hua, Z. Liu, Z. Xiong, Q. Wu, and K. Castleman, “Microarray BASICA: Background Adjustment, Segmentation, Image Compression and Analysis of microarray images,” *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 1, pp. 92–107, January 2004.

[3] Q. Xu, J. Hua, Z. Xiong, M. L. Bittner, and E. R. Dougherty, “The effect of microarray image compression on expression-based classification,” *Signal Image and Video Processing*, vol. 3, no. 1, pp. 53–61, February 2009.

[4] S. Lonardi and Y. Luo, "Gridding and compression of microarray images," in *Proceedings of the IEEE Computational Systems Bioinformatics Conference*, 2004, pp. 122–130.

[5] T. J. Peters, R. Smolikova-Wachowiak, and M. P. Wachowiak, "Microarray image compression using a variation of singular value decomposition," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1-16, 2007, pp. 1176–1179.

[6] M. R. N. Avanaki, A. Aber, and R. Ebrahimpour, "Compression of cDNA microarray images based on pure-fractal and wavelet-fractal techniques," *ICGST International Journal on Graphics, Vision and Image Processing, GVIP*, vol. 11, pp. 43–52, March 2011.

[7] K. Blekas, N. Galatsanos, A. Likas, and I. Lagaris, "Mixture Model Analysis of DNA Microarray Images," *IEEE Trans. Med. Imag.*, vol. 24, no. 7, pp. 901–909, Jul. 2005.

[8] J. Ho and W.-L. Hwang, "Automatic Microarray Spot Segmentation Using a Snake-Fisher Model," *IEEE Trans. Med. Imag.*, vol. 27, no. 6, pp. 847–857, Jun. 2008.

[9] E. Zacharia and D. Maroulis, "An Original Genetic Approach to the Fully Automatic Gridding of Microarray Images," *IEEE Trans. Med. Imag.*, vol. 27, no. 6, pp. 805–812, Jun. 2008.

[10] L. Rueda and I. Rezaeian, "A fully automatic gridding method for cDNA microarray images," *BMC Bioinformatics*, vol. 12, no. 1, p. 113, 2011.

[11] G.-F. Shao, F. Yang, Q. Zhang, Q.-F. Zhou, and L.-K. Luo, "Using the maximum between-class variance for automatic gridding of cDNA microarray images," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 10, no. 1, pp. 181–192, Jan 2013.

[12] L. Rueda, Ed., *Microarray image and data analysis: theory and practice*. CRC Press, 2014.

[13] L. Srinivasan, Y. Rakvongthai, and S. Oraintara, "Microarray image denoising using complex Gaussian scale mixtures of complex wavelets," *IEEE J. Biomed. Health. Inform.*, vol. 18, no. 4, pp. 1423–1430, 2014.

[14] M. Hernández-Cabronero, M. W. Marcellin, and J. Serra-Sagristà, "Compression of DNA microarray images," in *Microarray image and data analysis: theory and practice*. CRC Press, 2014, pp. 193–222.

[15] Y. Luo and S. Lonardi, "Storage and transmission of microarray images," *Drug Discovery Today*, vol. 10, no. 23-24, pp. 1689 – 1695, 2005.

[16] Y. H. Yang, M. J. Buckley, and T. P. Speed, "Analysis of cDNA microarray images," *Briefings in Bioinformatics*, vol. 2, no. 4, pp. 341–349, Jan. 2001.

[17] 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, no. 5235, pp. 467–470, 1995.

[18] J. Sun and V. Goyal, "Scalar quantization for relative error," in *Proceedings of the IEEE International Data Compression Conference, DCC*, March 2011, pp. 293–302.

[19] Molecular Devices, "GenePix Pro [Online]. Available http://moleculardevices.com/."

[20] Y. Zhang, R. Parthe, and D. Adjeroh, "Lossless compression of DNA microarray images," in *Proceedings of the IEEE Computational Systems Bioinformatics Conference*, August 2005, pp. 128 – 132.

[21] Speed Berkeley Research Group, "ApoA1 corpus [Online]. Downloaded from stat.berkeley.edu/users/terry/zarray/Html/apodata.html."

[22] SIB Computational Genomic Group, "ISREC corpus [Online]. Downloaded from: http://www.isrec.isb-sib.ch/DEA/module8/P5_chip_image/images/."

[23] Stanford Microarray Database, "Stanford corpus [Online]. Downloaded from: ftp://smd-ftp.stanford.edu/pub/smd/transfers/Jenny."

[24] David Galbraith Laboratory, "Arizona corpus [Online]. Available: http://deic.uab.es/~mhernandez/materials."

[25] IBB Genomics Service, "IBB corpus [Online]. Available: http://deic.uab.es/~mhernandez/materials."

[26] A. J. R. Neves and A. J. Pinho, "Lossless compression of microarray images using image-dependent finite-context models," *IEEE Trans. Med. Imag.*, vol. 28, no. 2, pp. 194–201, February 2009.

[27] M. J. Weinberger, G. Seroussi, and G. Sapiro, "The LOCO-I lossless image compression algorithm: principles and standardization into JPEG-LS," *IEEE Trans. Image Process.*, vol. 9, no. 8, pp. 1309–1324, 2000.

[28] X. Wu and N. Memon, "Context-based, adaptive, lossless image coding," *IEEE Trans. Commun.*, vol. 45, no. 4, pp. 437–444, Apr 1997.

[29] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image compression fundamentals, standards and practice*. Kluwer Academic Publishers, Boston, 2002.

[30] "High Efficiency Video Coding (HEVC) reference software (HM) [Online]. Available: http://hevc.hhi.fraunhofer.de."

[31] "ITU-T H.264 Recommendation [Online]. Available: http://www.itu.int/rec/T-REC-H.264."

[32] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, April 2004.

[33] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions," in *ACM Transactions on Graphics (TOG)*, vol. 30, no. 4. ACM, 2011, p. 40.

[34] V. Sanchez, F. Aulí-Llinàs, R. Vanam, and J. Bartrina-rapesta, "Rate Control for Lossless Region of Interest Coding in HEVC Intra-Coding with Applications to Digital Pathology Images," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Dec. 2015, In Press.