Universitat Autònoma de Barcelona

Departament d'Enginyeria de la Informació i de les Comunicacions

DNA Microarray Image Compression

Submitted to Universitat Autònoma de Barcelona
in partial fulfillment of the requirements for the
degree of Doctor of Philosophy in Computer Science

by Miguel Hernández-Cabronero

Bellaterra, June 2015

Supervisor:
Dr. Joan Serra Sagristà

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Bellaterra, June 2015

Dr. Joan Serra Sagristà (supervisor)

Committee:

Dr. Manuel Perez Malumbres

Dr. Joan Bartrina Rapesta

Dr. António José Ribeiro Neves

Dr. Javier Ruiz Hidalgo (substitute)

Dr. Carles Garrigues Olivella (substitute)

Abstract

Medical imaging methods –*e.g.*, ultrasound, computer tomography (CT) or X-rays– are crucial tools for the diagnosis and study of many diseases. DNA microarrays are one of such methods, commonly employed in biological and biomedical laboratories around the world. In DNA microarray experiments, two grayscale images are produced as an intermediate step. These DNA microarray images are then analyzed to obtain the genetic data of interest. Since these analysis algorithms are in constant development, it is convenient to save these images for future, more accurate re-analysis. Thus, image compression emerges as a particularly useful tool to alleviate the associated storage and transmission costs. This dissertation aims at improving the state of the art of the compression of DNA microarray images.

A thorough investigation of the characteristics of DNA microarray images has been performed as a part of this work. DNA microarray images exhibit much larger dynamic ranges, very different pixel distributions and use a smaller fraction of all possible intensities, as compared to natural images. Hence, algorithms not adapted to DNA microarray images typically attain only mediocre lossless compression results. By analyzing the first-order and conditional entropy present in these images, it is possible to determine approximate limits to their lossless compressibility. Even though context-based coding and segmentation provide modest improvements over generic-purpose algorithms, conceptual breakthroughs in data coding are arguably required to achieve compression ratios exceeding 2:1 for most images.

Prior to the start of this thesis, several lossless coding algorithms that have performance results close to the aforementioned limit were published. However, none of them is compliant with existing image compression standards. Hence, the availability of decoders in future platforms –a requisite for future re-analysis– is not guaranteed. Moreover, the adhesion to standards is usually a requisite in clinical scenarios. To address these problems, a fast reversible transform compatible with the JPEG2000 standard –the Histogram Swap Transform (HST)– is proposed. The HST improves the average compression performance of JPEG2000 for all tested image corpora, with gains ranging from 1.97% to 15.53%. Furthermore, this transform can be applied with only negligible time complexity overhead. With the HST, JPEG2000 becomes arguably the most competitive alternatives to microarray-specific, non-standard compressors. The similarities among sets of microarray images have also been studied as a means to improve the compression performance of standard and microarray-specific algorithms. An optimal grouping of the images which maximizes the inter-group

correlation is described. Average correlations between 0.75 and 0.92 are observed for the tested corpora. Thorough experimental results suggest that spectral decorrelation transforms can improve some lossless coding results by up to 0.6 bpp, although no single transform is effective for all copora.

Lossy coding algorithms can yield almost arbitrary compression ratios at the cost of modifying the images and, thus, of distorting subsequent analysis processes. If the introduced distortion is smaller than the inherent experimental variability, it is usually considered acceptable. Hence, the use of lossy compression is justified on the assumption that the analysis distortion is assessed. In this work, a distortion metric for DNA microarray images is proposed to predict the extent of this distortion without needing a complete re-analysis of the modified images. Experimental results suggest that this metric is able to tell apart image changes that affect subsequent analysis from image modifications that do not. Although some lossy coding algorithms were previously described for this type of images, none of them is specifically designed to minimize the impact on subsequent analysis for a given target bitrate. In this dissertation, a lossy coder –the Relative Quantizer (RQ) coder– that improves upon the rate-distortion results of previously published methods is proposed. Experiments suggest that compression ratios exceeding 4.5:1 can be achieved while introducing distortions smaller than half the inherent experimental variability. Furthermore, a lossy-to-lossless extension of this coder –the Progressive RQ (PRQ) coder– is also described. With the PRQ, images can be compressed once and then reconstructed at different quality levels, including lossless reconstruction. In addition, the competitive rate-distortion results of the RQ and PRQ coders can be obtained with computational complexity slightly smaller than that of the best-performing lossless coder of DNA microarray images.

Contents

Chapter 1

Introduction

1.1 DNA Microarrays

In recent years, medical imaging has become an indispensable tool for diagnosis and disease understanding. Great efforts are being made to improve the available imaging modalities –*e.g.*, ultrasound, magnetic resonance, computer tomography (CT), X-rays or visible light– and explore new ones. In spite of the constant struggle, important challenges remain open in all modalities and stages of the imaging process, from registration to analysis, also including storage and transmission of the images.

DNA Microarrays are an important imaging modality, widely employed in biological and medical research. In a single DNA microarray experiment, it is possible to monitor the behavior of thousands of genes simultaneously, and even the whole genome of an organism [1, 2]. Any species for which its genomic sequence is known is eligible for being subject to a DNA microarray experiment. Human (*Homo sapiens*), mouse (*Mus musculus*) and yeast (*Saccharomyces cerevisiae*) are among the most common such species. The information obtained via these experiments can be used to study the function and regulation mechanisms of virtually any gene. DNA microarrays can also be used to analyze the physiological reaction to a given drug, pathogen or environmental condition, or to compare species and subspecies in evolutionary biology. For this reason, DNA microarrays have been regularly employed in the research against Cancer [3, 4], HIV [5] or Malaria [6], among many other topics.

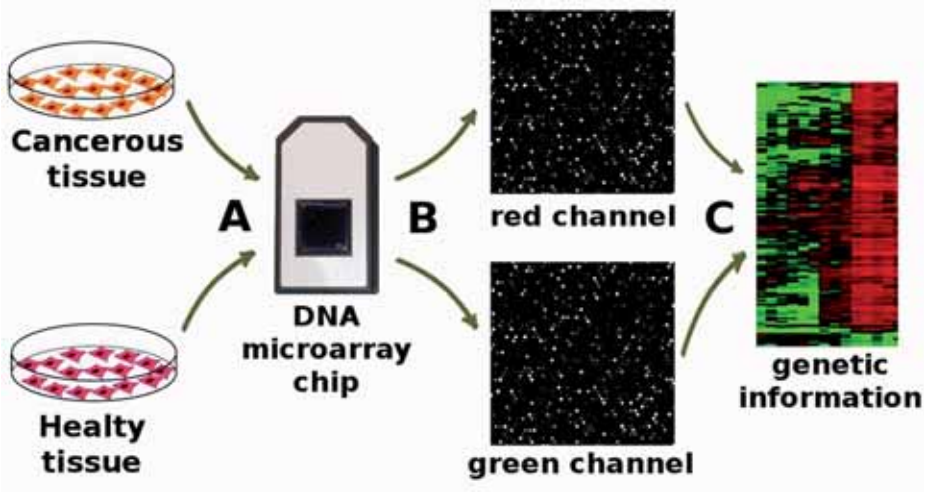
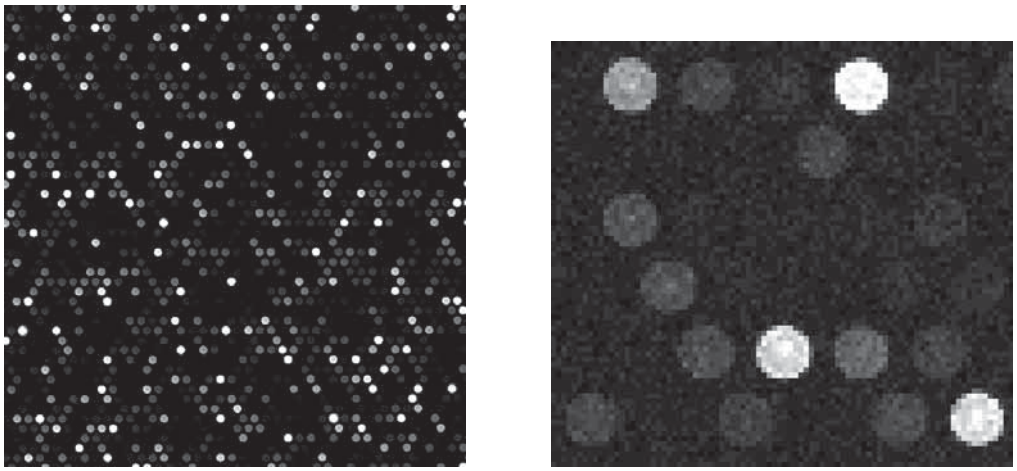


Figure 1.1: Diagram of a DNA microarray image experiment.

The main steps of a typical DNA microarray experiment are depicted in Fig. 1.1. Two biological samples, for instance coming from healthy and tumoral tissue, are put on a DNA microarray chip (step A in the Figure). The chip is then scanned to produce two grayscale images, the so-called *green* and *red* channels (step B in the Figure). The name green and red for these images is due to the color of the fluorescent markers that are applied to the biological samples prior to putting them on the chip. Each of the images is obtained by laser stimulation of one of the markers and contains information about one of the biological samples. Finally, the pair of images is then analyzed jointly in order to generate *genetic expression data* (step C in the Figure). Once these data are available, several statistical tools including normalization and classification are applied on them depending on the scope of that particular DNA microarray experiment.

The analysis of DNA microarray images is a very active research topic. Different parts of the analysis process have been explored –and improved– in recent publications [7, 8, 9, 10, 11, 2, 12]. A review of the state of the art on the analysis of microarray images can be found in [2]. As new, more accurate analysis techniques are developed, it will be interesting to reanalyze the microarray images to obtain more precise genetic expression data that can lead to more significant research results. However, repeating the whole DNA microarray experiment is most usually not



(a) (b)

Figure 1.2: Crops of a DNA microarray image with gamma levels adjusted for visualization: (a) 800×800 crop at original resolution; (b) 100×90 crop at a $8\times$ magnification.

a viable option because the needed biological samples may not be available anymore. The biological samples are much less likely to be accessible if the reanalysis is to be performed a long time after the first experiment or in a different laboratory. Therefore, storing the DNA microarray images is paramount to guarantee the applicability of future image analysis techniques. A single experiment using a last-generation DNA microarray platform can be carried out in about half an hour, generating over 250 MB of uncoded image data. At full capacity, a single DNA microarray scanner can produce over 5 GB of uncoded information everyday. As a result, large amounts of data are being produced in laboratories around the world and a necessity for efficient storage and transmission of DNA microarray images arises. Data compression –in particular, image compression– is a natural approach to this problem. If microarray images are represented in a more compact way, the costs associated to the management of these data decrease and a fast sharing of the images among geographically distant laboratories becomes feasible. The main goal of this thesis is to provide a significant contribution to the state of the art in the compression of DNA microarray images.

1.2 DNA Microarray Image Compression

DNA microarray images possess several properties that render their compression a challenging task. While natural images typically require 8 bits per sample, 16 bits are needed for microarray images and microarray pixels can take any of the 2^{16} possible values. Furthermore, microarray images present irregular round regions of varying brightness (the *spots*) on a dark background, as shown in Fig. 1.2. These abrupt changes are very difficult to code, as compared to the smooth regions that natural images typically present. In addition, between 6 and 9 of the least significant bit-planes exhibit random-like distributions with entropies close to the maximum of 1 bit per pixel (bpp). Altogether, these properties –further detailed in Chapter 2– make microarray images very different from natural images. Hence, the direct application of generic image compressors yields only poor results and microarray-specific techniques need be developed.

Prior to the development of this thesis, several methods had been proposed for both lossless [13, 14, 15, 16, 17, 18, 19, 20, 21] and lossy [14, 15, 22, 23] compression of DNA microarray images. A detailed review of all these methods is available in [24], and a brief description of their most important features is provided later.

Lossless compression guarantees perfect data fidelity. However, none of the publications previous to this thesis consistently attained compression ratios better than 2:1. Furthermore, none of the proposed methods is compliant with existing image compression standards, which can limit the availability of compatible decoders when new analysis techniques are developed in the future. Other desirable features such as quality scalability –the possibility of recovering a low resolution version of the image by decoding only part of the compressed information– and spatial scalability –the possibility of recovering parts of the image at full resolution– are also absent from previously published methods.

On the other hand, lossy compression allows almost arbitrary compression ratios at the cost of introducing changes in the original images. In turn, these changes can distort the results of analysis algorithms applied to microarray images, rendering the compressed images unusable. Nevertheless, if these changes are sufficiently small,

they can be considered acceptable [14, 16, 25]. However, previous publications dealing with lossy compression do not generally assess the acceptability of their impact in the analysis results in a rigorous way.

This work aims at finding solutions to the aforementioned issues regarding both lossless and lossy compression of DNA microarray images.

1.3 Contributions and Thesis Organization

This thesis consists of several contributions to the state of the art of the compression of DNA microarray images. The structure of the remainder of this document is detailed next.

Chapter 2 contains a thorough description of the DNA microarray images employed in this work. This description allows for a deeper knowledge of these images and sets some theoretical limits to their compressibility.

Chapter 3 describes the state of the art on the lossless compression of microarray images and the contributions on this field accomplished in this thesis. These contributions were originally presented in the following publications:

1. [26] **Miguel Hernández-Cabronero**, Juan Muñoz-Gómez, Ian Blanes, Joan Serra-Sagristà, Michael W. Marcellin, "DNA microarray image coding," In proceedings of the IEEE Data Compression Conference, DCC, pp 32-41, 2012.

2. [27] **Miguel Hernández-Cabronero**, Francesc Aulí-Llinás, Joan Bartrina-Rapesta, Ian Blanes, Leandro Jiménez-Rodríguez, Michael W. Marcellin, Juan Muñoz-Gómez, Victor Sanchez, Joan Serra-Sagristà, Zhongwei Xu, "Multicomponent compression of DNA microarray images," In Proceedings of the CEDI Workshop on Multimedia Data Coding and Transmission, WMDCT, 2012.

3. [28] **Miguel Hernández-Cabronero**, Victor Sanchez, Michael W. Marcellin, Joan Serra-Sagristà, "Compression of DNA Microarray Images", In Book "Microarray Image and Data Analysis: Theory and Practice", CRC Press, ch. 8, pp 193-225, 2014.

The first publication proposes a lossless point transform that is able to reduce the performance gap between the standard, DICOM-compatible JPEG2000 [29] and the best-performing lossless microarray-specific algorithms. The second contribution, first presented in a Workshop in 2012, and then included as part of a book chapter in 2014, is an exploration of the effect of multicomponent decorrelation transform on the lossless compression of microarray images.

Chapter 4 addresses the contributions of this thesis to the lossy compression of these images, addressed in the following works:

1. [30] **Miguel Hernández-Cabronero**, Victor Sanchez, Michael W. Marcellin, Joan Serra-Sagristà, "A distortion metric for the lossy compression of DNA microarray images," In proceedings of the IEEE International Data Compression Conference, DCC, pp 171-180, 2013.

2. [31] **Miguel Hernández-Cabronero**, Ian Blanes, Armando J. Pinho, Michael W. Marcellin, Joan Serra-Sagristà, "Analysis-Driven Lossy Compression of DNA Microarray Images," Submitted to IEEE Transactions on Medical Imaging.

3. [32] **Miguel Hernández-Cabronero**, Ian Blanes, Armando J. Pinho, Michael W. Marcellin, Joan Serra-Sagristà, "Progressive Lossy-to-Lossless Compression of DNA Microarray Images," Submitted to IEEE Signal Processing Letters.

The first contribution proposes an image distortion metric able to predict the impact on the analysis process based only on an original and a modified pair of microarray images. The second contribution –currently under review– is a lossy compression algorithm, which is based on an original quantization scheme. It is designed to limit its impact on the image analysis results and the introduced distortion is assessed by means of realistic analysis experiments. The third article describes a progressive lossy-to-lossless extension to the aforementioned lossy compression algorithm.

Finally, Chapter 5.2 draws global conclusions for this thesis and provides some insight on the future of the field of DNA microarray image compression.

A brief note on the quality and relevance of the publications is provided now. The first two publications on lossless and lossy compression appear in the proceedings of arguably the most important conference on data compression, *i.e.*, the Data

Compression Conference (DCC). This annual conference is ranked as A* (the highest possible) in the Computer Research and Education (CORE) Conference Ranking[1]. The second publication on lossless compression was included as a chapter in a book edited by Luis Rueda, who has authored over 30 works on DNA microarrays since the apparition of this technique, some of them in top-level Journals such as BMC Bioinformatics (first quartile of the "Mathematical & Computational Biology" category in the standard ISI Web of Science[2]). The second publication on lossy compression is currently submitted to the IEEE Transactions on Medical Imaging, which had an impact factor of 3.799 and ranked in the first quartile of five categories –including "Computer Science: Interdisciplinary Applications"– in 2013 (the last year for which data are currently available). Finally, the third publication on lossy compression is currently under review at the IEEE Signal Processing Letters, which had an impact factor of 1.639 and ranked in the second quartile of the "Engineering, Electrical & Electronic" category of the ranking.

[1] http://www.core.edu.au/coreportal
[2] https://webofknowledge.com

Chapter 2

DNA Microarray Images

When a image compression algorithm is designed, several assumptions about the image properties are made. For instance, most efficient algorithms rely on the similarity among neighboring pixels and the pixel intensity distribution. In many coding schemes, it is assumed that the input data are natural images such as pictures taken by a digital camera. However, when these schemes are applied to other types of images, their efficiency is hindered sensibly. DNA microarray images differ from natural images in many aspects. Therefore, deep knowledge of their characteristics is required to design efficient methods for their compression. In this chapter, a thorough description of the DNA microarray images employed throughout this thesis is provided.

2.1 Image Corpora

Several image corpora have been employed in literature for the benchmarking of DNA microarray image compression algorithms. To the best of our knowledge, all related works published prior to the beginning of this thesis employed one or more of the following corpora: Yeast [33], ApoA1 [34], ISREC [35], or MicroZip [36]. During the development of this thesis, four additional corpora –representative of different or more modern scanners– were also considered: Stanford [37], Omnibus [38], Arizona [39] and IBB [40]. The properties of all these corpora are discussed next. In what follows,

Table 2.1: Image corpora employed throughout this work.

Property	**Yeast**	**ApoA1**	**ISREC**	**Stanford**
Year	1998	2001	2001	2001
Images	109	32	14	20
Size	1024×1024	1044×1041	1000×1000	$> 2000\times2000$
Spot count	$\sim 9 \cdot 10^3$	$\sim 6 \cdot 10^3$	$\sim 2 \cdot 10^2$	$\sim 4 \cdot 10^3$
Spot layout	square	square	square	square
Avg. intensities	5.39%	39.51%	33.34%	28.83%
Property	**MicroZip**	**Omnibus**	**Arizona**	**IBB**
Year	2004	2006	2011	2013
Images	3	25	6	44
Size	$> 1800\times1900$	12200×4320	4400×13800	2019×6235
Spot count	$\sim 9 \cdot 10^3$	$\sim 2 \cdot 10^5$	$\sim 2 \cdot 10^5$	$\sim 1.4 \cdot 10^4$
Spot layout	square	hexagonal	hexagonal	square
Avg. intensities	37.71%	97.64%	82.82%	54.07%

they are employed for benchmarking the different compression techniques proposed in this work.

A summary of some of the most important properties of the aforementioned image corpora is provided in Table 2.1. The registration year, the number of images and the image size of each set in pixels are shown in the corresponding rows of the table. One main difference between regular images and DNA microarray images is their size. Even though early image scanners produced relatively small images, state-of-the-art scanners generate images significantly larger than most digital cameras.

As described in Chapter 1, DNA microarray images typically exhibit irregular round regions of varying intensity –known as spots– over a dark background. These spots are usually packed following a rectangular (matrix-like) grid or an hexagonal (bee-hive-like) grid. The approximate number of spots and the spot layout is provided in the *Spot count* and *Spot layout* rows, respectively. The abrupt intensity changes induced by the spots are not commonly found in natural images and are difficult to code and predict.

DNA microarray image pixels require 16 bits to be stored, whereas natural images

typically employ 8 bpp for each color component. Therefore, microarray pixels can take 65536 different values, *i.e.*, 256 times more than natural images. Moreover, DNA microarray image pixel distributions are strongly biased towards low intensities, as opposed to the more uniform distributions of natural images. Histograms of example images from all corpora are shown in Fig. 2.3. Due to this distribution, not all possible intensities are employed in each image. The percentage of intensities used in each image has been calculated. The average usage fraction across each set is shown in the *Avg. intensities* row of the Table. As can be observed, a significant fraction of the possible intensities remains unused in each image, except for the Omnibus set. Note that any given intensity is typically present in several images of the same corpus, although not necessarily in a given image. The large amount of possible intensities along with the important fraction is in stark contrast with natural images, where all intensities are typically employed.

2.2 Entropy

The entropy of a data source provides knowledge about the amount of information present in the data. In particular, it is usually considered as the optimal bitrate required to represent the data with a general-purpose coder. Hence, entropy is a very relevant aspect in the field of data compression.

When dealing with images, pixels can be considered as the output of a discrete random variable X with support $\mathrm{supp}(X) = \{0, \ldots, N-1\}$. The first-order entropy of an image is defined as

$$H(X) = -\sum_{x \in \mathrm{supp}(X)} p(x) \log_2 p(x), \tag{2.1}$$

where $p(x)$ is the probability of a pixel having value x. The average first-order entropy of all aforementioned corpora is provided in Table 2.2. It can be observed that all image sets except for the Yeast set have entropies close to or larger than 8 bpp. Since uncoded images require 16 bpp, the use of image compression techniques for the storage and transmission of DNA microarray images is justified.

Table 2.2: Average entropy results in bpp before segmentation (first-order entropy $H(X)$ and conditional entropy $H(X|C)$) and after segmentation (first-order entropy $H_\theta(I)$ and conditional entropy $H_\theta(I|C)$). The optimal threshold θ_{best}, used in $H(I)$ and $H(I|C)$ for each corpus is also provided.

Corpus	**No segmentation**		**With segmentation**		
	$\boldsymbol{H(X)}$	$\boldsymbol{H(X\|C)}$	$\boldsymbol{\theta}_{\text{best}}$	$\boldsymbol{H(I)}$	$\boldsymbol{H(I\|C)}$
Yeast	6.63	5.68	2^9	5.86	5.28
ApoA1	11.03	10.38	2^{10}	10.42	9.97
ISREC	10.44	10.09	2^8	9.70	9.42
Stanford	8.29	7.46	2^6	7.70	7.11
MicroZip	9.83	9.20	2^8	9.28	8.85
Omnibus	7.87	6.86	2^6	6.91	6.32
Arizona	9.31	8.45	2^8	8.58	8.11
IBB	8.50	7.97	2^7	7.54	7.19

Many compression algorithms code the different bitplanes sequentially. Hence, it is also interesting to measure the entropy of each bitplane. In this case, $\text{supp}(X) = \{0, 1\}$ and $H(X)$ is contained in $[0, 1]$ The entropy of the bitplanes of sample images of all corpora is provided in Fig. 2.2. It can be observed that between 6 and 9 of the least significant bitplanes exhibit entropies close to the maximum of 1 bpp. Exceptionally, the three least significant bitplanes of the Yeast corpus are almost constant (entropy close to 0 bpp). This is due to the properties of the scanner employed to extract this corpus only, and should not be taken as a general property of DNA microarray images. These observations suggests that large amounts of uniform-like noise are present in these bitplanes. As it is well known, the lossless coding of data with this distribution is a very hard task. Thus, large lossless compression ratios should not be expected unless a significant breakthrough in lossless compression technology is made.

When calculating the first-order entropy, each pixel value is considered independently from other neighboring pixels. In real compression algorithms, however, information about nearby pixels –the conditioning event or *context*– is usually employed. Therefore, conditional entropy can be a more accurate prediction of the maximum compressibility of an image. In this scenario, the context can be modeled as the

output of a random variable C, and the conditional entropy is defined as

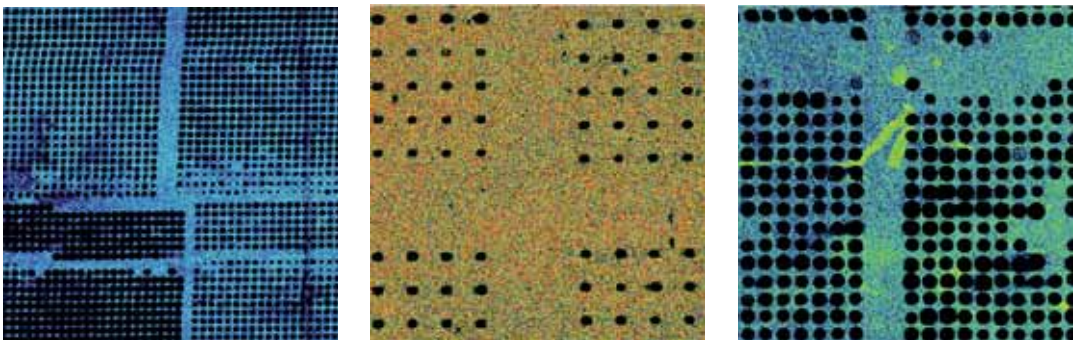
$$H(X|C) = -\sum_{\substack{x\in\text{supp}(X)\\ c\in\text{supp}(C)}} p(x,c)\log_2\frac{p(x)}{p(x,c)}, \tag{2.2}$$

where $p(x,c)$ is the probability of a pixel having value x with a context –*i.e.*, nearby pixels– with value c. For this work, the conditional entropy the context has been defined as $\lceil \log_2 \mu_{\text{neighbors}} \rceil$, where $\mu_{\text{neighbors}}$ is the average intensity of the 8 nearest neighbors. In the case of edge and corner pixels, only 5 and 4 neighbors are considered in $\mu_{\text{neighbors}}$. Therefore, the context can be modeled with a scalar random variable C such that $\text{supp}(C) = \{0,\ldots,15\}$. The average conditional entropy for all corpora is shown in Table 2.2. It can be seen that the conditional entropy is consistently smaller than the first-order entropy for all sets. In particular, reductions of up to 1 bpp can be observed. Hence, the use of context-based approaches for compression is justified. Note that other context definitions with different impact on the entropy results are possible. For instance, a larger or a smaller number of neighboring pixels can be employed. According to our experiments, all tested alternatives produce very similar entropy results. Thus, Table 2.2 shows results only for the aforementioned context definition.

As detailed later in Chapter 3, many successful lossless DNA microarray image compression methods segment the images into spots and background and then compress each part separately. Since spots and background exhibit different statistical properties, better compression performance is usually attained. An estimation of the compressibility of an image using this approach can be obtained by calculating independently the entropy of each part of the image, and then computing the weighted average of the entropies. If an image I is segmented into two sets of pixels modeled by the output of the random variables X and Y, then the entropy after segmentation is be defined as

$$H(I) = \frac{|X|}{|X|+|Y|}H(X) + \frac{|Y|}{|X|+|Y|}H(Y), \tag{2.3}$$

where $H(X)$ and $H(Y)$ are, respectively, the first-order entropy of X and Y, as



(a) (b) (c)

Figure 2.1: Sample images segmented with $\theta = 512$. Pixel values smaller than θ are represented in false color. Pixels larger than θ are shown in black. (a) *y744n89_ch1* from the Yeast set; (b) *Def667cy3* from the ISREC set; (c) *2001-01-18_0010* from the Stanford set;

defined in (2.1) and $|X|$ and $|Y|$ are the number of pixels in each set. Analogously, the conditional entropy after segmentation is defined as

$$H(I|C) = \frac{|X|}{|X|+|Y|}H(X|C) + \frac{|Y|}{|X|+|Y|}H(Y|C), \tag{2.4}$$

In this work, images have been segmented using a hard-thresholding algorithm, as suggested in [18]. The two segmented sets X and Y are defined as the pixels smaller than and greater or equal than θ, respectively. As can be observed in Fig. 2.1, this method yields acceptable segmentation results. The average first-order and conditional entropy after segmentation, along with the best choice of $\theta \in \{2^1, \ldots, 2^{16}\}$, are also provided in Table 2.2. As obvious from the table, both the first-order and the conditional entropy are smaller after segmenting the images. Notwithstanding, entropy reductions are always smaller than 0.8 bpp. Thus, it is not very likely to obtain great compression performance improvements due solely to segmentation.

In light of all gathered results, most corpora exhibit average entropies larger than 8 bits per pixel, even when segmentation and context information is considered. Therefore, it will be very difficult to attain lossless compression ratios exceeding 2:1 for most images. This is consistent with the compression limit predicted by Jörnsten [14] and with all existing experimental evidence, as described in Chapter 3.

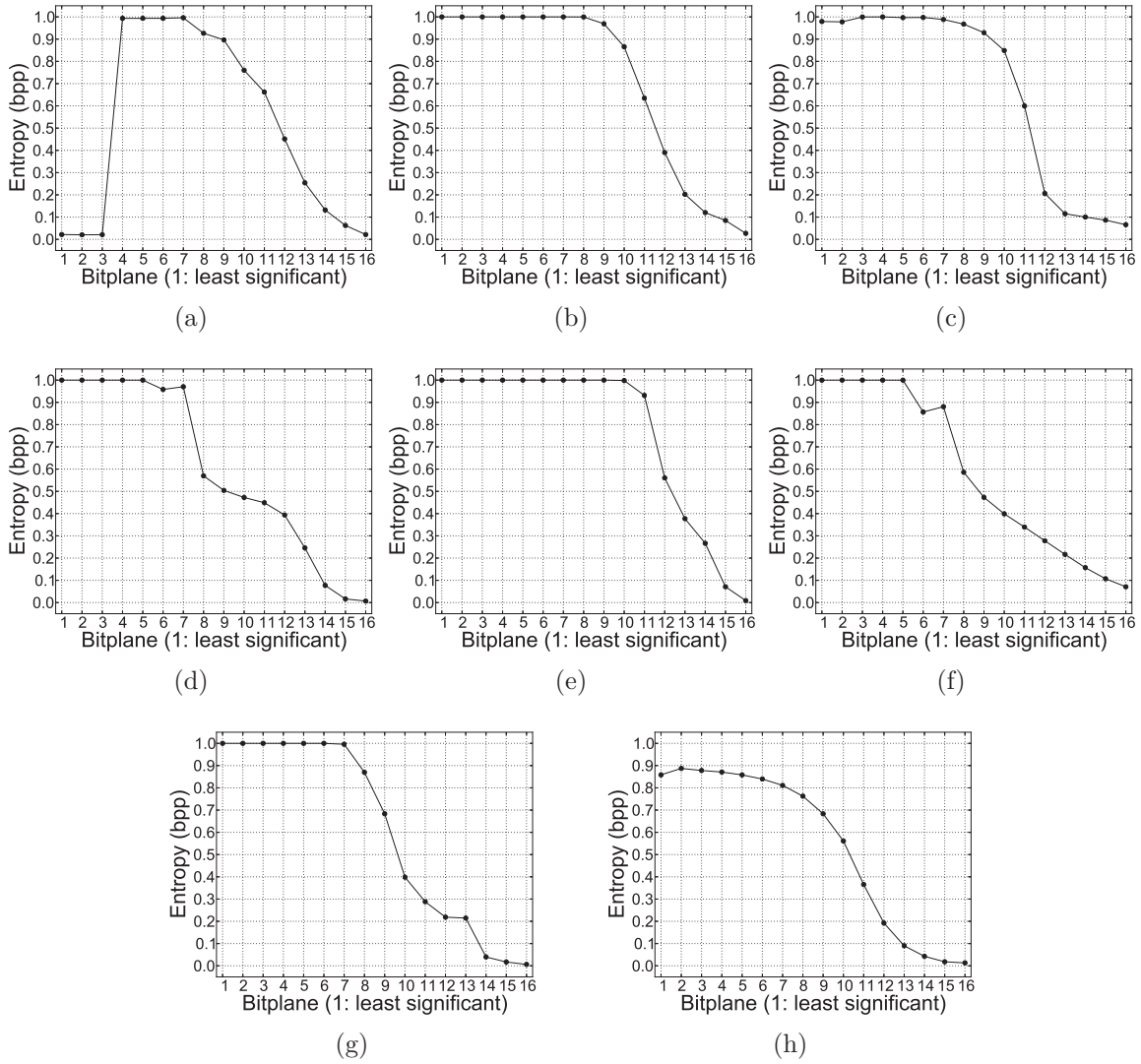


Figure 2.2: Bitplane entropy in bpp for sample DNA microarray images of all corpora. (a) *y744n101_ch1* from the Yeast set; (b) *1230ko1G* from the ApoA1 set; (c) *Def661Cy3* from the ISREC set; (d) *TB3_95_llama_005* from the Stanford set; (e) *array1* from the MicroZip set; (f) *GSM346097* from the Omnibus set; (g) *slide_1-red* from the Arizona set; (h) *134044018_Cy3* from the IBB set.

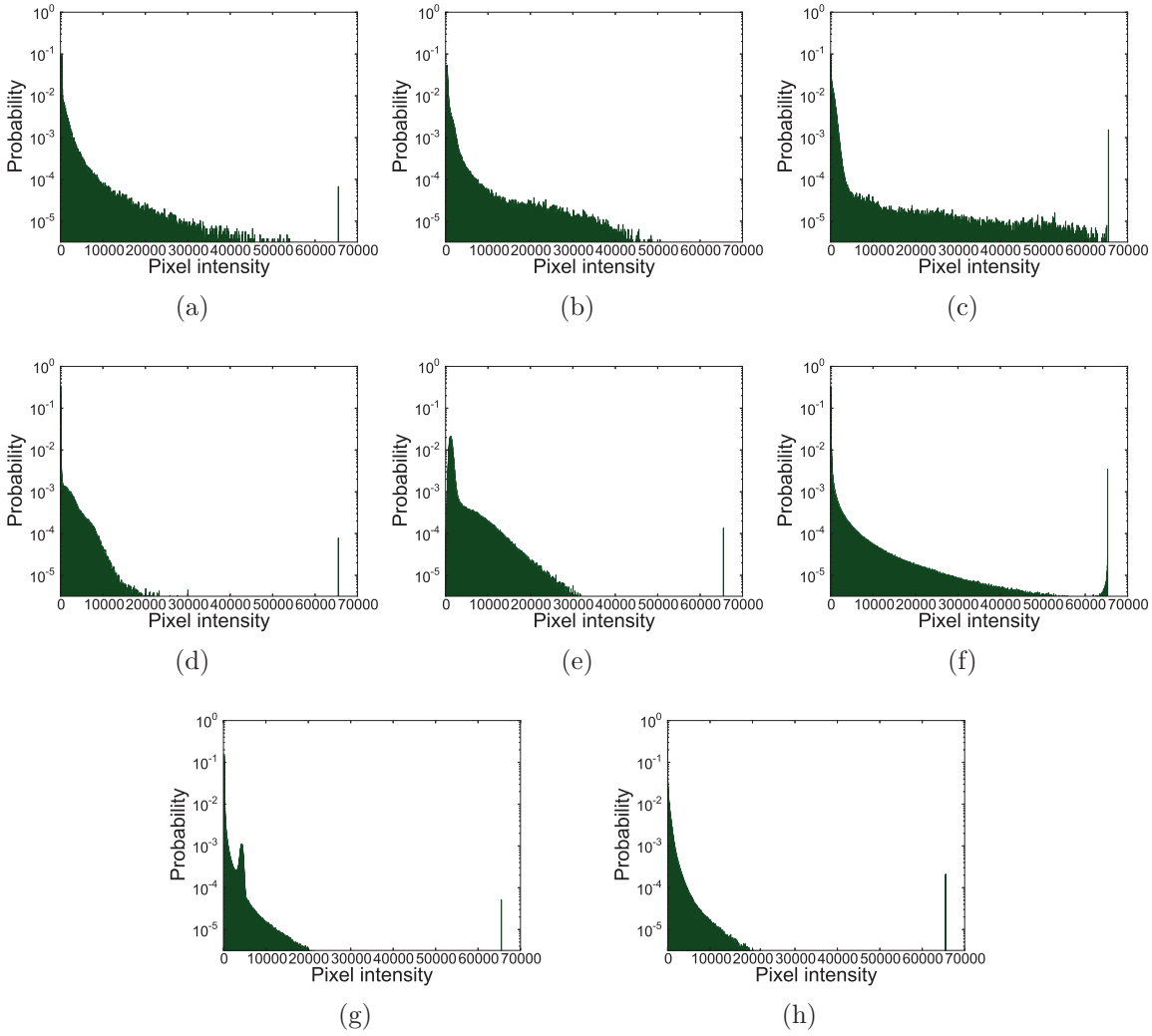


Figure 2.3: Pixel intensity distribution of different DNA microarray images using a semilogarithmic scale. (a) *y744n101_ch1* from the Yeast set; (b) *1230ko1G* from the ApoA1 set; (c) *Def661Cy3* from the ISREC set; (d) *TB3_95_llama_005* from the Stanford set; (e) *array1* from the MicroZip set; (f) *GSM346097* from the Omnibus set; (g) *slide_1-red* from the Arizona set; (h) *134044018_Cy3* from the IBB set.

Chapter 3

Lossless Compression

The lossless compression of DNA microarray images had been addressed in several publications before the beginning of this thesis. The two main approaches employed in them are segmentation and context-based coding. As discussed in Chapter 2, there exist a solid theoretical justification of using these approaches.

In the first approach, microarray image pixels are divided into spots and background. The underlying hypothesis is that spot pixels have an intensity distribution essentially different from that of background pixels. Hence, if these two types of pixels are coded separately, a performance gain can be expected. In all these methods, the image is segmented into spot and background pixels, which are coded separately using a lossless compressor. A binary mask signaling the position of the foreground also needs to be losslessly coded so that the decoder can reconstruct the original image. Fig. 3.1 depicts a general compression process using this approach. In 2003, Jörsten *et el.* proposed a fixed segmentation based on a-priori information of the spot positions and sizes [14]. Also in 2003, Faramarzpour *et al.* proposed a lossless coder whose segmentation stage consists of two steps [13]. First, a square region is obtained for each spot using the grid spatial regularity of microarray images, apparent in Fig. 3.2. Second, the centroid of that region is found and a spiral scanning is started from that point, which is assumed to be inside the spot. All pixels in the spiral path are tagged as spot until a low-intensity pixel is found. Then the spiral scanning is stopped and other remaining pixels are tagged as background. Later, in 2004, Lonardi and Luo

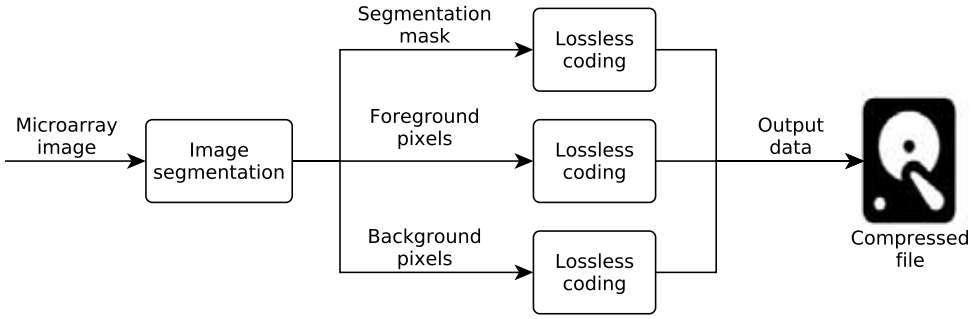
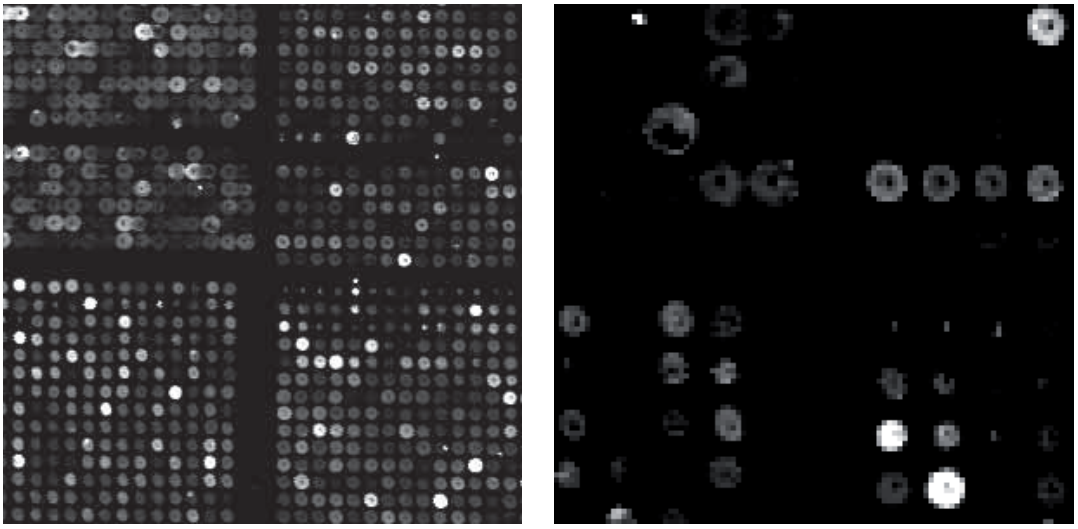


Figure 3.1: Diagram of segmentation-based compression.

presented their MicroZip compression software [15]. They used a variation of Faramarzpour's spot region finding idea, but they considered the existence of subgrids, *i.e.*, groups of spots isolated from one another, to improve the segmentation accuracy. The boundary of four such subgrids can be observed in Fig. 3.2. Also in 2004, Hua *et al.* proposed a segmentation technique that applies a statistical approach to decide whether two independent sets of pixels share a common distribution [16]. Using this technique, background pixels are grouped together and separated from background pixels. In 2006, Bierman *et al.* described a simple segmentation method based on thresholding. A threshold θ is selected from $\{2^8, 2^9, 2^{10}, 2^{11}\}$ so that approximately 90% of the pixels have intensities smaller than θ. These pixels are tagged as background, and the rest as foreground. In 2007, Neekabadi *et al.* proposed another threshold-based technique for segmentation [19], with the particularity that it divided the image into three subsets –background, edge and spot pixels–, each of which is coded separately. First, a threshold θ is chosen to minimize the variance of pixels with intensities smaller than θ. The spot pixels are those with intensities larger than θ. Then morphological growing, erosion and intersection operations are applied so that each connected group of spot pixels is totally surrounded by edge pixels. Finally, Battiato *et al.* described a techniqued that employs cellular neural networks (CNN) for segmentation, and then palette reindexing for improving the compression efficiency of the popular compressor PNG [20].

In the second main approach, context-based coding, each pixel is encoded using information present in neighbor pixels, *i.e.*, the context. This information is used to



(a) (b)

Figure 3.2: Crop of a microarray image with gamma levels adjusted, exhibiting its grid and subgrid structure. (a) 300×300 crop at $1\times$ magnification; (b) 100×100 crop at $3\times$ magnification.

estimate the probability of the pixel being encoded and drive an entropy coder –*e.g.*, an arithmetic coder–, whose efficiency depends on the accuracy of that prediction. A diagram of this general approach is shown in Fig. 3.3. In 2005, Zhang *et al.* defined a mixture probability model based on the gamma transform which is also based on segmentation [17]. First, the image is segmented into spot and background pixels and a different probability distribution is defined for each subgroup. When coding a pixel, the probability distributions of the neighbors are combined –hence the term *mixture*– to estimate the probability of that pixel. Later, in 2009, Neves and Pinho [21] proposed an image-dependent context modeling algorithm. Before coding the pixels, a greedy search is performed to find a nearly optimal finite Markov model in which the number of considered neighbors and their position depend on the particular image. The goodness of each model is measured by the conditional entropy associated to the resulting contexts for that model in the image. Conditional entropy is known to be a good predictor of the efficiency of context-driven arithmetic coding. Once the Markov model is decided, it is signaled as header information so that both

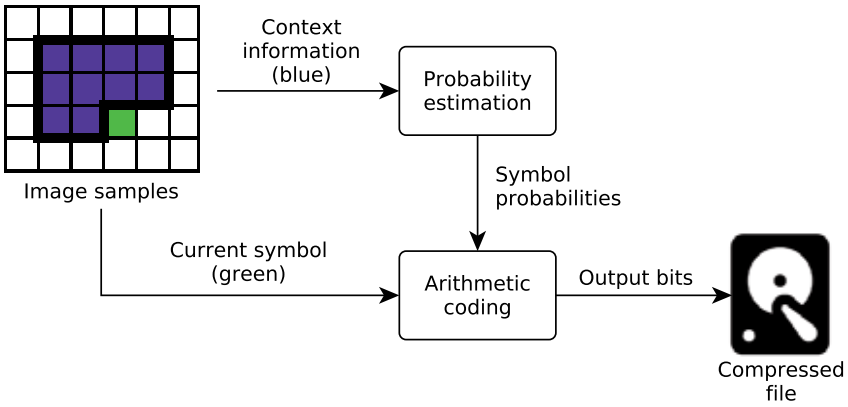


Figure 3.3: Coding of a single symbol (shown in green) using information from the context (shown in blue).

coder and decoder can make the same probability estimations, a necessary condition for lossless coding. To the best of our knowledge, Neves and Pinho's method yields the best reproducible lossless compression results for DNA microarray images.

As new analysis software is developed, working implementations of the decoders of these algorithms will be required to recover the image data. However, all the aforementioned algorithms require *ad-hoc* decoders that do not comply with existing compression standards. Hence, the availability of decoders for future platforms is not guaranteed. On the other hand, standard-compliant decoders are more likely to be ported to future platforms. Hence, standard-compliance is an important feature for the long-term storage of images such as DNA microarray images. In addition, compatibility with the Digital Imaging and Communications in Medicine (DICOM) standard [41] is paramount for compressed microarray images to be used in clinical scenarios. Therefore, in spite of the competitive compression results of previous lossless algorithms, which reach the theoretical limit described in Chapter 2, efficient standard algorithms for DNA microarray images are highly desirable. In order to simultaneously comply with this standard and to provide valuable features such as quality scalability and spatial random access to the compressed images, special attention has been paid to the JPEG2000 [29] compression standard. The two main contributions of this part of the thesis, consisting in the adaption and improvement of standard compressors to DNA microarray images, are the following:

- The lossless coding performance of different image compression standards was previously addressed in the literature [42]. When applied directly to microarray images, JPEG2000 exhibits poor compression results as compared to other standard compressors. As described in Chapter 2, this is due to microarray images having properties very different from those of natural images, for which JPEG2000 was designed. In particular, the abrupt intensity changes and the 16 bit sample precision precludes the discrete wavelet transform (DWT) from yielding good results. Furthermore, the pixel distribution of microarray images is very different from that for which JPEG2000 produces optimal results. Section 3.1 describes a lossless point transform –the Histogram Swap Transform (HST)– able to improve the coding performance of JPEG2000 by adapting microarray image histograms with a very low computational cost.

- In Section 3.2, multicomponent decorrelation, a not previously explored approach to improving the lossless compression performance of standard or microarray-specific coding algorithms, is researched. The redundancy present in similar images is exploited in order to reduce the overall entropy and, consequently, to improve the overall compression performance. In that section, an optimal grouping of images is proposed to maximize the inter-group correlation. Furthermore, the performance results of several spectral decorrelation transforms are evaluated.

3.1 The Histogram Swap Transform

```
@inproceedings{Hernandez12DCC,
               Title      = {DNA Microarray Image Coding},
               Author     = {Miguel Hern{\'a}ndez-Cabronero, and Juan Mu{\~n}oz-G{\'o}mez,
                             and Ian Blanes, and Joan Serra-Sagrist{\`a}, and Michael W. Marcellin,}
               Booktitle  = {Proceedings of the IEEE Data Compression Conference, DCC},
               Year       = {2012},
               Pages      = {32-41},
               ISBN       = {9781612842790},
               doi        = {10.1109/DCC.2012.11},
}
```

2012 Data Compression Conference

DNA microarray image coding

Miguel Hernández-Cabronero[†], Juan Muñoz-Gómez[†], Ian Blanes[†]
Michael W. Marcellin[†‡] and Joan Serra-Sagristà[†]

[†] Department of Information and Communications Engineering,
Universitat Autònoma de Barcelona, Barcelona, Spain.

[‡] Department of Electrical and Computer Engineering
University of Arizona, Tucson, AZ, USA.

Abstract

DNA microarrays are useful to identify the function and regulation of a large number of genes in a single experiment, even whole genomes. In this work, we analyze the relationship between DNA microarray image histograms and the compression performance of lossless JPEG2000. Also, a reversible transform based on histogram swapping is proposed. Intensive experimental results using different coding parameters are discussed. Results suggest that this transform improves previous lossless JPEG2000 results on all DNA microarray image sets.

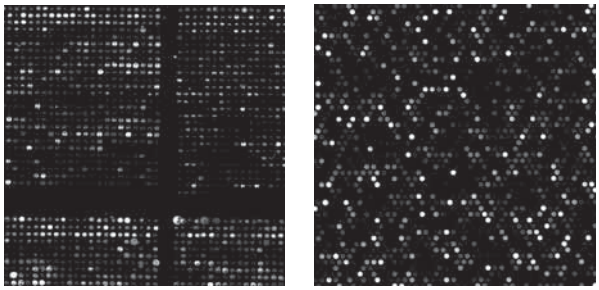
1 Introduction

1.1 DNA microarrays

DNA microarrays are a state of the art tool in medicine and biology for the study of genetic function, regulation and interaction [1]. Genome-wide monitoring is possible with existing DNA microarrays, which are used in research against cancer [2] and HIV [3], among many other applications. DNA microarrays consist of a solid surface on which thousands of different known genetic sequences, the oligonucleotides, are bound. Each sequence is contained in a single microscopic hole or *spot* and all spots are arranged conforming to a regular pattern, usually a rectangular or hexagonal grid. Example images for these two layouts are shown on Figure 1. Two samples dyed with fluorescent markers, usually Cy3 and Cy5 of the cyanine family, are made to react on the microarray. When one sample has expressed a gene, part of it is hybridized and adhered to the spot corresponding to that gene. The rest is washed away so that each dye is present in a spot proportionally to the activity of a gene in the corresponding sample. After the hybridization, the microarray is exposed to laser beams and the emissions from the fluorescent Cy3 and Cy5 dyes are recorded independently as so-called green and red channel images, respectively. Comparing the relative intensity of the green and red channels, it is possible to detect expression differences between two samples, which can be employed to make hypothesis about the function and regulation of thousands of individual genes.

Each microarray experiment outputs a pair of monochrome, single component images corresponding to the green and red channels. Due to the microscopic size of the spots, the produced images have a high spatial resolution: images from 1000×1000 onwards are typically described in the literature, but sizes over 4000×13000 are common nowadays. Since gene expression can vary in a very wide range and a high degree of precision is desired, DNA microarray images have a intensity resolution of 16 bits per pixel (bpp).

1068-0314/12 $26.00 © 2012 IEEE
DOI 10.1109/DCC.2012.11



(a) (b)

Figure 1: Example DNA microarray image: 600×600 crop with different spot layouts. a) *array3* image from the MicroZip set with square grid spot layout; b) *slide_1-red* from the Arizona set with hexagonal grid spot layout. Gamma levels have been adjusted for better viewing.

After the images have been recorded, they are computer analyzed to extract the genetic information present in them. However, analysis techniques are not fully mature or universally accepted, so it is preferable to keep the original images and not only the extracted genetic data because repeating an experiment is expensive and not always possible. Because of the high spatial and intensity resolutions, raw data for a single DNA microarray image can require from a few to hundreds of Megabytes. Most experiments are carried out under several different conditions, and with the increasing interest in DNA microarrays, very large amounts of data are created each year around the world. DNA microarray images need to be kept and shared, so efficient storage and transmission methods are required. In consequence, compression emerges as a natural approach.

Both lossy and lossless techniques have been proposed in the literature. Lossy approaches exhibit better compression performance on microarray images, but information loss is not globally accepted because it could affect reanalysis with future techniques. On the other hand, purely lossless methods guarantee perfect fidelity of the data, which is preferable for future reanalysis, at the cost of poorer compression performance as compared to lossy techniques. The efficient lossless compression of this type of image has proved to be a difficult task. This is partly due to the considerable amount of noise and the abundance of high frequencies present in this type of image. For this reason, original approaches like the one proposed in this work are needed to achieve the storage and transmission requirements for DNA microarray images.

1.2 State of the art in lossless compression

Many different techniques have been proposed for the lossy and lossless compression of DNA microarray images. In this subsection, we discuss lossless schemes that have been published in the literature. The typical image compression process consists of up to five stages: preprocessing, transform, quantization, entropy coding and postprocessing. Microarray image compression can be modeled likewise, but not all stages are equally relevant if we focus only on lossless compression: the quantization stage, which consists of dividing sets of values or vectors into groups, effectively reducing the total number of symbols needed to represent them, is not usually considered for lossless compression; the same happens for the postprocessing stage, consisting of processing images after compression to enhance their visual quality, to provide new features or to analyze their properties. Lossless techniques belonging to the rest of the stages are addressed next. A more exhaustive review of the state of the art can be found in the literature [4].

1.2.1 Preprocessing

The preprocessing stage comprises any computation performed on an image to prepare it for the compression or analysis processes. It is very important in DNA microarray images because many of the existing techniques rely heavily on the results of this stage to obtain competitive coding performance. The main preprocessing method is segmentation and consists in determining which of the image pixels belong to spots (i.e., the foreground), as opposed to those that do not (i.e., the background).

In 2003, Faramarzpour et al. proposed a segmentation stage consisting of two steps [5]: first spot regions are located by studying the row intensity sum minima, and then region centroids are used to estimate the spot centers. Simpler versions of this spot region location idea had already been used by Jörnsten and Yu in 2002 [6]. Later, in 2004, Lonardi and Luo presented their MicroZip software [7], which used a variation of Faramarzpour's spot region finding idea, but considering the existence of subgrids, which can be appreciated in Figure 1a. In 2004, Hua et al. proposed a scheme with a segmentation technique based on the Mann-Whitney U test [8]. In 2006, Bierman et al. described a simple thresholding method for dividing microarray images into low and high intensities [9], determining the lowest of the threshold values from 2^8, 2^9, 2^{10} or 2^{11} such that at least 90% of the pixels fall within it. In 2007, Neekabadi et al. proposed another threshold-based technique for segmentation [10] in three subsets (background, edge and spot pixels), using a threshold that minimizes the total standard deviation of pixels above and below it. In 2009, Battiato and Rundo published an approach based on Cellular Neural Networks (CNNs) [11].

1.2.2 Transform

The transform stage consists of changing the image domain from the spatial domain to a domain where it can be more efficiently processed or coded. However, transform based compression is not typically as efficient for DNA microarray images as it is for other types of images not containing such sharp edges [12]. For this reason, transformations are not frequently researched in microarray image compression, although they are used in some works.

In 2004, Hua et al. [8] published a modification of the EBCOT algorithm, the basis of the JPEG2000 standard [13], that included a tailored integer odd-symmetric transform. In 2004, Lonardi and Luo [7] made use of the Burrows-Wheeler transform [14] for lossy or lossless compression in their MicroZip software.

Table 1: Classification of lossless microarray-specific techniques discussed on Subsection 1.2, sorted chronologically.

Preprocessing	**Transform**	**Entropy coding**	
Segmentation		Segmentation	Context
[6], 2002	[8], 2004	[5], 2003	[15], 2005
[5], 2003	[7], 2004	[9], 2006	[16], 2006
[8], 2004		[11], 2009	[17], 2006
[7], 2004			[18], 2009
[9], 2006			
[10], 2007			
[11], 2009			

1.2.3 Entropy coding

In this stage, data obtained from previous stages are expressed in an efficient manner to produce a compact bitstream. Many techniques segment the image before compression,

while others build contexts or try to predict the intensity of the next pixels based on the previous ones. Purely lossless techniques using each approach are described next.

At least three different works that use segmentation can be found on lossless compression of DNA microarrays. In 2003, Faramarzpour et al. presented a prediction-based technique [5]. The image is gridded, and a linear prediction scheme is applied after creating a spiral path from the estimated spot center. In 2006, Bierman et al. presented their MACE (Micro Array Compression and Extraction) software [9]. The image is divided first using a threshold-based method. The low intensity pixels are coded using standard dictionary-based techniques, while the high intensity pixels are processed with a sparse matrix algorithm and then compressed. In 2009, Battiato and Rundo published an algorithm [11] based on image color reindexing after segmentation. Segmentation is made by means of a CNN-based system to produce two complementary subimages. The foreground image is compressed with a generic lossless algorithm and stored separately. The background image is first transformed into an indexed image. Then its color palette is reindexed with an algorithm that reduces the zero-order entropy of local differences, which are losslessly coded.

In no less than four publications, context building is used to perform lossless DNA microarray image compression. In 2005 and in 2006 Zhang et al. [15, 16] proposed a context-based lossless approach that also employs segmentation. Once the image is divided, a simple predictive scheme is used for the most significant bytes of each pixel, while the least significant bytes are coded using prediction by partial approximate matching (PPAM), also proposed by Zhang and Adjeroh [19]. In 2006, Neves and Pinho [17] proposed another context-based lossless approach. It is a bitplane-based technique that uses 3D finite-context models to drive an arithmetic coder. In 2009, they improved this scheme so that specific contexts are built for each image [18].

Table 1 presents a summary of all discussed methods classified attending to the stage of the image compression process in which they make their contribution.

1.3 Paper structure

This paper is organized as follows. We discuss the use of lossless JPEG2000 on DNA microarray images in Section 2. In Section 3, we analyze the typical histogram of a DNA microarray image and propose a point transform based on histogram swapping. Results for the application of this transform with lossless JPEG2000 are presented. Finally, we draw some conclusions in Section 4.

2 Lossless JPEG2000 coding of DNA microarray images

In this section, we study the performance of lossless JPEG2000 compression on DNA microarray images. We describe the image sets used for benchmarking in the literature and some of their properties in Subsection 2.1. We show lossless JPEG2000 compression results and compare them to previous data and other techniques in Subsection 2.2. We analyze the impact of the number of DWT decomposition levels and quality layers on the compression performance in Subsection 2.3.

2.1 Benchmark image sets

A number of different DNA microarray image sets have been used for benchmarking compression performance. No set has been used across all publications on DNA microarray image compression, but the MicroZip, ApoA1 and ISREC sets are employed more

Table 2: Image sets used in the literature.

Image set	Images	Size (px)
MicroZip [21]	3	$> 1800 \times 1900$
Yeast [22]	109	1024×1024
ApoA1 [23]	32	1044×1041
ISREC [24]	14	1000×1000
Stanford [20]	20	$> 2000 \times 2000$
Arizona	6	4400×13800

Table 3: Compression results in bpp for the Kakadu and JJ2000 implementation of the JPEG2000 standard, and for some other generic and microarray-specific techniques. Best results are highlighted in **green** and worst results in **red**.

Algorithm	MicroZip	Yeast	ApoA1	ISREC	Stanford	Arizona
Kakadu (5 levels)	9.508	**9.082**	11.052	11.360	8.007	9.099
JJ2000 (5 levels)	**9.515**	9.079	11.063	**11.366**	**8.010**	**9.106**
Bzip2	9.394	**6.075**	**11.067**	10.921	7.503	8.944
CALIC	9.281	8.502	10.515	10.615	7.248	8.767
JBIG	9.297	6.888	10.851	10.925	7.411	8.858
JPEG-LS	8.974	8.580	10.608	11.145	**7.204**	**8.646**
Battiato's [11]	**8.369**	–	**9.52**	**9.49**	–	–

frequently. The Stanford set was obtained from the Stanford Microarray Database public FTP [20] and the Arizona set has been kindly provided by David Galbraith and Megan Sweeney from the University of Arizona. Table 2 shows key properties of all sets documented in the literature. All images are monochrome, unsigned, 16 bits per pixel (bpp), and contain a single component per red/green channel.

2.2 Lossless compression performance

In this section, we report an experiment that we have conducted to test lossless JPEG2000 compression performance on DNA microarray images. We have compressed all images from the sets described in Subsection 2.1 using the Kakadu v6.0 [25] and the JJ2000 v5.1 [26] implementations of the JPEG2000 standard. In both cases, we have used lossless compression, 33 quality layers and 5 DWT decomposition levels. The number of quality layers was chosen to be the same as in a previous work by Pinho [27]. All codestreams are JPEG2000 part 1 compliant.

Table 3 shows compression results for the Kakadu and the JJ2000 implementation employing the mentioned configuration, and also for a generic compressor (Bzip2), some general image compressors (CALIC, JBIG and JPEG-LS) and the best microarray-specific compressor (Battiato's), as reported in a previous work [4]. These results have been computed dividing the total size in bits for all compressed files by the total number of pixels in the images.

It can be seen that both JPEG2000 implementations exhibit very similar compression performances, which are poor compared to the best microarray-specific technique, and generally to the other compressors as well. The results obtained with the JJ2000 implementation are consistent to the ones published by Pinho [27].

2.3 DWT decomposition levels and quality layers

In our experiments, we analyze the impact of varying the number of DWT decomposition levels and quality layers when using Kakadu JPEG2000 on DNA microarray images.

Table 4: Lossless compression results in bpp for different DWT decomposition levels using Kakadu JPEG2000. Best results are highlighted in **green** and worst results in **red**.

DWT levels	MicroZip	Yeast	ApoA1	ISREC	Stanford	Arizona
0	**10.027**	**6.829**	**11.525**	**10.888**	**8.567**	**9.548**
1	9.542	**9.089**	11.088	**11.476**	8.146	9.221
3	9.472	9.042	**10.999**	11.312	7.985	9.068
5	**9.467**	9.038	**10.999**	11.314	**7.969**	**9.064**

Table 4 shows results for 0, 1, 3 and 5 DWT decomposition levels and 1 quality layer.

For most sets, compression is improved by approximately 0.5 bpp when the number of decomposition levels is increased from 0 to 5, but most of that improvement is yielded when increasing from 0 to 1 level. Only the Yeast and ISREC sets break that pattern: for these two sets, using 0 decomposition levels produces the best results. Increasing to 1 decomposition level degrades measurably the performance, but further level increments improve the performance slightly as happens for the other sets. For all sets, using more than 5 decomposition levels does not modify the compression performance.

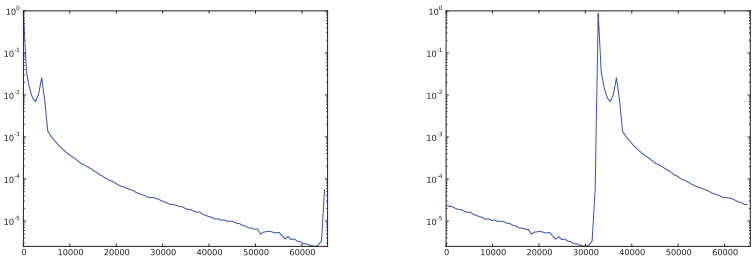
Increasing the number of quality layers from 1 to 33 decreases compression performance slightly, between 0.03 bpp and 0.04 bpp for all tested DWT decomposition levels and sets.

3 DNA microarray image histograms and JPEG2000

This section is organized as follows. In Subsection 3.1, we describe the typical DNA microarray image histogram and compare it to the histogram that JPEG2000 implicitly assumes. In Subsection 3.2 we propose an original transform based on histogram swapping and report compression results when employing lossless JPEG2000 after this transform.

3.1 Typical DNA microarray image histograms

DNA microarray images exhibit similar pixel intensity distributions across all data sets in Table 2. In a typical image, most pixels belong to the background or to low-activity spots and have very low values. Higher intensities are several orders of magnitude less frequent, but among them the intensities close to the maximum are usually over ten times more abundant. The histogram of a representative DNA microarray image from the Arizona set, *slide_1-red*, is shown in Figure 2a.



(a) Original histogram. (b) Histogram after the proposed transform.

Figure 2: Pixel value distribution for original and transformed image *slide_1-red* from the Arizona set using a semilog scale.

The characteristic distribution of this type of images partly explains the poor lossless compression results of JPEG2000 shown on Table 4. For unsigned image data with bit-depth B, the first step carried out by a JPEG2000 Part 1 encoder is to subtract 2^{B-1} from the value of each pixel [13]. This is known as the level offset stage and typically results in pixel values nominally distributed symmetrically about the origin. DNA microarray images are unsigned. Thus, the subtraction is performed resulting in a highly asymmetrical histogram, with the majority of pixel intensities taking values near -32768. Another problem is that microarray images have considerable high frequency content due to the many edge discontinuities between spots and background. This type of data is not well treated by the wavelet transform. In conclusion, JPEG2000 is receiving an input for which it is not designed, so a high compression performance cannot be expected.

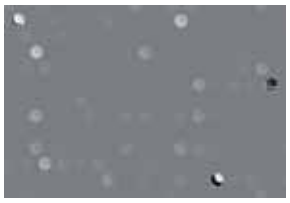
3.2 Histogram swapping and lossless JPEG2000 compression

DNA microarray images possess pixel value distributions that diverge from natural images. However, DNA microarray images can be modified so that their intensity histograms become more similar to what JPEG2000 implicitly assumes. If the most significant bit of each pixel of an image is flipped, the right half of the histogram is swapped for the left half. This transformation, which we will call the *histogram swap transform* (HST), can be easily reversed by flipping again the most significant bit of each pixel. Figure 2b shows the pixel distribution of the transformed version of image *slide_1-red*. This histogram is much more nearly symmetric about the origin.

We have conducted an experiment to test JPEG2000 lossless compression on DNA microarray images after applying the proposed transform. We have used 1 quality layer and 0, 1, 3, and 5 DWT decomposition levels, and we report the results on Table 5. It can be observed that the compression performance is always improved when using the HST. Comparing the results for the best choice of decomposition levels before and after the HST, rate improvements from 0.213 bpp to 0.918 bpp (1.97% to 15.53%) can be measured. It is also noteworthy that the compression results follow a different trend after altering the image histograms. When compressing the original images, the performance is generally improved as the number of DWT decomposition levels is increased, as previously shown on Table 4. However, after applying the HST, this pattern is reversed and performance is generally degraded when the number of DWT decomposition levels is increased. This behavior can be explained via two observations. First, the histogram of Figure 2b is very peaked near the origin, reminiscent of the Laplacian distribution often assumed for wavelet transform coefficients [28]. This suggests that the bitplane coder of JPEG2000 may work well when applied directly to the data obtained via the HST (without further transformation). Second, when the HST is applied on an unsigned image with bit-depth B, pixel values slightly smaller than 2^{B-1} become close to 2^B, while values slightly greater than 2^{B-1} become close to zero. In other words, mid-gray values (pre HST) result in abrupt intensity differences between near black and near white (post HST). Examples of this behavior can be appreciated in Figure 3. This adds to the abundant discontinuities already present in DNA microarray images. Thus, increasing the number of DWT decomposition levels, which is not very effective when high frequencies are present [7], results in a performance reduction.

There are other point transforms that lead to histograms similar to Figure 2b. In fact, minor additional performance improvements might be obtained in this fashion. However, the HST has a distinct advantage in terms of implementation. In fact, it is possible to implement the HST without any explicit changes in JPEG2000 or the image data itself. Indeed, if the original unmodified data are simply interpreted as twos-complement signed values, they have exactly the same decimal values that result when the unsigned values are subjected to the HST followed by the JPEG2000 level offset process. Specifically when

(a) Detail of *array3* image from the MicroZip set after the HST.



(b) Detail of *slide-1_red* image from the Arizona set after the HST.

Figure 3: Details of sudden intensity changes after the HST.

Table 5: Lossless compression results in bpp applying Kakadu JPEG2000 after applying the HST, using 1 quality layer and different DWT decomposition levels. Rate differences in bpp as compared to compression of unmodified images. Differences between the best results before and after HST are shown at the bottom.

DWT levels	MicroZip	Yeast	ApoA1	ISREC	Stanford	Arizona
0	9.157 (-0.870)	5.911 (-0.918)	10.786 (-0.739)	10.624 (-0.264)	7.685 (-0.882)	8.795 (-0.753)
1	9.297 (-0.245)	8.862 (-0.227)	10.917 (-0.171)	11.238 (-0.238)	7.851 (-0.295)	8.967 (-0.254)
3	9.455 (-0.017)	9.026 (-0.016)	11.003 (+0.004)	11.300 (-0.012)	7.950 (-0.035)	9.058 (-0.010)
5	9.466 (-0.001)	9.035 (-0.003)	11.012 (+0.013)	11.313 (-0.001)	7.958 (-0.011)	9.070 (+0.006)
Best	9.157 (-0,310)	5.911 (-0,918)	10.786 (-0,213)	10.624 (-0,264)	7.685 (-0,284)	8.795 (-0,269)

interpreted as twos-complement values, pixels between 0x0000 and 0x7FFF yield decimal values between 0 and 32767. Pixels between 0x8000 and 0xFFFF yield values between -32768 and -1. On the other hand, interpreting the data as unsigned and applying the HST results in values between 0x0000 and 0x7FFF being transformed to values from 0x8000 to 0xFFFF, with decimal equivalents 32768 to 65535. After the JPEG2000 level offset, values from 0 to 32767 are obtained. Similarly, values from 0x8000 to 0xFFFF become 0x0000 to 0x7FFF (or 0 to 32767) after the HST and -32768 to -1 after the level offset. Thus, HST followed by JPEG2000 can be performed by simply applying JPEG2000 to the data as if it were signed, even though it is unsigned. Encoding, decoding and the resulting codestreams are all JPEG2000 Part 1 compliant.

4 Conclusion

DNA microarray images are becoming commonplace for genome-wide monitoring, employed intensively in many medical treatments and biological research. The large size of these images motivates the use of coding techniques to help storing and transmitting them. Lossy coding approaches provide better performance than lossless coding techniques, but they are not always accepted because of the possible negative influence on later classification processes.

Lossless compression results for two different JPEG2000 implementations as well as for other schemes have been discussed. We have tested the performance impact of using different numbers of quality layers and DWT decomposition levels and we have concluded that, for most image sets, the best parameters choice is 1 quality layer and 5 DWT decomposition levels. However, it has been observed that lossless JPEG2000 performance is poor when compared to the best microarray-specific technique, and even to some general image compressors.

A reversible transform based on histogram swapping, which draws images closer to JPEG2000 assumptions for context modeling, has been proposed. With this modifica-

tion, the performance of lossless JPEG2000 compression is improved for all image sets. Rate improvements from 0.213 bpp to 0.918 bpp, corresponding to percentage increases of, respectively, 1.97% and 15.53%, have been measured. The histogram swap transform is easily implemented in a JPEG2000 part 1 compliant manner.

Acknowledgements

MicroZip corpus was kindly provided by Neves and Pinho from the University of Aveiro. The Arizona image set was provided by David Galbraith and Megan Sweeney from the University of Arizona.

This work has been partially supported by the European Union, by the Spanish Government (MICINN), by FEDER, and by the Catalan Government under Grants FP7-PEOPLE-2009-IIF FP7-250420, TIN2009-14426-C02-01, UAB-BI3INT2006-08, and 2009-SGR-1224.

References

[1] S. Moore, "Making chips to probe genes," *IEEE SPECTRUM*, vol. 38, no. 3, pp. 54–60, MAR 2001.

[2] S. Satih, N. Chalabi, N. Rabiau, R. Bosviel, L. Fontana, Y.-J. Bignon, and D. J. Bernard-Gallon, "Gene Expression Profiling of Breast Cancer Cell Lines in Response to Soy Isoflavones Using a Pangenomic Microarray Approach," *OMICS-A JOURNAL OF INTEGRATIVE BIOLOGY*, vol. 14, no. 3, pp. 231–238, JUN 2010.

[3] M. S. Giri, M. Nebozhyn, L. Showe, and L. J. Montaner, "Microarray data on gene modulation by HIV-1 in immune cells: 2000-2006," *JOURNAL OF LEUKOCYTE BIOLOGY*, vol. 80, no. 5, pp. 1031–1043, NOV 2006.

[4] M. Hernández-Cabronero, I. Blanes, M. W. Marcellin, and J. Serra-Sagristà, "Standard and specific compression techniques for dna microarray images," *MDPI Algorithms*, vol. 4, 2012, In press.

[5] N. Faramarzpour, S. Shirani, and J. Bondy, "Lossless DNA microarray image compression," in *In Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers*, vol. 2, November 2003, pp. 1501–1504.

[6] R. Jornsten, Y. Vardi, and C. Zhang, "On the bitplane compression of microarray images," in *In Proceedings of the 4th International Conference on Statistical Data Analysis Based on the L1-Norm and Related Methods*, 2002.

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

[8] 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.

[9] R. Bierman, N. Maniyar, C. Parsons, and R. Singh, "MACE: lossless compression and analysis of microarray images," in *In Proceedings of the ACM symposium on Applied computing*, ser. SAC '06. ACM, 2006, pp. 167–172.

[10] A. Neekabadi, S. Samavi, S. A. Razavi, N. Karimi, and S. Shirani, "Lossless microarray image compression using region based predictors," in *In Proceedings of the*

International Conference on Image Processing, vol. 1-7. IEEE, 2007, Proceedings Paper, pp. 913–916.

[11] S. Battiato and F. Rundo, "A bio-inspired CNN with re-indexing engine for lossless dna microarray compression and segmentation," in *In Proceedings of the 16th International Conference on Image Processing*, vol. 1-6, IEEE. IEEE, 2009, Proceedings Paper, pp. 1717–1720.

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

[13] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Kluwer Academic Publishers, Boston, 2002.

[14] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm." HP, Tech. Rep. 124, 1994.

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

[16] D. A. Adjeroh, Y. Zhang, and R. Parthe, "On denoising and compression of DNA microarray images," *Pattern Recognition*, vol. 39, no. 12, pp. 2478–2493, December 2006.

[17] A. J. R. Neves and A. J. Pinho, "Lossless compression of microarray images," in *In Proceedings of the International Conference on Image Processing, ICIP*. IEEE, 2006, Proceedings Paper, pp. 2505–2508.

[18] ——, "Lossless compression of microarray images using image-dependent finite-context models," *IEEE Transactions on Medical Imaging*, vol. 28, no. 2, pp. 194–201, February 2009.

[19] Y. Zhang and D. Adjeroh, "Prediction by partial approximate matching for lossless image compression," *IEEE Transactions on Image Processing*, vol. 17, no. 6, pp. 924 –935, June 2008.

[20] "Stanford Microarray Database public FTP." [Online]. Available: ftp://smd-ftp.stanford.edu/pub/smd/transfers/Jenny

[21] "MicroZip test image set." [Online]. Available: http://www.cs.ucr.edu/~yuluo/MicroZip

[22] "Stanford Yeast Cell-Cycle Regulation Project." [Online]. Available: http://genome-www.stanford.edu/cellcycle/data/rawdata/individual.html

[23] "ApoA1 experiment data." [Online]. Available: http://www.stat.berkeley.edu/users/terry/zarray/Html/apodata.html

[24] "ISREC image set." [Online]. Available: http://www.isrec.isb-sib.ch/DEA/module8/P5_chip_image/images/

[25] D. Taubman, "Kakadu Software home page." [Online]. Available: http://www.kakadusoftware.com/

[26] "JJ2000 Software home page." [Online]. Available: http://code.google.com/p/jj2000/

[27] A. Pinho, A. Paiva, and A. Neves, "On the use of standards for microarray lossless image compression," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 563–566, March 2006.

[28] B. E. Usevitch, "A tutorial on modern lossy wavelet image compression: foundations of jpeg 2000," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 22–35, 2001.

3.2 Multicomponent Transformations

This section consists on the exploration of multicomponent transformation techniques for the compression performance improvement for DNA microarray images. It was published as a part of a book chapter in

```
@book{Hernandez13RuedaBook,
      title = {Compression of DNA Microarray Images},
      author = {Miguel Hern{\'a}ndez-Cabronero, and Victor Sanchez,
                and Michael W. Marcellin, and Joan Serra-Sagrist{\`a},}
      booktitle = {In Book "Microarray Image and Data Analysis: Theory and Practice", CRC Press},
      year = {2014},
      pages = {193-225},
      editor = {Luis Rueda},
      publisher = {CRC Press},
      url = {http://www.crcpress.com/product/isbn/9781466586826},
},
```

which also describes other contributions of this thesis. For space economy and to avoid duplicities among different parts of this thesis, only the contents of the following reference are included:

```
@inproceedings{Hernandez12Sarteco,
               title = {Multicomponent compression of DNA microarray images},
               author = {Miguel Hern{\'a}ndez-Cabronero, and Francesc Aul{\'i}-Llin{\`a}s,
                         and Joan Bartrina-Rapesta, and Ian Blanes,
                         and Leandro Jim{\'e}nez-Rodr{\'i}guez,
                         and Michael W. Marcellin, and Juan Mu{\~n}oz-G{\'o}mez,
                         and Victor Sanchez, and Joan Serra-Sagrist{\`a}, and Zhongwei Xu,}
               booktitle = {Proceedings of the CEDI Workshop on Multimedia Data Coding
                            and Transmission, WMDCT},
               year = {2012},
}.
```

This alternative reference contains an equivalent description of the contributions on spectral decorrelation, which were thereafter accepted as a part of the book chapter.

Multicomponent compression of DNA microarray images

Miguel Hernández-Cabronero[1], Francesc Aulí-Llinàs[1], Joan Bartrina-Rapesta[1], Ian Blanes[1], Leandro Jiménez-Rodríguez[1], Michael W. Marcellin[1,2], Juan Muñoz-Gómez[1], Victor Sánchez[1], Joan Serra-Sagristà[1] and Zhongwei Xu[1]

Abstract— **In this work, the correlation present among pairs of DNA microarray images is analyzed using Pearson's r as a metric. A certain amount of correlation is found, especially for red/green channel image pairs, with averages over 0.75 for all benchmark sets. Based on that, the lossless multicomponent compression features of JPEG2000 have been tested on each set, considering different spectral and spatial transforms (DWT 5/3, DPCM, R-Haar and POT). Improvements of up to 0.6 bpp are obtained depending on the transform considered, and these improvements are consistent to the correlation values observed.**

Keywords— **microarray images, microarray image compression, JPEG2000, multicomponent compression**

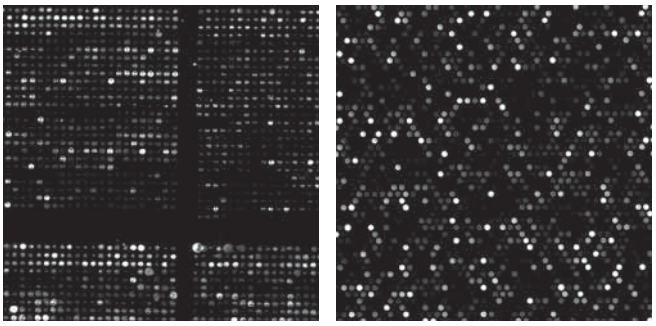
I. Introduction

A. DNA microarrays

DNA microarrays are a state of the art tool in medicine and biology for the study of genetic function, regulation and interaction [1]. Genome-wide monitoring is possible with existing DNA microarrays, which are used in research against cancer [2] and HIV [3], among many other applications. DNA microarrays consist of a solid surface on which thousands of different known genetic sequences, the oligonucleotides, are bound. Each sequence is contained in a single microscopic hole or *spot* and all spots are arranged conforming to a regular pattern, usually a rectangular or hexagonal grid. Example images for these two layouts are shown on Figure 1. Two samples dyed with fluorescent markers, usually Cy3 and Cy5 of the cyanine family, are made to react on the microarray. When one sample has expressed a gene, part of it is hybridized and adhered to the spot corresponding to that gene. The rest is washed away so that each dye is present in a spot proportionally to the activity of a gene in the corresponding sample. After the hybridization, the microarray is exposed to laser beams and the emissions from the fluorescent Cy3 and Cy5 dyes are recorded independently as so-called green and red channel images, respectively. Comparing the relative intensity of the green and red channels, it is possible to detect expression differences between two samples, which can be employed to make hypotheses about the function and regulation of thousands of individual genes.

[1]Dept. of Information and Communications Engineering, Universitat Autònoma de Barcelona, Barcelona, Spain. Contact e-mail: miguel.hernandez@uab.es

[2]Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ, USA



(a) (b)

Fig. 1: Example DNA microarray image: 600×600 crop with different spot layouts. a) *array3* image from the MicroZip set with square grid spot layout; b) *slide_1-red* from the Arizona set with hexagonal grid spot layout. Gamma levels have been adjusted for better viewing.

Each microarray experiment outputs a pair of monochrome, single component images corresponding to the green and red channels. Due to the microscopic size of the spots, the produced images have a high spatial resolution: images from 1000 × 1000 onwards are typically described in the literature, with sizes over 4000 × 13000 being common nowadays. Since gene expression can vary over a wide range, a high degree of precision is desired, DNA microarray images have a intensity resolution of 16 bits per pixel (bpp).

After the images have been recorded, they are computer analyzed to extract the genetic information present in them. However, analysis techniques are not fully mature or universally accepted, so it is preferable to keep the original images rather than only the extracted genetic data because repeating an experiment is expensive and not always possible. Because of the high spatial and intensity resolutions, raw data for a single DNA microarray image can require from a few to hundreds of Megabytes. Many DNA microarray studies consist of running several experiments at different moments on similar biological samples that have been exposed to different physical and chemical conditions. With the increasing interest in DNA microarrays, very large amounts of data are created each year around the world. DNA microarray images need to be kept and shared, so efficient storage and transmission methods are re-

quired. In consequence, compression emerges as a natural approach.

Both lossy and lossless techniques have been proposed in the literature. Lossy approaches exhibit better compression performance on microarray images, but information loss is not globally accepted because it could affect reanalysis with future techniques. On the other hand, purely lossless methods guarantee perfect fidelity of the data, which is preferable for future reanalysis, but at the cost of poorer compression performance as compared to lossy techniques. The efficient lossless compression of this type of images has proved to be a difficult task. Transform-based coding methods do not perform well on microarray images due to the considerable amount of noise and the abundance of high frequencies present in this type of images [4].

B. State of the art in lossless compression

Many different techniques have been proposed for the lossy and lossless compression of DNA microarray images. In this subsection, we discuss lossless schemes that have been published in the literature. The typical image compression process consists of up to five stages: preprocessing, transform, quantization, entropy coding and postprocessing. Microarray image compression can be modeled likewise, but not all stages are equally relevant if we focus only on lossless compression. For example, the quantization stage, which consists of dividing sets of values or vectors into groups, effectively reducing the total number of symbols needed to represent them, is not usually considered for lossless compression. On the other hand, the postprocessing stage is always addressed for microarray images, since they are always analyzed to extract genetic data, but is not described in this document. Lossless techniques belonging to the preprocessing, transform and entropy coding stages are addressed next. A more exhaustive review of the state of the art can be found in the literature [5].

B.1 Preprocessing

The preprocessing stage comprises any computation performed on an image to prepare it for the compression or analysis processes. It is very important in DNA microarray image compression because many of the existing techniques rely heavily on the results of this stage to obtain competitive coding performance. The main preprocessing method is segmentation and consists in determining which of the image pixels belong to spots (i.e., the foreground), as opposed to those that do not (i.e., the background).

In 2003, Faramarzpour et al. proposed a segmentation stage consisting of two steps [6]: first, spot regions are located by studying the minimum values of the pixel intensity sum by rows, and then region centroids are used to estimate the spot centers. Simpler versions of this spot region location idea had already been used by Jörnsten and Yu in 2002 [7]. Later, in 2004, Lonardi and Luo presented their MicroZip software [4], which used a variation of Faramarzpour's spot region finding idea, but considering the existence of subgrids, which can be appreciated in Figure 1a. In 2004, Hua et al. proposed a scheme with a segmentation technique based on the Mann-Whitney U test [8]. In 2006, Bierman et al. described a simple thresholding method for dividing microarray images into low and high intensities [9], determining the lowest of the threshold values from 2^8, 2^9, 2^{10} or 2^{11} such that at least 90% of the pixels fall within it. In 2007, Neekabadi et al. proposed another threshold-based technique for segmentation [10] in three subsets (background, edge and spot pixels), using a threshold that minimizes the total standard deviation of pixels above and below it. In 2009, Battiato and Rundo published a segmentation approach based on Cellular Neural Networks (CNNs) [11].

B.2 Transform

The transform stage consists of changing the image domain from the spatial domain to a domain where it can be more efficiently processed or coded. However, transform based compression is not typically as efficient for DNA microarray images as it is for other types of images not containing such sharp edges [12]. For this reason, transformations are not frequently researched in microarray image compression, although they are used in some works.

In 2004, Hua et al. [8] published a modification of the EBCOT algorithm, the basis of the JPEG2000 standard [13], that included a tailored integer odd-symmetric transform. In 2004, Lonardi and Luo [4] made use of the Burrows-Wheeler transform [14] for lossy or lossless compression in their MicroZip software. In 2012, we proposed a novel reversible point transform consisting in swapping the left and right halves of the image histogram [15].

TABLE I: Classification of lossless microarray-specific techniques discussed in Subsection I-B, sorted chronologically.

Preprocessing	**Transform**	**Entropy coding**	
Segmentation		Segmentation	Context
[7], 2002	[8], 2004	[6], 2003	[16], 2005
[6], 2003	[4], 2004	[9], 2006	[17], 2006
[8], 2004	[15], 2012	[11], 2009	[18], 2006
[4], 2004			[19], 2009
[9], 2006			
[10], 2007			
[11], 2009			

B.3 Entropy coding

In this stage, data obtained from previous stages are expressed in an efficient manner to produce a compact bitstream. Many techniques segment the image before compression, while others build contexts or try to predict the intensity of the next pixels based on the previous ones. Purely lossless techniques using each approach are described next.

At least three different coding proposals that are based on segmentation can be found on lossless

compression of DNA microarrays. In 2003, Faramarzpour et al. presented a prediction-based technique [6]. The image is gridded, and a linear prediction scheme is applied after creating a spiral path from the estimated spot center. In 2006, Bierman et al. presented their MACE (Micro Array Compression and Extraction) software [9]. The image is divided first using a threshold-based method. The low intensity pixels are coded using standard dictionary-based techniques, while the high intensity pixels are processed with a sparse matrix algorithm and then compressed. In 2009, Battiato and Rundo published an algorithm [11] based on image color reindexing after segmentation. Segmentation is made by means of a CNN-based system to produce two complementary subimages. The foreground image is compressed with a generic lossless algorithm and stored separately. The background image is first transformed into an indexed image. Then its color palette is reindexed with an algorithm that reduces the zero-order entropy of local differences, which are losslessly coded.

In no less than four publications, context building is used to perform lossless DNA microarray image compression. In 2005 and in 2006 Zhang et al. [17], [16] proposed a context-based lossless approach that also employs segmentation. Once the image is divided, a simple predictive scheme is used for the most significant bytes of each pixel, while the least significant bytes are coded using prediction by partial approximate matching (PPAM), also proposed by Zhang and Adjeroh [20]. In 2006, Neves and Pinho [18] proposed another context-based lossless approach. It is a bitplane-based technique that uses 3D finite-context models to drive an arithmetic coder. In 2009, they improved this scheme so that specific contexts are built for each image [19].

Table I presents a summary of all discussed methods classified attending to the stage of the image compression process in which they make their contribution.

C. Paper structure

The rest of this paper is organized as follows. In Section II, the correlation present among microarray images that belong to the same set is analyzed. In Section III, several multicomponent compression experiments are described and discussed. Finally, in Section IV, some conclusions are drawn.

II. Correlation between DNA microarray images

As it was pointed out in Section I, many DNA microarray studies consist of running several experiments at different moments on similar biological samples that have been exposed to different physical and chemical conditions. In addition, each microarray experiment produces two monochrome images which are obtained by scanning the same microarray chip. For this reason, it is natural to assume that some kind of correlation is present among the images produced within a study. Subsection II-A describes the microarray image benchmark sets studied in this paper and Subsection II-B analyzes the correlation present among the images of each set.

TABLE II: Image sets used in the literature.

Image set		Images	Size (px)
Yeast	[21]	109	1024×1024
ApoA1	[22]	32	1044×1041
ISREC	[23]	14	1000×1000
Arizona		6	4400×13800

A. Datasets

A number of different DNA microarray image sets have been used for benchmarking compression performance. No set has been used across all publications on DNA microarray image compression, but the MicroZip, ApoA1 and ISREC sets are employed more frequently. Unfortunately, images of the MicroZip set do not have the same size, so they have not been used in our correlation and multicomponent compression experiments. We have included the Yeast set, which has been employed in a few publications. Furthermore, we have gathered additional larger images, closer to what is employed in laboratories today. These images, which come from the Arizona set, have been kindly provided by David Galbraith and Megan Sweeney from the University of Arizona.

Table II shows key properties of all sets documented in the literature. All images are monochrome, unsigned, 16 bits per pixel (bpp), and contain a single component per red/green channel.

B. Image correlation

In this subsection, the correlation present among images of the same set is analyzed. Two configurations for this experiment have been tested. First, all possible pairs of images from the same set are considered. After that, only red and green channels obtained from the same microarray chip are analyzed. The Pearson product-moment correlation coefficient is the correlation metric used in our experiments. It takes values in $[-1, 1]$ and is defined as

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}},$$

where X_i and Y_i are the sequences of pixels obtained by scanning two images in the same order, and $\bar{X}$ and $\bar{Y}$ are the average pixel values of the first and second images, respectively.

Figures 2 and 3 show the distribution of the measured Pearson's r values among all pairs and red/green channel pairs, respectively. In these figures it can be observed which ranges of Pearson's r values are more common. Additionally, Tables III and IV display statistical information about each of

the experiment configurations. It can be easily observed that, in general, pairs of images of the same set are not very correlated: most Pearson's r values are under 0.2 except for the Arizona set, which shows larger measurements, but most of them are still below 0.8. On the other hand, when considering only red/green channel pairs from the same chip, correlation is considerably larger. Attending to the average Pearson's r values of the sets, tenfold increases can be observed, while variance is reduced. Comparing the *Max* column in Tables III and IV, we see that the largest correlation value in each set corresponds to a red/green channel pair.

These results show that a certain amount of correlation actually exists among images from the same data set, to a larger extent when considering only red/green channel image pairs, for which average values larger than 0.75 are consistently observed for all sets.

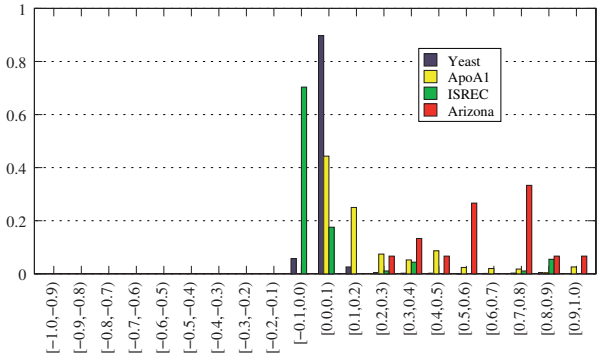


Fig. 2: Pearson's r value distribution for all image pairs from the same chip. The sum of the bars referring to each set is one.

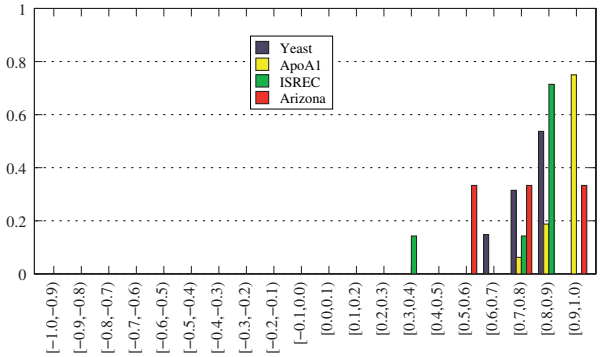


Fig. 3: Pearson's r value distribution for red/green channel image pairs from the same set.

III. Multicomponent compression of DNA microarray images

Attending to the results discussed in Section II, it seems reasonable to exploit the correlation present among pairs of images from the same set, especially among red/green channel image pairs. To do so, we have designed several experiments using lossless JPEG2000, since it is the only progressive lossy-to-lossless scheme supported by the DICOM medical image standard.

TABLE III: Statistical properties of the Pearson's r values obtained comparing all pairs of images in each set. The sum of the bars referring to each set is one.

Set	Max	Min	Average	Variance
ISREC	0.8901	-0.0186	0.0735	0.0501
ApoA1	0.9678	0.0105	0.2043	0.0468
Yeast	0.8961	-0.0329	0.0390	0.0070
Arizona	0.9332	0.2845	0.6153	0.0383

TABLE IV: Statistical properties of the Pearson's r values obtained comparing red/green channel pairs of images in each set.

Set	Max	Min	Average	Variance
ISREC	0.8901	0.3250	0.7822	0.0359
ApoA1	0.9678	0.7969	0.9229	0.0029
Yeast	0.8961	0.6084	0.7821	0.0048
Arizona	0.9332	0.5937	0.7508	0.0195

In these experiments, we have benchmarked the lossless multicomponent compression performance of Kakadu JPEG2000 when applied to DNA microarray images using different spatial and spectral transforms. For the spatial transformations, we have tested different numbers of 5/3 DWT decomposition levels. For the spectral transform, we have also used different numbers of 5/3 DWT decomposition levels as well as the reversible Haar transform (R-Haar), differential pulse code modulation (DPCM) and the pairwise orthogonal transform (POT [24]). Table V shows the average compression performance expressed in bits per pixel when compressing all images of a set together as a single multicomponent image. The order in which the images are arranged for the multicomponent compression affects the performance only to a little extent, generally less than 0.5 bpp. In this table, we show results only for the $R_1G_1 \cdots R_NG_N$ arrangement, where R_i and G_i are the red and green channel images from the i-th pair of a set, respectively. Table VI shows the average results for compressing red/green channels together as a 2-component image. For brevity's sake, these tables display only a representative set of results from all the data obtained in our experiments. A full description of the experiment with data for all tested configurations can be downloaded at `http://deic.uab.es/~mhernandez/media/reports/multicomponent_compression.pdf`. Table VII of this document displays compression performance results of other lossless schemes, including the best-performing microarray-specific technique [11], for ease of reference.

It can be observed that one level of DWT spectral transform does not generally improve compression performance, as compared to using zero wavelet decomposition levels. When it does, the gain does not exceed 0.3 bpp. The POT shows slightly better

results compared to using zero DWT levels for the ApoA1 set, with improvements up to 0.6 bpp. For the Yeast and Arizona sets, the DPCM transform produces improvements of up to 0.4 bpp, better than any of the other transforms. For the ISREC set, no transform is able to improve upon zero DWT decomposition levels. It is also noteworthy that all spectral transforms work better when considering only red/green channel image pairs that when all images in the set are compressed as a single multicomponent image. This is consistent to the results discussed in Section II because red/green channel image pairs exhibit larger amounts of correlation.

IV. Conclusions and future work

In this paper, we have described DNA microarray images and we have motivated the importance of their compression. We have briefly described the state of the art of the lossless compression of this type of imagery. We have analyzed the correlation present on pairs of microarray images from the same set using Pearson's r as a metric. From that we have concluded that there exists a certain amount of correlation among image pairs, especially among green/red channel image pairs. Based on these results, several lossless multicomponent compression tests have been run and described. The DWT 5/3, DCPM, the reversible Haar transform and the POT have been employed using several different configurations in the experiments. No single spectral transform is able to improve upon zero wavelet decomposition levels in the spatial and spectral domain for all the sets, even though the DPCM transform does so except for the ISREC set. For all sets and transforms, the observed compression performance is better when considering only red/green channel image pairs. This is consistent to the correlation values observed.

As future work, we plan to quantitatively analyze the relationship between correlation and multicomponent compression. In addition, we also plan to apply the histogram swap transform [15] together with multicomponent compression.

Acknowledgements

This work has been partially supported by the European Union, by the Spanish Government (MECD) and by the Catalan Government, under Grants TIN2009-14426-C02-01, SGR2009-1224, TEC2010-11776-E, FPU AP2010-0172, China Scholarship Council, UAB-BI3INT2006-08, UAB-472-01-2/09, RYC-2010-05671 and FP7-PEOPLE-2009-IIF FP7-250420.

References

[1] S. Moore, "Making chips to probe genes," *IEEE SPECTRUM*, vol. 38, no. 3, pp. 54–60, MAR 2001.

[2] S. Satih, N. Chalabi, N. Rabiau, R. Bosviel, L. Fontana, Y.-J. Bignon, and D. J. Bernard-Gallon, "Gene Expression Profiling of Breast Cancer Cell Lines in Response to Soy Isoflavones Using a Pangenomic Microarray Approach," *OMICS-A JOURNAL OF INTEGRATIVE BIOLOGY*, vol. 14, no. 3, pp. 231–238, JUN 2010.

[3] M. S. Giri, M. Nebozhyn, L. Showe, and L. J. Montaner, "Microarray data on gene modulation by HIV-1 in immune cells: 2000-2006," *JOURNAL OF LEUKOCYTE BIOLOGY*, vol. 80, no. 5, pp. 1031–1043, NOV 2006.

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

[5] M. Hernández-Cabronero, I. Blanes, M. W. Marcellin, and J. Serra-Sagristà, "Standard and specific compression techniques for DNA microarray images," *MDPI Algorithms*, vol. 4, pp. 30–49, 2012.

[6] N. Faramarzpour, S. Shirani, and J. Bondy, "Lossless DNA microarray image compression," in *In Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers*, vol. 2, November 2003, pp. 1501–1504.

[7] R. Jornsten, Y. Vardi, and C. Zhang, "On the bitplane compression of microarray images," in *In Proceedings of the 4th International Conference on Statistical Data Analysis Based on the L1-Norm and Related Methods*, 2002.

[8] 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.

[9] R. Bierman, N. Maniyar, C. Parsons, and R. Singh, "MACE: lossless compression and analysis of microarray images," in *In Proceedings of the ACM symposium on Applied computing*, ser. SAC '06. ACM, 2006, pp. 167–172.

[10] A. Neekabadi, S. Samavi, S. A. Razavi, N. Karimi, and S. Shirani, "Lossless microarray image compression using region based predictors," in *In Proceedings of the International Conference on Image Processing*, vol. 1-7. IEEE, 2007, Proceedings Paper, pp. 913–916.

[11] S. Battiato and F. Rundo, "A bio-inspired CNN with re-indexing engine for lossless dna microarray compression and segmentation," in *In Proceedings of the 16th International Conference on Image Processing*, vol. 1-6, IEEE. IEEE, 2009, Proceedings Paper, pp. 1717–1720.

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

[13] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Kluwer Academic Publishers, Boston, 2002.

[14] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm." HP, Tech. Rep. 124, 1994.

[15] M. Hernández-Cabronero, J. Muñoz-Gómez, I. Blanes, J. Serra-Sagristà, and M. W. Marcellin, "DNA microarray image coding," in *In Proceedings of the IEEE International Conference on Data Compression, DCC*, IEEE, Ed., 2012.

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

[17] D. A. Adjeroh, Y. Zhang, and R. Parthe, "On denoising and compression of DNA microarray images," *Pattern Recognition*, vol. 39, no. 12, pp. 2478–2493, December 2006.

[18] A. J. R. Neves and A. J. Pinho, "Lossless compression of microarray images," in *In Proceedings of the International Conference on Image Processing, ICIP*. IEEE, 2006, Proceedings Paper, pp. 2505–2508.

[19] ——, "Lossless compression of microarray images using image-dependent finite-context models," *IEEE Transactions on Medical Imaging*, vol. 28, no. 2, pp. 194–201, February 2009.

[20] Y. Zhang and D. Adjeroh, "Prediction by partial approximate matching for lossless image compression," *IEEE Transactions on Image Processing*, vol. 17, no. 6, pp. 924 –935, June 2008.

[21] "Stanford Yeast Cell-Cycle Regulation Project: (http://genome-www.stanford.edu/cellcycle/data/rawdata/individual.html)."

[22] "ApoA1 experiment data: (http://www.stat.berkeley.edu/users/terry/zarray/Html/apodata.html)."

[23] "ISREC image set: (http://www.isrec.isb-sib.ch/DEA/module8/P5_chip_image/images/)."

[24] I. Blanes and J. Serra-Sagristà, "Pairwise orthogonal transform for spectral image coding," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 3, pp. 961–972, March 2011.

TABLE V: Average lossless multicomponent compression results in bpp considering all images of each set as a single multicomponent image. Zero spatial DWT decomposition levels are assumed. In every case, using more than zero levels of spatial DWT yields worse performance.

Set	Spectral transform	Spectral levels	Bpp
Yeast	DWT 5/3	0	6.828
Yeast	DWT 5/3	1	8.877
Yeast	POT	1	9.883
Yeast	DPCM	1	6.447
Yeast	R-Haar	1	6.999
ApoA1	DWT 5/3	0	11.524
ApoA1	DWT 5/3	1	11.417
ApoA1	POT	1	11.267
ApoA1	DPCM	1	11.260
ApoA1	R-Haar	1	11.160
ISREC	DWT 5/3	0	10.887
ISREC	DWT 5/3	1	11.932
ISREC	POT	1	11.864
ISREC	DPCM	1	11.870
ISREC	R-Haar	1	11.451
Arizona	DWT 5/3	0	9.548
Arizona	DWT 5/3	1	9.804
Arizona	POT	1	9.556
Arizona	DPCM	1	9.575
Arizona	R-Haar	1	9.629

TABLE VI: Average lossless multicomponent compression results in bpp for red/green channel image pairs. Zero spatial DWT decomposition levels are assumed. In every case, using more than zero levels of spatial DWT yields worse performance.

Set	Spectral transform	Spectral levels	Bpp
Yeast	DWT 5/3	0	6.829
Yeast	DWT 5/3	1	6.786
Yeast	POT	1	9.279
Yeast	DPCM	1	6.439
Yeast	R-Haar	1	6.790
ApoA1	DWT 5/3	0	11.524
ApoA1	DWT 5/3	1	11.217
ApoA1	POT	1	10.956
ApoA1	DPCM	1	11.289
ApoA1	R-Haar	1	11.218
ISREC	DWT 5/3	0	10.887
ISREC	DWT 5/3	1	11.451
ISREC	POT	1	11.468
ISREC	DPCM	1	11.203
ISREC	R-Haar	1	11.452
Arizona	DWT 5/3	0	9.548
Arizona	DWT 5/3	1	9.649
Arizona	POT	1	9.439
Arizona	DPCM	1	9.386
Arizona	R-Haar	1	9.649

TABLE VII: Average compression results in bpp for generic image compressors and individual image compression. Results for the best microarray-specific technique (by Battiato) and the best general compressor are also included at the bottom for ease of reference.

Algorithm	MicroZip	Yeast	ApoA1	ISREC	Arizona
CALIC	9.582	8.502	10.515	10.615	–
JBIG	9.747	6.888	10.852	10.925	8.858
JPEG-LS	9.441	8.580	10.608	11.145	8.646
JPEG2000 (0 DWT 5/3 levels)	10.063	6.863	11.566	10.930	9.582
JPEG2000 (1 DWT 5/3 level)	9.577	9.128	11.134	11.517	9.253
JPEG2000 (5 DWT 5/3 levels)	9.508	9.082	11.052	11.360	9.099
Battiato [11]	8.369	–	9.52	9.49	–
Bzip2	9.841	6.075	11.067	10.921	8.944

3.3 Result comparison

For ease of reference, lossless compression results for the best-performing previously published algorithms and for the methods proposed in Sections 3.1 and 3.2 are provided in Table 3.1 and discussed next. Results for the Omnibus and IBB corpora, included in our experiments after the publication of the publications in these sections, and for the HEVC/H.265 compression standard are also considered for completeness.

It can be observed that JPEG2000 with the proposed HST is the best-performing generic or standard algorithm for three of the corpora with differences between 0.16 bpp and 0.69 bpp. For the other image sets, JPEG2000 with the HST also produces competitive bitrates, between 0.01 bpp and 0.27 bpp larger than the best-performing generic or standard algorithm. Even though the spectral decorrelation techniques from Section 3.2 can improve the best average results for JPEG2000 for 2 of the 6 in which they can be applied, the HST improves upon the best spectral decorrelation techniques and can be employed for all methods. Hence, JPEG2000 with the HST is arguably better than any other generic or standard algorithm.

When non-standard microarray-specific are also considered, the methods by Battiato and Rundo [20] and by Neves and Pinho [21] yield the overall best results of all coders. Significantly better results are reported in [20] for the ApoA1, ISREC and MicroZip sets. For this reason, the only microarray-specific algorithm for which data are reported in Sections 3.1 and 3.2 is [20]. Notwithstanding, the author has not been able to replicate these results nor to obtain an implementation of this algorithm from the authors. On the other hand, all results reported in [21] could be replicated and extended to other corpora. If [20] is not considered, [21] produces the most competitive compression results for 7 of the 8 corpora. JPEG2000 with the HST is only up to 0.6 bpp behind. For the Omnibus set, all standard compression algorithms –including JPEG2000 with the HST– are able to improve upon [21]. Therefore, JPEG2000 with the HST –with all advantages of standard algorithms– can be considered a competitive alternative to microarray-specific techniques.

Although not shown here, combining the HST and the proposed spectral decorrelation techniques does not yield better compression results. Furthermore, none of

these two approaches improves the results yielded by the best-performing microarray-specific algorithm.

Table 3.1: Average lossless compression results in bits per pixel gathered for this thesis. Results for standard image compressors include JPEG2000 using the best number of spatial decomposition levels, the results for the HST (in the *JPEG2000 (HST)* row, as reported in [26]) and for the best spectral decorrelation transform (in the *JPEG2000 (MCT)* row, as reported in [27]). Note that spectral decorrelation results could not be computed for some of the sests because they do not consist of pairs of images generated in the same experiment. Results for microarray-specific techniques include results as reported in the original papers except for [21], for which data were gathered using the author's codec width default parameters. Best results for each category are highlighted in bold font.

Algorithm	**Yeast**	**ApoA1**	**ISREC**	**Stanford**	**MicroZip**	**Omnibus**	**AZ**	**IBB**
Generic algorithms and standard image compressors								
Bzip2	6.075	11.067	10.921	7.867	9.394	7.523	8.944	9.081
CALIC	8.502	**10.515**	**10.615**	7.592	9.582	**6.929**	8.767	9.327
JBIG2	6.888	10.852	10.925	7.776	9.747	7.198	8.858	9.344
JPEG-LS	8.580	10.608	11.145	**7.571**	9.441	6.952	**8.646**	9.904
HEVC/H.265	10.660	14.482	14.876	8.897	11.179	8.350	10.592	12.262
JPEG2000	6.829	10.999	10.888	7.969	9.467	8.121	7.549	9.064
JPEG2000 (HST)	**5.911**	10.786	10.624	7.685	**9.157**	7.103	8.795	**8.392**
JPEG2000 (MCT)	6.439	10.956	11.203	N/A	N/A	N/A	9.386	9.602
Microarray-specific techniques								
[14] SLOCO	8.556	—	—	—	—	—	—	—
[13] Faramarzpour	9.091	—	—	—	—	—	—	—
[16] Hua	6.985	—	—	—	—	—	—	—
[15] MicroZip	—	—	—	—	9.843	—	—	—
[17] PPAM	6.601	—	—	—	9.587	—	—	—
[19] Neekabadi	—	10.250	10.202	—	8.856	—	—	—
[20] Battiato	—	**9.520**	**9.490**	—	**8.369**	—	—	—
[21] Neves	**5.521**	10.223	10.199	**7.335**	8.667	**7.743**	**8.275**	**8.039**

Chapter 4

Lossy Compression

The lossy compression of DNA microarray images can produce arbitrary compression ratios. This is in stark contrast to the limited performance of lossless coders, described in Chapter 3. However, the lossy compression of microarray images has received less attention than the lossless compression. The main reason for this is the fact that making changes in the images generally introduces variations in the results of the analysis process. If the introduced variations are too large, the images become unusable for practical purposes. On the other hand, data extracted from DNA microarrays is inherently subject to experimental variability. In theory, if a given microarray experiment is replicated under the same conditions, the results should be identical. In practice, they differ. Thus, analysis result distortions smaller than the inherent experimental variability are considered acceptable [14, 16, 25]. In particular, acceptable distortion results can be yielded by lossy compression methods [25]. Hence, the research of lossy compression methods for DNA microarray images is justified as long as the impact on the analysis results is assessed.

As discussed in Chapter 1, microarray images present thousands of round regions called *spots*. Each of these spots contains information about a single gene of the species being tested in the experiment. When a pair of images are analyzed, they are first gridded into regions, each of which contains exactly one spot. Each region of one image is then compared to the co-located region in the other image. Note that the spot contained in both regions is associated to the same gene. As a result of these

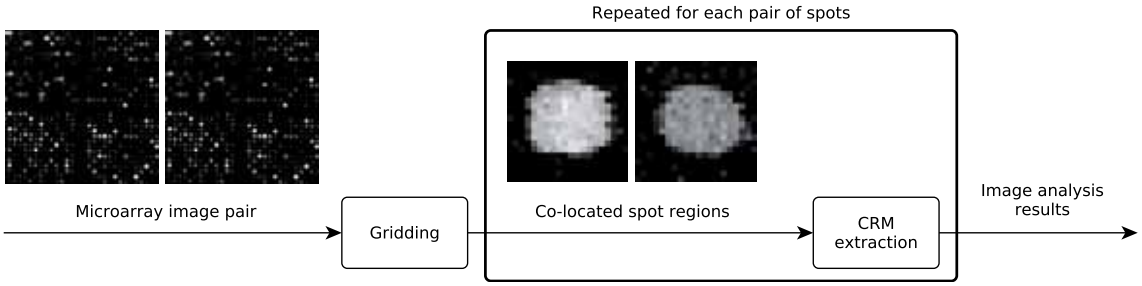


Figure 4.1: Diagram of the extraction of the corrected ratio of means (CRM) values.

comparisons, a numerical value –the *corrected ratio of means* (CRM)– is produced for each spot positively detected by the griding algorithm. These CRM values constitute the main output of microarray image analysis techniques. A diagram of the CRM extraction process is shown in Fig. 4.1.

The main contributions of this thesis to the lossy compression of DNA microarray images are the following:

- In the literature, the main way of assessing the impact on the analysis results is by comparing the CRM values extracted from the original and the distorted images. The most common way to achieve this is by analyzing plots of the dispersion of the distorted results, as compared to the original results [14, 16, 25]. Measures based in the mean absolute error (MAB) and the mean squared error (MSE) such as

$$\frac{1}{N}\sum_{i=1}^{N}\left|\log(\mathrm{CRM}_i)-\log(\widehat{\mathrm{CRM}}_i)\right| \tag{4.1}$$

and

$$\frac{1}{N}\sum_{i=1}^{N}\left(\log(\mathrm{CRM}_i)-\log(\widehat{\mathrm{CRM}}_i)\right)^2 \tag{4.2}$$

have also been employed for this purpose [16]. Here, CRM_i and $\widehat{\mathrm{CRM}}_i$ are the CRM values associated to the i-th spot in the original and the distorted images, respectively. Depending on the application of the DNA microarray experiment, the extracted CRM values are subject to different classification algorithms. The

disagreement introduced in these classifications by the modification of the images has also been employed to assess the impact on the analysis process [16, 25]. Even though these methods give very precise information about the distortion introduced in the analysis results, they all require that the original and distorted images are analyzed so that the CRM values can be compared. Furthermore, if the distortion assessment is based on some classification process, it has to be carried out for original and distorted CRM values. Therefore, these types of distortion metric may not be useful for lossy compression methods such as JPEG2000, which requires to evaluate repetitively the distortion metric to be minimized. A main contribution of this thesis is the definition of a distortion metric which does not require a full analysis of the original and distorted images. Instead, simulations suggest that the impact on the analysis process can be predicted by considering only pixel values of the original and distorted images. The proposed metric –the microarray distortion metric (MDM)– appeared first in [30] and it is fully described in Section 4.1.

- Before the beginning of this thesis, several methods for lossy or lossy-to-lossless compression methods for DNA microarray images had been published. In 2003, Jörnsten *et al.* proposed a modification of the near-lossless mode of the LOCO-I algorithm [43], basis for the JPEG-LS standard [44]. The modification consists in segmenting the image and coding spot and background pixels separately, as described in Chapter 3. Once the image is coded allowing a maximum absolute error of δ, their lossy-to-lossless scheme allows the coding of successive refinement data until $\delta = 0$, equivalent to lossless compression. In 2004, Hua *et al.* proposed a similar approach in which the EBCOT algorithm (basis for the JPEG2000 standard [29]) is adapted to better code the segmented microarray images. Also in 2004, Lonardi proposed an algorithm based on segmentation in which the 8 least significant bits of the pixels tagged as background are coded with a generic lossy image codec [15]. In their implementation, the SPIHT [45] algorithm is employed. In 2007, Peters *et al.* presented a lossy compression method [22] based on a direct application of the singular value decomposition

(SVD) [22]. Finally, in 2011, Avanaki *et al.* directly applied fractal and wavelet-fractal lossy compression methods [23]. As can be observed, all these works are based on generic image compressors not specific for DNA microarray images. However, generic compressors are designed to minimize the visual impact on the images and not the impact on the analysis process that is applied to them. Hence, the rate-distortion results (considering the distortion of the analysis results) are unlikely to be optimal. The second main contribution of this thesis to the lossy compression of microarray images is the proposal of a lossy compression scheme designed to minimize the impact on the analysis process. The proposed method is based on a novel quantization scheme –the relative quantizer (RQ)– that limits the maximum *relative* error introduced in the image and devotes additional precision to low-intensity pixels. The rate-distortion results yielded by this method are significantly superior than those produced by generic image compressors used in the aforementioned publications. A manuscript describing this proposal is under revision at the IEEE Transactions on Medical Imaging, and is reproduced in Section 4.2.

- In spite of the competitive results of the aforementioned RQ coder, the images compressed with this algorithm can only be reconstructed with a fixed quality, determined by the parameter k. Therefore, if researchers need to access or transmit the DNA microarray images with different qualities, several compressed versions of the images need be stored. Clearly, this approach multiplies the space and transmission time requirements and is not practical. In the last contribution, this problem is solved. The regular structure of the quantizer intervals for different values of k is exploited to create a progressive lossy-to-lossless scheme based on [21]. This scheme is hereinafter referred to as the Progressive Relative Quantizer (PRQ). The original context modeling system of that work has been improved and support for a region of interest (ROI) containing only spots has been added. The resulting coder is referred to as the PRQ-ROI coder. With these enhancements, the rate distortion results of the original RQ are significantly improved without diminishing the lossless compres-

sion results. That is, the PRQ introduces smaller distortion in the microarray image analysis than the RQ for equally large compressed file sizes. A paper describing this contribution to the state-of-the art on the compression of DNA microarray images has been submitted to the IEEE Signal Processing Letters and is reproduced in Section 4.3.

4.1 Microarray Distortion Metric

```
@InProceedings{Hernandez13DCC,
               Title     = {{A distortion metric for the lossy compression of DNA microarray images}},
               Author    = {Miguel Hern{\'a}ndez-Cabronero and Victor Sanchez
                            and Michael W. Marcellin and Joan Serra-Sagrist{\`a}},
               Booktitle = {Proceedings of the IEEE International Data Compression Conference, DCC},
               Year      = {2013},
               Editor    = {IEEE},
               Pages     = {171--180},
               ISBN      = {9781467307154}
               doi       = {10.1109/DCC.2013.26}
}
```

2013 Data Compression Conference

A distortion metric for the lossy compression of DNA microarray images

Miguel Hernández-Cabronero[†], Victor Sanchez[*], Michael W. Marcellin[‡] and Joan Serra-Sagristà[†]

[†] Department of Information and Communications Engineering,
Universitat Autònoma de Barcelona, Barcelona, Spain.

[*] Department of Computer Science,
University of Warwick, Coventry, UK.

[‡] Department of Electrical and Computer Engineering
University of Arizona, Tucson, AZ, USA.

Abstract

DNA microarrays are state-of-the-art tools in biological and medical research. In this work, we discuss the suitability of lossy compression for DNA microarray images and highlight the necessity for a distortion metric to assess the loss of relevant information. We also propose one possible metric that considers the basic image features employed by most DNA microarray analysis techniques. Experimental results indicate that the proposed metric can identify and differentiate important and unimportant changes in DNA microarray images.

I. Introduction

In this section, we motivate the need for introducing a distortion metric to evaluate the performance of lossy compression on DNA microarray images.

A. DNA microarrays

DNA microarrays are used to analyze the function and regulation of the genes of an organism [1]. They are state-of-the-art tools in biological and medical research, and are employed in many areas ranging from the study of metabolism [2] and evolution [3] to the fight against cancer [4], HIV [5] and malaria [6]. Interest in DNA microarrays has grown in the last few years, and an exponential increase of DNA microarray data has been observed [7].

In a DNA microarray experiment, the genetic expression level of two biological samples is compared. In these two samples –for instance one coming from a healthy tissue and the other coming from a tumoral tissue– the same genes might be expressed with different intensities. Analyzing these differences, it is possible to identify genes related to a particular biological process. A DNA microarray experiment consists of several steps, schematically summarized in Figure 1. The biological samples are first dyed with fluorescent markers, usually Cy3 and Cy5 of the cyanine family, and are then left to react on the surface of a DNA microarray chip (step A in Figure 1). The surface contains thousands of microscopic holes or *spots*, each of which is related to an individual gene. This surface is then washed so that each of the biological samples appears only inside the different spots. After that, the DNA microarray chip is scanned using two lasers, each exciting only one of the fluorescent markers, so that one 16 bpp monochrome image is produced for each of the two biological samples (step B in Figure 1). These two monochrome images are usually known as the green and red channels due to the color of the laser needed to excite the fluorescent dyes. The image intensity with which each spot is acquired is proportional to the amount of

1068-0314/13 $26.00 © 2013 IEEE
DOI 10.1109/DCC.2013.26

biological sample that is contained in the spot; that amount is also proportional to the expression intensity of the corresponding gene. Finally, these two images are analyzed to measure genetic expression intensity differences, which are then processed statistically to identify relevant genetic expression alterations (step C in Figure 1).

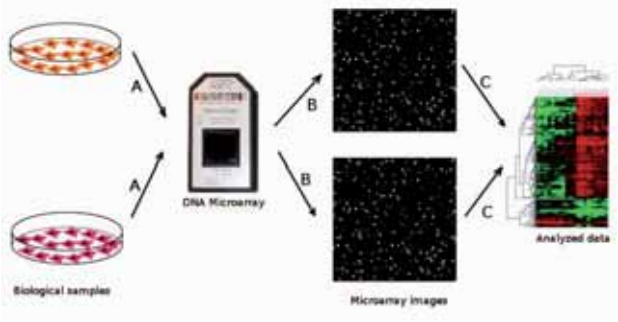


Figure 1: A DNA microarray experiment. The biological samples are first put on a microarray chip, then the chip is scanned to produce two microarray images and finally the images are processed to extract the genetic data.

B. DNA microarray image analysis

DNA microarray images are an intermediate product of DNA microarray experiments: image analysis is performed on these images to extract information about the genetic expression intensity. Unfortunately, these image analysis techniques are not fully mature or universally accepted [8], so they are likely to change in the future. As new image analysis techniques are developed, it will be highly desirable to reanalyze the images to obtain more accurate genetic data. However, in such cases, repeating the whole experiment is not an option because the needed biological samples are usually not available. For this reason, it is important to store the DNA microarray images along with the extracted genetic data.

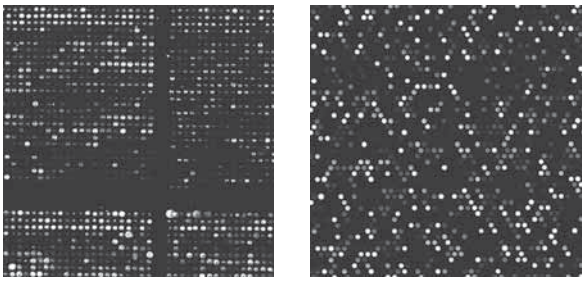
In spite of the increasing reproducibility of DNA microarray experiments, some variability is always present. When samples from the same two tissues are used in different experiments, the produced images, and thus the extracted genetic data, are not identical [9]. In addition, modern DNA microarray chips make use of biological replication, that is, they contain several spots associated to a single gene; even though theoretically these spots should express the same degree of gene activity, in practice they do not [8].

C. DNA microarray image lossless compression

In a DNA microarray experiment, two images –known as the green and red channels– are produced. Nowadays, these images easily exceed 4000×13000 pixels in size, with 16 bits per pixel (bpp) per channel, so that raw file sizes over 100 Megabytes per image are common. Besides, when a DNA microarray study is carried out, several DNA microarray experiments are performed. Thus, a considerable amount of data are produced in laboratories around the world. It is therefore paramount to design efficient storage and transmission methods for this type of images, and data compression comes as the best suited approach to this problem.

DNA microarray images possess several characteristics that make compression a challenging task. Figure 2 shows part of two example DNA microarray images. It can be observed that thousands of irregular round spots of varying intensities are displayed on a dark background. These abrupt and irregular intensity

changes produce high frequencies in the image which are hard to code or predict. Furthermore, the 16 bpp needed to represent DNA microarray images increase the data entropy due to the larger amount of possible pixel values. Altogether, these properties make the compression of DNA microarray images a challenging task, in particular when using lossless compression methods.



(a) (b)

Figure 2: Example DNA microarray images: 600×600 crops with different spot layouts. a) *array3* image from the MicroZip set with square grid spot layout; b) *slide_1-red* from the Arizona set with hexagonal grid spot layout. Gamma levels have been adjusted for visualization purposes.

Table I: Image sets used for benchmarking in the literature. All original images are 16 bpp.

Image set		Images	Size (pixels)
MicroZip	[10]	3	$> 1800\times1900$
Yeast	[11]	109	1024×1024
ApoA1	[12]	32	1044×1041
ISREC	[13]	14	1000×1000
Stanford	[14]	20	$> 2000\times2000$
Arizona	[15]	6	4400×13800

Several authors have proposed compression techniques for DNA microarray images in the last few years. A detailed review of the state of the art in the compression of DNA microarray images can be found in the literature [16]. Both lossy and lossless techniques have been proposed, but lossless techniques are more common. It has been argued that lossless compression is more suitable for DNA microarray images because it guarantees that no relevant information will be missing when reanalyzing the decompressed version of the images [17]. However, this data fidelity is obtained at the expense of poor compression performance results. Table I displays basic information about the sets of images used for benchmarking while Table II reports the lossless compression results yielded by both standard and the best performing microarray-specific techniques.

It can be observed that even the best-performing techniques specific for DNA microarray images are only approximately 1 bpp better than the best standard compression techniques. Moreover, compression ratios equal or better than 2:1 are only found for two of the datasets. Considering these lossless compression results, schemes with higher compression performance are needed.

D. Distortion metrics for DNA microarray image lossy compression

Lossy compression schemes can yield very good compression ratios and as long as the distortion introduced by the lossy compression of DNA microarray images falls below the variability of DNA

Table II: Average lossless compression results for different standard and microarray-specific techniques for the benchmark image sets. All values are expressed in bits per pixel and all images are 16 bpp. Bitrates for non-standard compression techniques are shown as reported by the original authors when data are available for a set. Results for JPEG2000 have been obtained with the best number of wavelet decomposition levels for each set [16].

Algorithm	**MicroZip**	**Yeast**	**ApoA1**	**ISREC**	**Stanford**	**Arizona**
Standard techniques						
Bzip2	9.394	6.075	11.067	10.921	7.867	8.944
JBIG	9.747	6.888	10.852	10.925	7.776	8.858
JPEG-LS	9.441	8.580	10.608	11.145	7.571	8.646
JPEG2000	9.508	6.863	11.050	10.930	8.007	9.099
JPEG2000 + HST [18]	9.157	5.911	10.786	10.624	7.685	8.795
Microarray-specific techniques						
MicroZip [19], 2004	9.843	—	—	—	—	—
PPAM [20], 2005	9.587	6.601	—	—	—	—
Neves [21], 2006	8.840	—	10.280	10.199	—	—
Neekabadi [22], 2007	8.856	—	10.250	10.202	—	—
Neves [17], 2009	8.619	—	10.194	10.158	—	—
Battiato [23], 2009	8.369	—	9.520	9.490	—	—

microarray experiments (Subsection I-B), lossy coding can be regarded as an alternative to lossless coding.

An important number of both standard [24], [25] and microarray-specific [9], [19], [26], [27] lossy or lossy-to-lossless compression techniques have been proposed and discussed in the literature. For lossy coding, it is necessary to assess the loss of relevant information by employing a distortion metric, since it can affect current and future DNA microarray analysis techniques.

To date, only full reference metrics –which need the original image to be calculated– have been considered. Existing metrics based on pixel-wise errors or visual fidelity are not suitable for this purpose [9] because they do not identify changes that could affect the subsequent analysis process. Metrics using pixel-wise errors like MSE, PSNR or SNR consider every pixel in the image equally important [28], hence they are unable to distinguish changes that significantly distort several spots from those that slightly affect unimportant parts of the background. Metrics based on visual fidelity like SSIM or CW-SSIM are designed to identify changes that can be detected by the human visual system [28]; however, DNA microarray images are always computer analyzed –most usually without human intervention– so changes applied to a spot can be difficult to recognize by the naked eye, but still affect greatly the analysis process.

At least two publications have addressed the problem of measuring the information loss due to the compression of DNA microarray images. In [25], the authors propose several methods to evaluate the image distortion based on comparing the results of different classification algorithms like linear discriminant analysis when applied to the extracted mean intensity of each spot. Their methods rely completely on the results of the selected image analysis and classification algorithms and do not explicitly consider image properties used in most analysis algorithms. For this reason, their results are subject to the particularities of the selected image analysis and data classification algorithms and it is unclear whether they would be applicable when other present and future DNA microarray analysis techniques are employed. Even though the authors provide intensive experimental results, they do not discuss which selection of intensity extraction and classification algorithms could be the best as a distortion metric. In [9], a similar approach

is used. Two different image analysis techniques are used on original and distorted DNA microarray images to calculate the spot intensities. After that, simple functions of the extracted intensities are plotted. These functions include the logarithm of the quotient of the intensities extracted for one spot in each of the two DNA microarray images, which is a measure of gene activity. The plots are analyzed to assess the variability using different intensity extraction and compression algorithms. As in [25], the authors do not explicitly use any commonly used image feature nor propose a way to combine the plotted values in order to construct a distortion metric.

E. Paper structure

The rest of this document is structured as follows. In Section II we describe the DNA microarray analysis pipeline and highlight the image features that are commonly used for this analysis. In Section III we propose a distortion metric that employs these image features to detect relevant changes in DNA microarray images. Finally, in Section IV we provide some concluding remarks.

II. Analysis pipeline of DNA microarrays

In this section, we describe the analysis of DNA microarrays. We also identify which processes are basic for any analysis process and what image features are considered in them.

A. Analysis pipeline

In Section I we briefly described how DNA microarray experiments are carried out, as schematically summarized in Figure 1. We now focus on the analysis pipeline that is performed once the DNA microarray images are obtained. The first two subsections describe how the images are analyzed while the third briefly explains subsequent procedures.

1) Gridding and segmentation: The analysis of DNA microarray images begins by locating where each spot is situated. As shown in Figure 1, spots are arranged following a regular grid which must be identified first. This process is known as *gridding* and can be done either automatically without any prior information [19], or using geometrical information provided by the DNA microarray manufacturer.

Once the grid is known, each spot is confined individually to a rectangular area. The next step is to determine which pixels belong to the spot, and which ones are background. This process is known as *segmentation* and is one of the most active research topics on the analysis of DNA microarray images [8]. Many different approaches have been used for this purpose, and it is possible to find clustering-based [29]–[33], threshold-based [34], graph-based [35] and even wavelet-based [36] proposals in the literature.

2) Intensity extraction: The next step in the analysis pipeline is calculating the expression intensity of each gene for each of the biological samples. As previously explained in Section I, the amount of dyed sample inside one spot is proportional to the intensity of the corresponding gene for that biological sample. For this reason, it is possible to estimate the genetic expression intensity by looking at the average value of pixels inside each spot in the DNA microarray images. This process is performed automatically by DNA microarray image analysis algorithms and is known as *intensity extraction* or *feature extraction*.

Unfortunately, in real DNA microarray experiments there are artifacts that distort the scanned images. When the DNA microarray chip is washed, some residues can remain on the surface. To address this and other artifacts, researchers have proposed several *background-correction* algorithms [37]. In their proposals the average intensity of the local background (the pixels outside but near each spot) is calculated and subtracted from the extracted intensity according to different algorithms. In addition, since the two images produced in a DNA microarray experiment are obtained using two different lasers, it is possible that one

of the two images produced in a DNA microarray experiment is globally brighter than the other. For this reason, some analysis techniques compute the global intensity to identify brightness changes and modify the extracted spot intensities accordingly.

3) Normalization and data analysis: After the spot intensities are extracted from the images associated with each of the two biological samples, they are compared to detect relevant differences in the genetic expression intensities. This is commonly know as *data analysis*. Before that, the intensities need to be further processed in order to remove any systematic variation, due for example to dye bias. This step –known as *normalization*– is an active research topic and several normalization techniques have been proposed [8]. Both data analysis and normalization techniques employ only the extracted spot intensities and do not directly consider the images, so they are out of the scope of this work.

B. Key image properties

Having in mind the processes described in Subsection II-A, it is possible to identify at least three key image features that can greatly affect the outcome of a DNA microarray experiment. Since it is our goal to design a distortion metric for DNA microarray images, we next describe these key features more fully and discuss changes that can most alter experimental outcomes.

1) Spot intensity: The value of the pixels inside spots is the most important DNA microarray image feature because it is employed to extract the genetic expression intensities that are employed in subsequent steps of DNA microarray experiments. Only average pixel values are used for this purpose, so it is more important to maintain these mean values unmodified than to achieve pixel-wise fidelity. If we call μ_R and μ_G to the average intensity of the co-located spot in each of the two image channels of an experiment, then the *mean intensity ratio* is defined as μ_R/μ_G. Later analysis steps [38] consider only the mean intensity ratio of the spots, so relative errors in the average intensity values are much more relevant than absolute errors.

2) Local background intensity: The mean intensity ratios are often corrected by subtracting the average intensity of the local background, that is, the pixels that are close but do not belong to a spot. Even though this subtraction can be done in different ways [37], the mean intensity of the local background is always used. Therefore, it is important to keep these local background mean intensity values unmodified.

3) Global intensity: Some normalization techniques rely on the global intensity of each DNA microarray image, i.e., the sum of all pixel intensities [39]. They are based on the assumption that this global intensity should be the same for the images corresponding to each biological sample. As a consequence, producing a deviation on the global intensity can affect the obtained mean intensity ratios proportionally.

To the best of our knowledge, these three features are the basis of every existing analysis technique to the extent of our knowledge, and they are likely to remain fundamental in future techniques as well.

III. Proposed distortion metric

In Section I, we motivated the compression of DNA microarray images and argued that traditional image distortion metrics like MSE or SSIM are not suitable for DNA microarray images. In this section, we propose a novel distortion metric for this type of imagery.

A. Metric definition

Our main goal is to define a distortion metric that is able to summarize information about the three main image features with primary effect on the analysis process. As will be shown in Subsection III-B, common natural image distortion metrics are not suitable for DNA microarray images because they do not

consider these important image features. Since the main purpose of our metric is to detect image changes that can affect present and future analysis, we have designed it to be as conservative as possible; we have also assumed that a segmentation of the image into individual spots and background is provided as input to the metric. Such segmentation may be produced using automatic methods or geometric information from the manufacturer.

In order to produce a compact output that can be more easily interpreted, we have defined our microarray distortion metric (MDM) after the well-known logarithmic metric PSNR:

$$\mathrm{MDM} = 10\log_{10}\frac{(\mathrm{max\_val})^2}{\mathrm{ME}}. \quad (1)$$

To calculate the "signal-to-noise ratio" of our metric, the maximum value in the image ($\mathrm{max\_val}$) is used as a measure of the signal, while the noise is estimated using our proposed microarray error (ME). The ME must be sensitive to relevant changes in any of the three main features than can affect the analysis process: the mean intensity ratio, the average intensity of the local background and the global image intensity. As explained in the previous section, relative errors in those features are much more important than absolute errors. To estimate the relative distortion in the mean intensity ratio of one spot, we do not need to process both the red and green channel images. If the average intensity inside one spot is multiplied by a factor q in the distorted image, it can be easily proved that the intensity ratio for that spot in the red and green channel images is multiplied by a factor smaller than $\bar{q} = max(q, 1/q)$. In order to detect changes that can affect analysis results even if they appear in isolated spots, we consider only the maximum value of $\bar{q}$ among all spots. A similar conservative reasoning can be applied to the estimations of the distortion of the local background and global intensity. For our metric, we employ the expressions in (2)-(4) to calculate the distortion of the three key image features:

$$r_{\mathrm{spot}} = \max(\mathrm{max\_spot\_ratio}, 1/\mathrm{min\_spot\_ratio}), \quad (2)$$

$$r_{\mathrm{localBG}} = \max(\mathrm{max\_localBG\_ratio}, 1/\mathrm{min\_localBG\_ratio}), \quad (3)$$

$$r_{\mathrm{global}} = \max(\mathrm{global\_intensity\_ratio}, 1/\mathrm{global\_intensity\_ratio}). \quad (4)$$

Ideally, the MDM should produce high signal-to-noise ratios when the image is not distorted enough to affect the analysis results. When such relevant distortions are introduced in any of the key image features, the MDM should decrease toward 0. To achieve this, the definition of ME is based on $\mathrm{max\_val}$ raised to p, a logistic function of r_{spot}, r_{localBG} and r_{global}:

$$p = 2/(1 + exp(-\alpha(r_{\mathrm{spot}} + r_{\mathrm{localBG}} + r_{\mathrm{global}} - 3))), \quad (5)$$

$$\mathrm{ME} = (\mathrm{max\_val})^p - \mathrm{max\_val} + \min(\mathrm{max\_val}, \mathrm{MSE}_{\mathrm{image}}). \quad (6)$$

As is obvious from (5), when r_{spot}, r_{localBG} and r_{global} are all close to 1 –a scenario of irrelevant or nonexistent distortions–, p is also close to 1. When distortions relevant to the analysis process are introduced, r_{spot}, r_{localBG} and r_{global} are increased and p approaches 2. In consequence, the first term of (6) varies between $\mathrm{max\_val}$ and $\mathrm{max\_val}^2$. The other terms in (6) are employed for normalization purposes so that the MDM outputs only meaningful values and is able to differentiate lossless and lossy compression by employing the global MSE of the image. The sensitivity of the MDM to changes in the three key image features can be adjusted through the arbitrary parameter α in (5), which controls the speed in which the signal-to-noise ratio is degraded. In our experiments, we have found $\alpha = 3$ to be a balanced choice. When smaller values of α are chosen, the MDM decreases too slowly when essential parts of the DNA microarray images are modified. When larger values are selected, the MDM is too sensitive and produces values close to 0 dB when the key image properties are only slightly modified.

Further experimentation using real compression and analysis scenarios is necessary to elucidate optimal values of α for realistic operating ranges.

B. Experimental results

In this subsection, we illustrate the behavior of our proposed metric and provide further evidence of its suitability for the assessment of the information loss in DNA microarray images.

In our experiments, we have distorted three images from each set shown in Table I in three different manners: we have modified the pixels inside each spot, the pixels inside the local background of each spot and the pixels not inside any spot or local background. In this experiment, we have not modified the images by applying lossy compression so that changes in the images can be more easily located and understood. To identify spot and local background areas, we have employed a Matlab implementation of the circular Hough transform [40], whose results have been further refined to obtain an accurate list of circle centers and radii describing the spots. These results are used to calculate the three types of distortions as well as the segmentation required as input to our proposed MDM. Figure 3a shows the results for the MDM and the PSNR metrics when pixels inside all spots are multiplied by different coefficients. For each spot, we define the *spot ratio distortion* as μ_d/μ_o, where μ_d and μ_o are the mean intensity for that spot in the distorted and original images, respectively. Because of the definition in (2), spot ratio distortions in $(0, 1)$ are substituted by their inverse without loss of generality. In this figure, the horizontal axis represents the average spot ratio distortion. In Figure 3b, we show the results for MDM and PSNR when only the pixels inside local background of each spot are modified. In this figure, the horizontal axis represents the average local background ratio distortion. In Figure 3c, we show the results for applying zero-mean additive white Gaussian noise to pixels outside spots and local backgrounds as a function of noise variance. All results shown in Figure 3 are for the ApoA1 set only. Results for other image sets are similar.

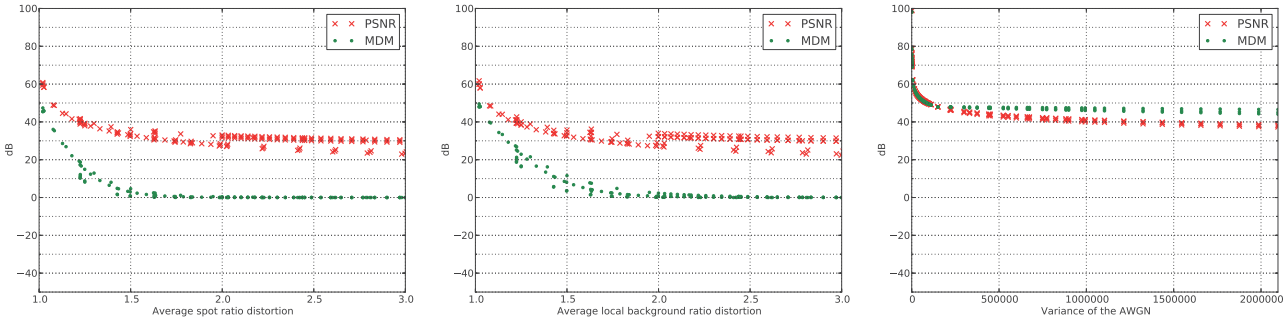
It can be observed that when features that are relevant to the DNA microarray analysis process are modified (Figures 3a and 3b), the MDM decreases rapidly toward 0 dB. The slope in which the MDM decreases when spot ratios are modified can be controlled by the α parameter in (5). Comparing Figures 3a and 3b, it can be seen that the slope is steeper in the former. This is due to the fact that there are more pixels inside spots than inside local backgrounds and they have larger average intensity, so r_{spot} in (5) grows faster. When unimportant changes are applied to the images (Figure 3c), the MDM decreases very slowly. In this case, if the global intensity of the image is not modified, the MDM remains constant at approximately 48 dB, independently of the actual MSE of the image.

These results suggest that the proposed MDM is able to detect changes in DNA microarray images that affect image analysis, whereas unimportant changes do not affect the output of the MDM.

IV. Conclusions and future work

DNA microarray images are broadly employed in biological and medical research to analyze the function of the genes of many different organisms. Large image file sizes motivate the use of coding techniques to help with storage and transmission, but lossless compression has not proved to be effective. On the other hand, lossy compression can provide more compression, but it is necessary to assess whether present and future analysis techniques are affected by the information loss. Distortion metrics like PSNR or SSIM are not suitable for this purpose, so microarray-specific metrics are needed.

The analysis pipeline of DNA microarrays has been discussed, and three key image features have been identified that are the foundation of most current analysis techniques, and very likely for future techniques as well. The identified features are the mean intensity of each spot, of each local background, and the overall intensity of the image. Based on these features, one possible microarray-specific metric has been



(a) (b) (c)

Figure 3: Distortion results for PSNR and our proposed MDM, when applied to three sample images from the ApoA1 set. a) Results after modifying pixels inside spots; b) Results after modifying pixels inside the spot local backgrounds; c) Results after applying additive white Gaussian noise to pixels outside spots and local backgrounds.

proposed, and evidence of its suitability to assess the information loss of DNA microarray images has been provided.

Our future work involves assessing the suitability of the proposed metric to detect changes in the output of standard DNA microarray analysis techniques when different types of distortion are introduced, including the distortion produced by lossy compression.

ACKNOWLEDGEMENTS

MicroZip corpus was kindly provided by Neves and Pinho from the University of Aveiro. The Arizona image set was provided by David Galbraith and Megan Sweeney from the University of Arizona. This work has been partially funded by the Spanish Ministry of Economy and Competitiveness (MINECO) and the Catalan Government under projects FPU AP2010-0172, TIN2009-14426-C02-01, TIN2012-38102-C03-03 (LIFE-VISION) and 2009-SGR-1224.

REFERENCES

[1] S. Moore, "Making chips to probe genes," *IEEE SPECTRUM*, vol. 38, no. 3, pp. 54–60, MAR 2001.

[2] M. Zaparty, A. Zaigler, C. Stamme, J. Soppa, R. Hensel, and B. Siebers, "DNA Microarray Analysis of Central Carbohydrate Metabolism: Glycolytic/Gluconeogenic Carbon Switch in the Hyperthermophilic Crenarchaeum Thermoproteus tenax," *Journal of Bacteriology*, vol. 190, no. 6, pp. 2231–2238, March 15, 2008.

[3] H. Maughan, C. W. Birky, and W. L. Nicholson, "Transcriptome divergence and the loss of plasticity in bacillus subtilis after 6,000 generations of evolution under relaxed selection for sporulation," *Journal of Bacteriology*, vol. 191, no. 1, pp. 428–433, January 1, 2009.

[4] S. Satih, N. Chalabi, N. Rabiau, R. Bosviel, L. Fontana, Y.-J. Bignon, and D. J. Bernard-Gallon, "Gene Expression Profiling of Breast Cancer Cell Lines in Response to Soy Isoflavones Using a Pangenomic Microarray Approach," *OMICS-A JOURNAL OF INTEGRATIVE BIOLOGY*, vol. 14, no. 3, pp. 231–238, JUN 2010.

[5] M. S. Giri, M. Nebozhyn, L. Showe, and L. J. Montaner, "Microarray data on gene modulation by HIV-1 in immune cells: 2000-2006," *JOURNAL OF LEUKOCYTE BIOLOGY*, vol. 80, no. 5, pp. 1031–1043, NOV 2006.

[6] Z. Bozdech, S. Mok, and A. P. Gupta, "DNA Microarray-Based Genome-Wide Analyses of Plasmodium Parasites," in *Malaria*, ser. Methods in Molecular Biology, R. Ménard, Ed. Humana Press, 2013, vol. 923, pp. 189–211.

[7] D. A. Adjeroh, Y. Zhang, and R. Parthe, "On denoising and compression of DNA microarray images," *Pattern Recognition*, vol. 39, no. 12, pp. 2478–2493, December 2006.

[8] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus." *Nature reviews. Genetics*, vol. 7, no. 1, pp. 55–65, Jan. 2006.

[9] R. Jornsten, 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.

[10] "MicroZip test image set (http://www.cs.ucr.edu/~yuluo/MicroZip)."
[11] "Stanford Yeast Cell-Cycle Regulation Project (http://genome-www.stanford.edu/cellcycle/data/rawdata/individual.html)."
[12] "ApoA1 experiment data (http://www.stat.berkeley.edu/users/terry/zarray/Html/apodata.html)."
[13] "ISREC image set (http://www.isrec.isb-sib.ch/DEA/module8/P5_chip_image/images/)."
[14] "Stanford Microarray Database public FTP (ftp://smd-ftp.stanford.edu/pub/smd/transfers/Jenny)."
[15] "Arizona test image set (http://deic.uab.es/~mhernandez/materials)."
[16] M. Hernández-Cabronero, I. Blanes, M. W. Marcellin, and J. Serra-Sagristà, "Standard and specific compression techniques for DNA microarray images," *MDPI Algorithms*, vol. 4, pp. 30–49, 2012.
[17] A. J. R. Neves and A. J. Pinho, "Lossless compression of microarray images using image-dependent finite-context models," *IEEE Transactions on Medical Imaging*, vol. 28, no. 2, pp. 194–201, February 2009.
[18] M. Hernández-Cabronero, J. Muñoz-Gómez, I. Blanes, J. Serra-Sagristà, and M. W. Marcellin, "DNA microarray image coding," in *Proceedings of the IEEE International Data Compression Conference, DCC*, IEEE, Ed., 2012, pp. 32–41.
[19] S. Lonardi and Y. Luo, "Gridding and compression of microarray images," in *Proceedings of the Computational Systems Bioinformatics Conference*. IEEE, 2004, pp. 122–130.
[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] A. J. R. Neves and A. J. Pinho, "Lossless compression of microarray images," in *Proceedings of the International Conference on Image Processing, ICIP*. IEEE, 2006, pp. 2505–2508.
[22] A. Neekabadi, S. Samavi, S. A. Razavi, N. Karimi, and S. Shirani, "Lossless microarray image compression using region based predictors," in *Proceedings of the International Conference on Image Processing*, 2007, pp. 349–352.
[23] S. Battiato and F. Rundo, "A bio-inspired CNN with re-indexing engine for lossless dna microarray compression and segmentation," in *Proceedings of the 16th International Conference on Image Processing*, vol. 1-6. IEEE, 2009, pp. 1717–1720.
[24] A. Pinho, A. Paiva, and A. Neves, "On the use of standards for microarray lossless image compression," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 563–566, March 2006.
[25] 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.
[26] 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.
[27] 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. IEEE, 2007, pp. 1176–1179.
[28] Z. Wang and A. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *Signal Processing Magazine, IEEE*, vol. 26, no. 1, pp. 98 –117, jan 2009.
[29] V. Uslan and I. O. Bucak, "Clustering-based Spot Segmentation of cDNA Microarray Images," in *2010 Annual International Conference of the IEEE in Medicine and Biology Society (EMBC)*, 2010, pp. 1828–1831.
[30] ——, "Microarray image segmentation using clustering methods," *Mathematical & Computational Applications*, vol. 15, no. 2, pp. 240–247, August 2010.
[31] Z. Li and G. Weng, "Segmentation of cDNA Microarray Image using Fuzzy c-mean Algorithm and Mathematical Morphology," in *Functional manufacturing technologies and CEEUSRO II*, 2011, pp. 159–162.
[32] S. Battiato, G. Farinella, G. Gallo, and G. Guarnera, "Neurofuzzy segmentation of microarray images," in *Proceedings of the 19th International Conference on Pattern Recognition, ICPR*, Dec 2008, pp. 1–4.
[33] S. Battiato, G. D. Blasi, G. M. Farinella, G. Gallo, and G. C. Guarnera, "Ad-hoc segmentation pipeline for microarray image analysis," in *S&T-SPIE Electronic Imaging*, 2006.
[34] I. Rezaeian and L. Rueda, "Sub-grid and Spot Detection in DNA Microarray Images Using Optimal Multi-level Thresholding," in *Pattern Recognition in Bioinformatics*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2010, vol. 6282, pp. 277–288.
[35] N. Karimi, S. Samavi, S. Shirani, and P. Behnamfar, "Segmentation of DNA microarray images using an adaptive graph-based method," *IET IMAGE PROCESSING*, vol. 4, no. 1, pp. 19–27, FEB 2010.
[36] E. Athanasiadis, D. Cavouras, D. Glotsos, P. Georgiadis, I. Kalatzis, and G. Nikiforidis, "Segmentation of Complementary DNA Microarray Images by Wavelet-Based Markov Random Field Model," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 6, pp. 1068–1074, nov 2009.
[37] J. Ambroise, B. Bearzatto, A. Robert, B. Govaerts, B. Macq, and J.-L. Gala, "Impact of the spotted microarray preprocessing method on fold-change compression and variance stability," *BMC Bioinformatics*, vol. 12, no. 1, p. 413, 2011.
[38] X. Chen and H. Duan, "A vector-based filtering algorithm for microarray image," in *Proceedings of the International Conference on Complex Medical Engineering*, vol. 1-4. IEEE/ICME, 2007, pp. 794–797.
[39] P. Hegde, R. Qi, K. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. Hughes, E. Snesrud, N. Lee, and J. Quackenbush, "A concise guide to cDNA microarray analysis," *BIOTECHNIQUES*, vol. 29, no. 3, pp. 548+, SEP 2000.
[40] T. Atherton and D. Kerbyson, "Size invariant circle detection," *Image and Vision Computing*, vol. 17, no. 11, pp. 795–803, sep 1999.

4.2 Relative Quantizer

```
@Article{Hernandez15RQ,
               Title     = {{Analysis-Driven Lossy Compression of DNA Microarray Images}},
               Author    = {Miguel Hern{\'a}ndez-Cabronero and Ian Blanes
                            and Armando J. Pinho and Michael W. Marcellin and Joan Serra-Sagrist{\`a}},
               journal   = {Submitted to the IEEE Transactions on Medical Imaging},
}
```

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.

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.

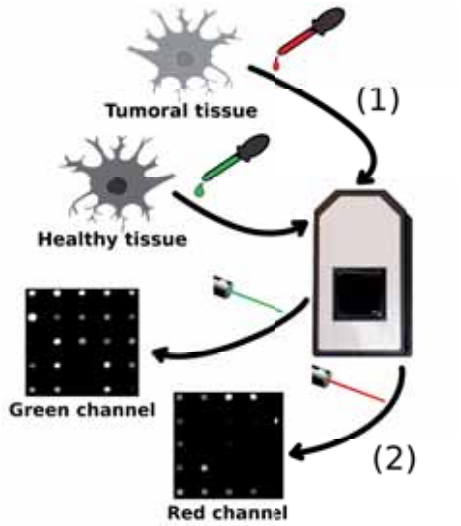


Fig. 1: Outline of an example DNA microarray image acquisition procedure. Samples from healthy and tumoral tissue are dyed with fluorescent pigments and put on a DNA microarray chip (1), which is optically scanned using two different wavelength lasers to produce two microarray images (2).

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.

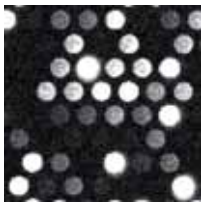
Samples coming from the healthy and tumoral tissues are first dyed with, respectively, green and red fluorescent markers (step (1) in Fig. 1). After that, the biological samples are left to react on the surface of the DNA microarray chip, which contains microscopic holes or *spots* arranged in a regular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

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.



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

(a) (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.

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 possess several properties 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 results (summarized here in Table IV) provide compression ratios less 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 aim of this work is to provide significant improvements, compared to existing lossy techniques, by the design of an approach informed by the process employed in DNA microarray analysis.

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 and no genetic information will be subsequently extracted from it. Even if a spot is correctly 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.

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 **0100**	5 **0101**	6 **0110**	7 **0111**	8 **1000**	9 **1001**	10 **1010**	11 **1011**	12 **1100**	13 **1101**	14 **1110**	15 **1111**
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	**0101** 5	**0101** 5	**0111** 7	**0111** 7	**1010** 10	**1010** 10	**1010** 10	**1010** 10	**1110** 14	**1110** 14	**1110** 14	**1110** 14

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. To this end, professionals working with DNA microarrays usually employ the *corrected ratio of means* (CRM) of each spot [12], defined as

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

Here, μ_{spot} and μ_{localBG} are the average pixel intensity within a spot and its *local background*, respectively. The latter is a region of background pixels near the spot of interest. The exact shape and size of the local background is determined by the segmentation algorithm. The *red* and *green* superscripts refer to each of the image channels being analyzed. The μ_{localBG} is subtracted from μ_{spot} to compensate for background noise and unavoidable inaccuracies in the segmentation process. Since posterior statistical analyses rely on the CRM, lossy coding methods applied to DNA microarray images should minimize their impact on it.

In what follows, the error introduced in the CRM is taken as a measure of distortion introduced by lossy compression within 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. 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.

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.

The fixed-rate scalar quantizer that minimizes relative error for continuous-amplitude sources has been described in the literature [16]. 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. 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. 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 integer of bitdepth $B \geq 1$. The RQ is parameterized by an integer k in $\{1, \ldots, B\}$, which controls the precision of the quantizer. 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 $B-1, \ldots, N-k+1$. Note that, by definition, only the k bits in positions $N, ..., N-k+1$, can be different from 0. From this observation, it follows that if $N < k$, then all bits of the pixels are preserved. That is, all pixels having values in $\{0, 1, \ldots, 2^k - 1\}$ 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 $s \in \{2^1, 2^2, \ldots, 2^{B-k}\}$, 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 $B = 16$.

k	1	2	3	4	5	6	7
I_k $(B = 4)$	5	8	12	16	N/A	N/A	N/A
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}, $j = 0, 1, \ldots, B-(k+1)$. 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.

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 detected spots may be distorted.

CRM values are usually classified into one of three categories: *a)* CRM $\in R = [\alpha, \beta]$, *b)* CRM $< \alpha$ or *c)* CRM $> \beta$. Typically, $\alpha = 0.5$ and $\beta = 2$ [3] so that category *a)* indicates roughly equal expression levels, while the other two categories indicate that the gene of interest is more highly expressed in one biological sample or the other. 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.

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* (ARE_{CRM}). Given the analysis results of an original and a distorted (*e.g.*, quantized) image, it is defined as

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

Here, n is the number of spots positively detected in both the original and the distorted images. The CRM extracted from the *i*-th such spot in the original and distorted images are denoted as CRM_i and $\widehat{\text{CRM}_i}$, respectively. The parameter δ is set to 0.001 to stabilize the case $\text{CRM}_i = 0$. As an example, a value of $\text{ARE}_{\text{CRM}} = 0.5$ 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].

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

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

where d is the number of spots that are detected differently in the original and quantized images, c is the number of spots that are positively detected in both the original and quantized images but are classified differently, and m is the total number of spots. Note that m includes both detected and not 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 $R = [\alpha, \beta] = [0.5, 2]$ interval 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, 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 [17]. 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, the *replicate ARE*$_{\mathrm{CRM}}$ (Rep-ARE$_{\mathrm{CRM}}$) is defined as

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

where $\mathrm{CRM}_i^{\mathrm{f}}$ and $\mathrm{CRM}_i^{\mathrm{s}}$ are, respectively, the CRM of the first and second spots of the i-th replicated pair, p is the number of such 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

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

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$_{\mathrm{CRM}}$. Nevertheless, rapid improvement is observed as the parameter k is increased. For $k \geq 4$, the ARE$_{\mathrm{CRM}}$ and the FWDOC are below 8.0% and 4.5%, respectively. Significantly, for all $k > 1$, the ARE$_{\mathrm{CRM}}$ and the FWDOC metrics show a better behavior than the Rep-ARE$_{\mathrm{CRM}}$ 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, this leads to the choice of $k = 3$ or $k = 4$.

Results for the test described in this section have not been obtained for DNA microarray images from other corpora.

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 44 images 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$_{\mathrm{CRM}}$**	**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$_{\mathrm{CRM}}$**	**Rep-FWDOC**
Original	0.254	0.212

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

In this section, several techniques are considered for the coding of RQ indices. Only lossless coding strategies are taken into account, so that the distortion of the entire resulting system is due only to the RQ. Prior to describing the techniques employed, the image corpora used in subsequent experiments are discussed.

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 ApoA1, the ISREC 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 properties. In particular, the total number of *grayscale* images in each corpus is provided in the *Images* 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. The

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

Property	Yeast [18]	ApoA1 [19]	ISREC [20]	Stanford [21]	MicroZip [4]	Arizona [22]	IBB [23]
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) [24]	5.521	10.223	10.199	7.335	8.667	8.275	8.039

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 [24] 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 [24]. 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) for each $k \in \{1, \ldots, 7\}$, and the quantization *indices* were stored as an image. The resulting index images were then subjected to lossless coding using several algorithms. 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 [25], CALIC [26], lossless JPEG2000 [27] and lossless HEVC/H.265 [28]), and the best lossless microarray-specific image compressor (Neves and Pinho's method [24]). Note that publicly available CALIC codecs support 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 [29].

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 $k = 4$, 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[2] 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 [24]. This should be expected, at least for the case of the original DNA microarray images, for which it was designed.

C. Rate-distortion Analysis

The previous sections have demonstrated that compression systems based on the proposed RQ can provide significant compression with negligible effect on the DNA microarray analysis. In what follows, we compare the performance of RQ-based compression with more classical 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 [30] and HDR-VDP 2 [31] may not be useful, since microarray images are analyzed by algorithms and not by human observers.

To the best of the authors' knowledge, all lossy algorithms that report the distortions introduced in the DNA microarray image analysis process [1]–[3] are based on either lossy JPEG2000 or near-lossless JPEG-LS. The direct study of the actual lossy microarray-specific methods in the literature [1], [2], [4]–[6] is not possible due to the lack of available software implementations and the image corpora used for benchmarking. Therefore, standard lossy JPEG2000 and near-lossless JPEG-LS are applied to the original images and used

[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.

[2]Invocation instructions and configuration file are available at http://deic.uab.es/~mhernandez/media/software/hevc_lossless.cfg.

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

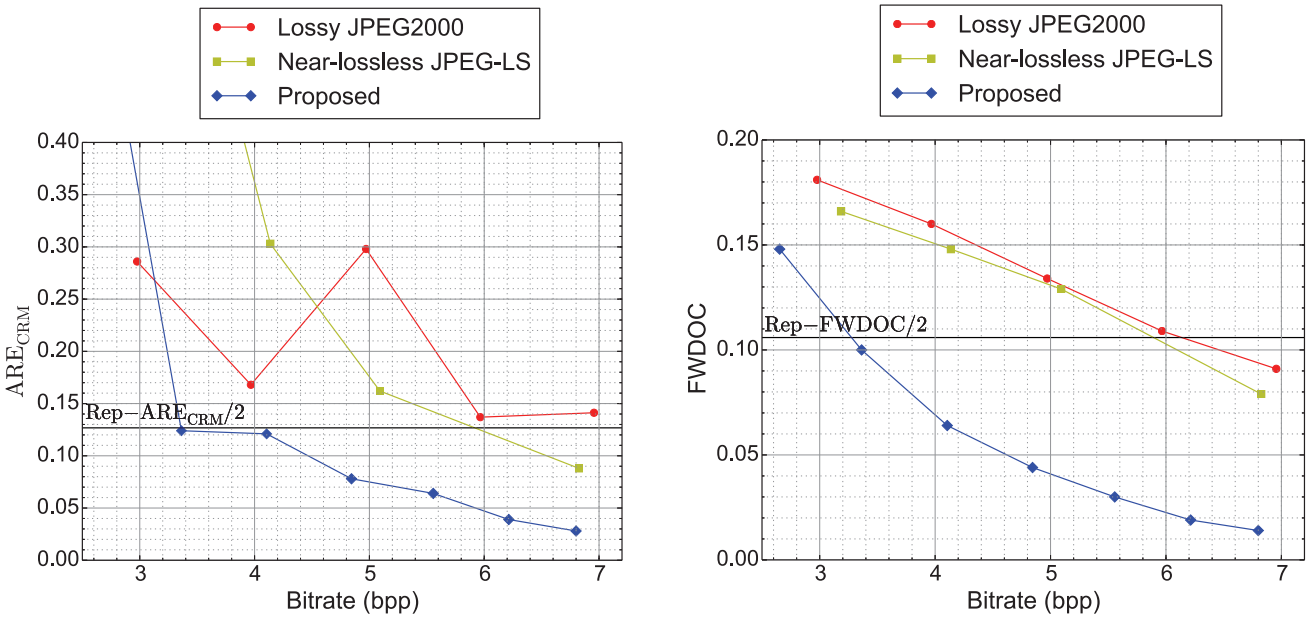


Fig. 3: Distortion metrics versus bitrate: (a) average relative error in the CRM (ARE_{CRM}); (b) fraction of spots incorrectly detected or classified (FWDOC). Half the average replicate CRM relative error (Rep-$\text{ARE}_{\text{CRM}}/2$) and half the fraction of replicated spots wrongly detected or classified (Rep-FWDOC / 2) are shown as horizontal lines in (a) and (b), respectively.

to provide comparisons with the proposed RQ-based coder (using Neves and Pinho's lossless compressor).

The resulting rate-distortion curves for the IBB corpus are shown in Fig. 3. The average results for each value of $k \in \{1, 2, \ldots, 7\}$ (the proposed scheme), each target bitrate $R \in \{3, 4, \ldots, 7\}$ bpp (lossy JPEG2000), and each maximum absolute error $\varepsilon_{\text{abs}} \in \{4, 16, 32, 64\}$ (near-lossless JPEG-LS) are shown. Note that the high ARE_{CRM} yielded by $k = 1$ has been omitted in Fig. 3a. 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 both lossy JPEG2000 and near-lossless JPEG-LS for both the ARE_{CRM} and FWDOC metrics. At about only 3.4 bpp, the proposed algorithm produces less than Rep-ARE_{CRM} / 2 and Rep-FWDOC / 2, *i.e.*, half the 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 (see Table V).

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 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] 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.

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

[18] 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.

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

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

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

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

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

[24] 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.

[25] 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.

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

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

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

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

[30] 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.

[31] 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.

4.3 Progressive Relative Quantizer

```
@Article{Hernandez15PRQ,
               Title     = {{Progressive Lossy-to-Lossless Compression of DNA Microarray Images}},
               Author    = {Miguel Hern{\'a}ndez-Cabronero and Ian Blanes
                            and Armando J. Pinho and Michael W. Marcellin and Joan Serra-Sagrist{\`a}},
               journal   = {Submitted to the IEEE Signal Processing Letters},
}
```

Progressive Lossy-to-Lossless 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*—The analysis techniques applied to DNA microarray images are under active development. As new techniques become available, it will be useful to apply them to existing microarray images to obtain more accurate results. The compression of these images can be a useful tool to alleviate the costs associated to their storage and transmission. The recently proposed Relative Quantizer (RQ) coder provides the most competitive compression ratios while introducing only acceptable changes in the images. However, images compressed with the RQ coder can only be reconstructed with a limited quality, determined before compression. In this work, a progressive lossy-to-lossless scheme is presented to solve this problem. The regular structure of the RQ intervals is exploited to define a lossy-to-lossless compression algorithm called the PRQ coder. An enhanced version that prioritizes a region of interest –the PRQ-ROI coder– is also described. Experiments indicate that the proposed algorithms achieve lossless compression results almost identical to those of the non-progressive RQ coder. Moreover, the PRQ-ROI coder yields better rate-distortion results than both the RQ and PRQ coders.**

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

I. Introduction

DNA microarrays are a state-of-the-art tool in biology and biomedicine. Laboratories around the world employ microarrays to monitor in parallel the function and regulation of thousands of genes of an organism [1]. When a DNA microarray experiment is performed, two biological samples are put on a microarray chip, which is then scanned to produce two grayscale images. Finally, the images are analyzed to extract the genetic data of interest. The different parts of the analysis process are under active development [2]–[8]. This work does not focus on the analysis of DNA microarray images and, hence, a discussion of previous works is out of scope. The interested reader can find a complete review of the state of the art of this topic in [7]. As new analysis techniques are developed, it will be desirable to apply them to obtain more accurate genetic data from previously performed experiments. However, repeating all parts of the experiment is usually not an option because the required biological samples may not be available some time after performing the original experiment, or because the re-analysis may need to be performed in another laboratory. Therefore, storing the DNA microarray images is paramount to enable future, more accurate genetic data extraction. In order to facilitate the management and transmission of these images, image compression emerges as a valuable tool.

The lossless compression of DNA microarray images has proven to be a very challenging task. Compression ratios significantly better than 2:1 are not generally obtained even by algorithms specifically designed for such images [9]. On the other hand, lossy coders can yield arbitrary compression ratios at the cost of modifying the images. Even though subsequent analysis techniques may be distorted by these modifications, sufficiently small distortions can be considered acceptable [10]–[12]. Several generic image compression approaches (or adaptations thereof) have been applied to DNA microarray images [10], [11], [13]–[16]. Since these coding techniques are not specifically designed with the analysis of microarray images in mind, results for these methods may not be optimal. A lossy compression method expressly designed for this type of images was recently proposed [9]. In spite of its very competitive rate-distortion results, superior to existing lossy and lossy-to-lossless compressors, this technique does not offer a progressive reconstruction of the images. In this work, a lossy-to-lossless compression scheme for DNA microarray images is proposed.

The rest of this paper is structured as follows. Section II describes the most relevant features of the technique presented in [9]. A progressive lossy-to-lossless coding approach is proposed in Section III and its compression performance is analyzed in Section IV. Finally, Section V draws some conclusions.

II. The Relative Quantizer

The lossy compression method presented in [9] is based on a non-uniform scalar quantizer called the *Relative Quantizer* (RQ). This quantizer is applied independently to unsigned pixels of images of bitdepth $B \geq 1$. The quantization intervals of the RQ are fully determined by an integer parameter $k \in \{1, \ldots, B\}$ that controls the precision of the quantization process. The first 2^k intervals have size 1 and, hence, pixel intensities in $\{0, \ldots, 2^k - 1\}$ are preserved losslessly. The next 2^{k-1} intervals have each size 2^1 and the following 2^{k-1} intervals have size 2^2. Each successive group of 2^{k-1} intervals has size $2^3, 2^4$, etc until the last group, which contains intervals of size 2^{B-k}. A diagram of the quantization intervals of the

This work has been partially funded by FEDER, the Spanish Government (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 Universitat Autònoma de Barcelona, Bellaterra 08193, Spain (e-mail: mhernandez@deic.uab.cat).

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

M. W. Marcellin is with the University of Arizona, Tucson, AZ 85721-0104, USA.

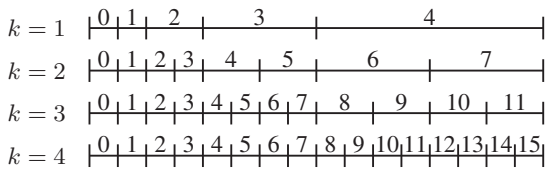


Fig. 1: Quantization intervals of the RQ for $B = 4$ and all possible values of k. The index of each interval is also indicated.

RQ for $B = 4$ and all possible values of k is shown in Fig. 1. The index associated to each of the intervals is also provided in the figure. Note that the RQ for $k = B$ corresponds to a uniform quantizer of step size 1, *i.e.*, not performing any quantization. This definition allows the RQ to yield very competitive compression performance while introducing only an acceptable distortion in subsequent analysis results [9]. Thus, the lossy-to-lossless scheme proposed in Section III is based on the RQ.

In [9], the quantization indices produced by the RQ are coded with the lossless compressor described in [17]. The bitplanes are compressed in raster order using an arithmetic coder (AC), beginning with the most significant bitplane. The probabilities used to drive the AC are computed based on a model that employs information from previously encoded bits. The position of the bits employed to extract that information is referred to as the *context*. A key property of this algorithm is the fact that the actual context employed in each bitplane is calculated at compression time. To calculate the best context, a greedy algorithm evaluates each candidate context by compressing a rectangular region of the center of the image using that context. After that, the best-performing candidate is selected. Thence, this image-dependent algorithm is able to very accurately adapt to the individual properties of each bitplane. As a result of this flexibility, this compressor exhibits the best performance for this type of images and the RQ-quantized versions thereof. Therefore, a version of this algorithm is employed in the lossy-to-lossless coder proposed in Section III.

The main drawback of the RQ coder is the fact that compressed images can only be reconstructed up to a certain precision determined by the chosen value of k. In particular, the original image cannot be recovered for any $k < B$. If the user wants to reconstruct the image losslessly or at different qualities, several compressed versions of the image need be kept. This approach multiplies the storage requirements and, thus, is not practical. In what follows, a microarray-specific progressive lossy-to-lossless scheme based on the RQ and the compression algorithm from [17] is presented and its performance is analyzed.

III. Progressive Lossy-to-Lossless Coding

A. Progressive Representation of DNA Microarray Images

In this section, the relationship between the quantization intervals of the RQ for different values of k is exploited to define a lossless, progressive representation of DNA microarray image pixels. As described in Section II, the quantization intervals of the RQ for a given image bitdepth B are determined by an integer parameter $k \in \{1, \ldots, B\}$. The parameter $k = 1$ corresponds to the most aggressive (least precise) quantization and $k = B$ corresponds to not performing any quantization. A key observation is that the RQ with parameter $k = k_0$ with $k_0 > 1$ is actually a refined version of the RQ with parameter $k = k_0 - 1$. More specifically, all quantizer intervals that include more than one pixel intensity for $k = k_0 - 1$ are divided in half for $k = k_0$. For instance, let us consider the case $B = 4$, shown in Fig. 1. The interval with index 3 for $k = 1$ is divided into the intervals with indices 4 and 5 for $k = 2$. Likewise, the interval with index 6 for $k = 2$ is divided into the intervals with indices 8 and 9 for $k = 3$. On the other hand, intervals that contain only one pixel intensity cannot be further refined and their index is identical to the contained pixel intensity. For example, for $k = 3$, only intervals with index $i > 7$ can be refined. For $k = B$, all intervals contain only one pixel intensity and their index is identical to that intensity. Analogous relationships between the quantization intervals apply for any $B > 1$ including $B = 16$, the bitdepth of DNA microarray images.

The previous observations can be formalized to enable a lossless, progressive representation of these type of images. Given a pixel intensity p, let $\mathrm{RQ}_k(p)$ be the quantization interval corresponding to p for the RQ with parameter k and let $|\mathrm{RQ}_k(p)|$ be the number of intensities assigned to that interval. Clearly, $\mathrm{RQ}_B(p) \subset \mathrm{RQ}_{B-1}(p) \subset \cdots \subset \mathrm{RQ}_1(p)$. If the index of the $\mathrm{RQ}_1(p)$ interval is encoded, the decoder knows that $p \in \mathrm{RQ}_1(p)$. However, if $|\mathrm{RQ}_1(p)| > 1$, the exact value of p is not known. In this scenario, additional information can be encoded to allow a more precise reconstruction of that pixel. Since $|\mathrm{RQ}_1(p)| > 1$, the interval $\mathrm{RQ}_1(p)$ is divided into two intervals of size $|\mathrm{RQ}_1(p)|/2$ for $k = 2$, as in the example above. Therefore, only one *refinement bit* is needed to signal which of these two intervals corresponds to $\mathrm{RQ}_2(p)$. Hereinafter, a refinement bit equal to 0 (resp. 1) is used when $\mathrm{RQ}_{k+1}(p)$ equals the lower (resp. upper) half of $\mathrm{RQ}_k(p)$. By encoding this bit, the range of possible reconstruction values is halved and, hence, the precision is doubled. Likewise, if $\mathrm{RQ}_2(p)$ comprises more than one value, another refinement bit can be encoded so that the decoder can determine which candidate interval corresponds to $\mathrm{RQ}_3(p)$. By successively applying this refinement process, it is possible to sequentially determine $\mathrm{RQ}_1(p), \ldots, \mathrm{RQ}_B(p)$, *i.e.*, the quantization indices corresponding to p for all values of k. Recall that, by definition, $\mathrm{RQ}_B(p)$ allows a lossless reconstruction of the original pixel intensity p.

Based on this, we define here a lossless, progressive RQ-based (PRQ) representation of the pixel p as

$$\mathrm{PRQ}(p) = \mathrm{RQ}_1(p), \Delta_{1\to 2}(p), \ldots, \Delta_{B-1\to B}(p), \qquad (1)$$

where $\Delta_{k\to k+1}(p)$ is the refinement bit needed to obtain $\mathrm{RQ}_{k+1}(p)$ from $\mathrm{RQ}_k(p)$. For instance let $B = 4$ and $p = 11$. As can be seen in Fig. 1, $\mathrm{RQ}_1(11) = 4$, $\mathrm{RQ}_2(11) = 6$, $\mathrm{RQ}_3(11) = 9$ and $\mathrm{RQ}_4(11) = 11$. Therefore, the refinement bits are $\Delta_{1\to 2}(11) = 0$ (lower half), $\Delta_{2\to 3}(11) = 1$ (upper half) and $\Delta_{3\to 4}(11) = 1$ (upper half). Thus, $\mathrm{PRQ}(11) =$

$4, 0, 1, 1$. Note that $|\mathrm{RQ}_n(p)| = 1$ implies that the interval need not be refined and, hence, $\Delta_{m\to m+1}(p)$ need not be signaled for any $m \geq n$. For instance, for $B = 4$ and $p = 3$, $\mathrm{PRQ}(3) = 2, 1$.

The PRQ representation of any pixel p can also be expressed in binary form. As shown in [9], the total number of quantization intervals of the RQ with parameter k is given by $(B - k + 2)2^{k-1}$. Hence, for DNA microarray images ($B = 16$), 17 quantization intervals are employed for $k = 1$. Thus, the index of $\mathrm{RQ}_1(p)$ can be expressed using $\lceil \log_2 17 \rceil = 5$ bits. Once the first element of the PRQ is signaled, each refinement bit $\Delta_{k\to k+1}(p)$ provides enough information to recover the quantization interval index for the RQ with the next value of k. Therefore, at most $B - 1 = 15$ such refinement bits need be coded to enable the recovery of the original pixel value p. By sequentially appending the refinement bits to the index of $\mathrm{RQ}_1(p)$, any pixel can be expressed in a progressive lossy-to-lossless way by signaling at most 20 bits.

B. Progressive Compression

In what follows, a progressive lossy-to-lossless coder for DNA microarray images based on the PRQ representation is introduced. This coder is hereinafter referred to as the *PRQ coder*.

When compressing an image, its 20-bpp PRQ representation is first computed. The resulting data are coded with a version of the algorithm introduced in [17]. This version includes two modifications to adapt the original algorithm to the particularities of the PRQ representation and improve its coding efficiency. As described in Section II, the compressor in [17] proceeds by sequentially coding each of the image bitplanes, beginning with the most significant bit. Hence, the elements of the PRQ representation are coded in the order described in Equation (1). As explained in III-A, some of the 15 refinement bits are not needed for a given pixel p when $|\mathrm{RQ}_n(p)| = 1$ for $n < 16$. Therefore, the first modification consists in not coding any unneeded refinement bit of the PRQ representation. As described in II, the original algorithm selects the optimal context by comparing several candidates, each of which is evaluated by compressing a rectangular $N \times M$ region in the center of the image. If that region is not representative of the whole image, a sub-optimal candidate context would be selected. Thus, the second modification to [17] consists in evaluating the candidate contexts by compressing NM pixels uniformly sampled across the image. Since sampled pixels are not confined in a relatively small region of the image, an overall more precise context can be selected, which can improve the compression performance at a similar computational cost.

When the image is decompressed, the 20-bpp PRQ representation of the image is first obtained by applying a version of the decoder presented in [17]. This version includes modifications analogous to those described above to make it compatible with the output of the encoder. For each pixel p, the index of the $\mathrm{RQ}_1(p)$ interval is then obtained from the 5 most significant bits of its PRQ representation. Finally, the refinement information contained in subsequent bitplanes is successively applied until $\mathrm{RQ}_{16}(p)$ is recovered. By definition, the index of this interval is identical to the original pixel intensity p and, thus, the image can be losslessly recovered. As discussed later in Section IV, the 20 bpp PRQ representation can be losslessly coded with approximately the same performance as the original 16 bpp pixels.

The correct decoding of truncated data is required to enable a progressive lossy-to-lossless coding pipeline. Since the algorithm proposed in [17] is designed for purely lossless coding, it needs be adapted to accept truncated versions of the encoded data produced at the encoder. If 5 or more complete bitplanes are decoded before the end of file (EOF) is reached, $\mathrm{RQ}_1(p)$ and possibly some refinement bits are available for each pixel p. Hence, $\mathrm{RQ}_n(p)$ can be computed for some n with $1 \leq n \leq 16$, depending on the number of refinement bits available. As in the original RQ, the recovered value of p is calculated using the interval midpoint of $\mathrm{RQ}_n(p)$, rounded up to the next integer. If less than 5 complete bitplanes are available, then the index of $\mathrm{RQ}_1(p)$ –corresponding to $k = 1$– would need to be estimated for some pixels. Even though it is possible to do so, this scenario should be generally avoided due to the relatively high distortion introduced in subsequent analysis processes for $k = 1$ [9].

An additional enhancement is described now to improve the coding performance of the progressive lossy-to-lossless PRQ coder described above. The original algorithm introduced in [17] assigns equal priority to all pixels of the image. Thus, all bits of a bitplane are coded before proceeding to the next bitplane. However, almost all information relevant to the analysis of microarray images is contained in pixels inside the so-called *spots* [1], which can be enclosed inside rectangular *regions of interest* (ROIs). Due to the regular layout of the spots in rectangular grids, whose geometry can be determined *a priori*, defining the ROIs is a fast and easy operation. The rest of the image –*i.e.*, the *background*– contains data relatively unimportant for subsequent analysis processes. If all bitplanes of pixels inside a ROI are coded before the bitplanes of background pixels, relevant information is placed closer to the beginning of the compressed file. Therefore, the rate-distortion performance of the PRQ coder can be improved by including this enhancement. The version of the PRQ coder that includes the ROI prioritization enhancement is hereinafter referred to as the *PRQ-ROI coder*.

IV. Compression Performance

The compression performance of the proposed PRQ and PRQ-ROI coders is addressed in this section. First, the bitrate required to obtain a lossless compression is surveyed. After that, the rate-distortion results yielded by these lossy-to-lossless algorithms is discussed.

To test the lossless compression efficiency, 228 real DNA microarray images where compressed with the proposed coders. This corpus contains most images used for the benchmarking of microarray image compressors in the literature. The average compression results in bpp –calculated as the total number of compressed bits required for a lossless recovery of the images divided by the total number of pixels

TABLE I: Average lossless compression bitrate in bpp and execution time expressed in relation to [17].

	Neves and Pinho [17]	PRQ coder	PRQ-ROI coder
Bitrate	7.909	7.871	7.892
Time	100%	97.38%	99.76%

in all images– is provided in Table I. Results for the best-performing lossless compressor for DNA microarray images –first published in [17]– are also provided for comparison. The average time for compressing and decompressing 4 times each of the 228 images, expressed as a percentage of the execution time of [17], is also provided in the table.

It can be observed that both the PRQ and the PRQ-ROI coders achieve a slightly better lossless coding efficiency than the best state-of-the-art lossless compressor [17], even though the latter does not offer lossy-to-lossless capabilities. This can be explained by the modifications described in Section III-B (skipping of unneeded refinement bits and improved candidate context evaluation). It can also be observed that the PRQ-ROI coder yields a lossless compression performance almost identical to that of the PRQ coder. This suggests that the lossless coding overhead due to the ROI prioritization capabilities is negligible. As can be seen in the table, the PRQ and PRQ-ROI coders are, respectively, 2.62% and 0.24% faster than the non-progressive algorithm from [17]. These differences are due to the fact that the PRQ-based coders skip the coding of all unneeded refinement bits, which compensates for the larger amount of bitplanes that need be processed.

Since the proposed PRQ coder is a progressive lossy-to-lossless algorithm, it is paramount to analyze its rate-distortion performance. To do so, it is necessary to assess the amount of distortion introduced in the results of subsequent analysis processes. Traditional image distortion metrics such as MSE are not suitable for DNA microarray images because they do not characterize the analysis result distortion [9]. Instead, the microarray-specific metrics introduced in [9] –the $\mathrm{ARE}_{\mathrm{CRM}}$ and the FWDOC metrics– are hereinafter employed to assess the distortion introduced in the images. Unlike traditional metrics, these microarray-specific metrics directly compare the data yielded by real analysis software when applied to the original or the modified images. Hence, the $\mathrm{ARE}_{\mathrm{CRM}}$ and the FWDOC metrics provide an accurate measure of the distortion introduced in subsequent analysis processes. In this work, 44 of the 228 aforementioned DNA microarray images are considered. The rest of the images could not be used because of the lack of publicly available analysis software compatible with these images.

First, the 44 images were compressed with the PRQ-Uniform and the PRQ-ROI coders and the resulting codestreams were truncated at 7 different lengths. The first truncation point was selected so that only the first element of the PRQ representation of each pixel is available. The next truncation point was chosen so that the first refinement bit is also available for all pixels. Each of the successive truncation points was selected so that exactly one more refinement bit is available for all pixels, as compared to the previous truncation point. For each truncation point and coder, a reconstructed version of the 44 images was then obtained. Finally, the two aforementioned distortion metrics were calculated for each reconstructed version. The average distortion results for the PRQ and the PRQ-ROI coders as a function of the average bitrate after truncation is provided in Fig. 2 for the $\mathrm{ARE}_{\mathrm{CRM}}$ and the FWDOC metrics, respectively. The rate-distortion results for the non-progressive RQ coder for $k \in \{1, \ldots, 7\}$ are also provided for comparison.

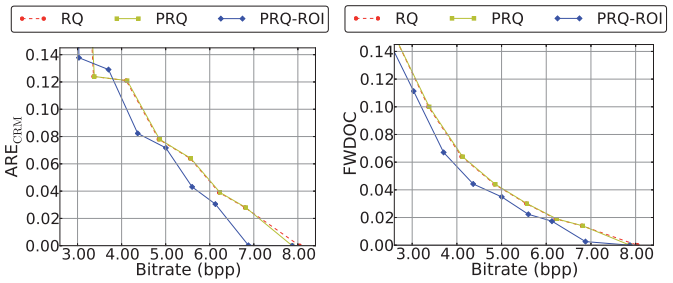


Fig. 2: Rate distortion results for the $\mathrm{ARE}_{\mathrm{CRM}}$ metric (left) and the FWDOC metric (right).

It can be observed that the PRQ and the RQ coders yield almost identical results for all tested bitrates. This suggests that the PRQ representation introduces only a negligible overhead even when only some of the refinement bits are coded. It can also be seen that the PRQ-ROI coder generally yields significantly better rate-distortion results than both the PRQ and RQ coders. This can be explained by the fact that the information important for subsequent analysis is coded before the relatively unimportant information of the background. In light of these data, it appears that the ROI-prioritization is an effective way of enhancing the rate-distortion performance of the PRQ coder.

V. Conclusions

Better analysis techniques for DNA microarray images are being actively investigated. Hence, it is convenient to store the images to enable future re-analysis of the data. The compression of this type of images is a useful tool to reduce the storage and management costs and to accelerate the sharing of these images. Lossy coding algorithms can yield high compression ratios introducing only acceptable distortion in subsequent analysis processes. A lossy compression method called Relative Quantization (RQ) was recently proposed. In spite of its competitive compression performance, an image coded with the RQ can only be reconstructed with a certain quality level determined before compression and it is not possible to recover the original image. This work introduces an original solution to this problem. First, a lossless representation of DNA microarray images is defined. Then the PRQ coder, a progressive lossy-to-lossless coder based on this representation, is proposed. Finally, an enhanced version of this coder that includes region-of-interest prioritization – the PRQ-ROI coder– is described. The proposed PRQ-ROI exhibits significantly better rate-distortion results than the non-progressive RQ coder without introducing any overhead in the lossless compression bitrate nor in the average execution time.

REFERENCES

[1] S. Moore, "Making chips to probe genes," *IEEE Spectr.*, vol. 38, no. 3, pp. 54–60, Mar. 2001.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] L. Rueda, Ed., *Microarray Image and Data Analysis: Theory and Practice*. CRC Press, 2014.

[8] L. Srinivasan, Y. Rakvongthai, and S. Oraintara, "Microarray image denoising using complex Gaussian scale mixtures of complex wavelets," *Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1423–1430, 2014.

[9] M. Hernández-Cabronero, I. Blanes, A. J. Pinho, M. W. Marcelling, and J. Serra-Sagristà, "Analysis-Driven Lossy Compression of DNA microarray images," *Submitted to IEEE Transactions on Medical Image*, 2015.

[10] R. Jornsten, 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, Apr. 2003.

[11] 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, Jan. 2004.

[12] 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, Feb. 2009.

[13] N. Faramarzpour and S. Shirani, "Lossless and lossy compression of DNA microarray images," in *Proceedings of the IEEE International Data Compression Conference, DCC*, 2004, pp. 538–538.

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

[15] 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.

[16] 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.

[17] 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, Feb. 2009.

Chapter 5

Conclusions

5.1 Summary

Medical imaging has gained increasing importance in the diagnosis and research of many diseases. DNA microarrays are state-of-the-art biomedical imaging tools extensively employed in laboratories across the world for the study of the function and regulation of thousands of genes in parallel. The long-term storage of DNA microarray images enables future, more accurate study of these genes. Thus, image compression emerges as a natural approach to reducing the costs associated to their storage and transmission.

DNA microarray images possess many properties that are different from those present in natural images. Microarray images usually exhibit larger dimensions and each pixel requires 16 bits to be stored without coding, whereas natural images usually require 8. Moreover, microarray images have very different intensity distributions and typically employ only a fraction of all possible pixel values. All these differences explain the relatively low compression performance obtained by general-purpose or standard image compression algorithms when applied directly to DNA microarray images. Hence, coding techniques specific for this type of images is required to attain competitive results. The measurement of the entropy present in DNA microarray images suggests that context-based segmentation-based approaches can yield performance gains, as compared to generic data compression algorithms. Notwithstanding,

according to the entropy results, compression ratios greatly exceeding 2:1 should not be expected unless a significant breakthrough in lossless coding technology is made.

Lossless compression is the most commonly employed approach for DNA microarray images. Several publications have addressed the problem of creating efficient lossless coding algorithms, most of which rely on image segmentation or context-based approaches. However, none of these microarray-specific techniques is compliant with existing image compression standards. Therefore, it is not possible to guarantee the availability of compatible decoders in future platforms, hindering the main purpose of storing the images. In [26], a compression approach fully compatible with the JPEG2000 standard is proposed. By exploiting the statistical distribution of DNA microarray image pixels, a reversible low-complexity transform –the Histogram Swap Transform or HST– is able to consistently improve the lossless compression performance of JPEG2000. Depending on the set, gains between 1.97% and 15.53% are observed. With this transform, JPEG2000 becomes the overall best image compression standard, with results closer to microarray-specific algorithms. For convenience, Table 5.1 summarizes the improvements due to the HST.

Table 5.1: Compression performance of JPEG2000 in bits per pixel with and without the Histogram Swap Transform. All results are shown for the best number of spatial DWT decorrelation levels.

Corpus	**Without the HST**	**With the HST**	**Difference**
Yeast	6.829	5.911	0.918 (15.53%)
ApoA1	10.999	10.786	0.213 (1.97%)
ISREC	10.888	10.624	0.264 (2.48%)
Stanford	7.969	7.685	0.284 (3.70%)
MicroZip	9.467	9.157	0.310 (3.39%)
Omnibus	7.549	7.103	0.446 (6.28%)
Arizona	9.064	8.795	0.269 (3.06%)
IBB	9.182	8.392	0.790 (9.41%)

In order to improve the lossless compression performance of standard and specific techniques, the multicomponent compression of DNA microarray images is also investigated. It was found that grouping together the two images produced in each experiment is the optimal configuration in terms of correlation. With this approach,

average correlations between 0.75 and 0.92 are observed for the different corpora. Several methods for exploiting the spectral redundancy –including the DWT, the RKLT, DPCM and the RHaar transforms– were surveyed in [28, 27]. The attained results in bits per pixel for JPEG2000 using these transforms are provided in Table 5.2. Results for the spectral transforms that improve the results for a given set, as compared to not applying any spectral decorrelation transform, are highlighted in bold font. Note that only the Yeast, ApoA1, ISREC, Arizona and IBB corpora from Table 5.1 are included. The other corpora did not consist of pairs of images of the same size, so the spectral decorrelation could not be applied. In spite of the improvements yielded by some transform/corpus combination, it was not possible to find any single decorrelation transform that consistently enhances the compression performance.

Table 5.2: Compression performance of JPEG2000 for different spectral decorrelation transforms and no spatial decorrelation.

Corpus	**No transform**	**5/3 DWT**	**RKLT**	**DPCM**	**RHaar**
Yeast	6.829	**6.786**	9.279	**6.439**	**6.790**
ApoA1	11.524	**11.217**	**10.956**	**11.289**	**11.218**
ISREC	10.888	11.451	11.468	11.203	11.452
Arizona	9.548	9.649	**9.439**	**9.386**	9.649
IBB	9.182	9.948	10.269	9.602	9.948

The lossy compression of DNA microarray images can provide significantly better compression ratios than lossless algorithms. Moreover, the changes introduced in the images can be small enough to produce only acceptable alterations in subsequent analysis processes, which justifies the use of lossy compression. Notwithstanding, these alterations need to validated. In [30], a distortion metric –the Microarray Distortion Metric (MDM)– was proposed to predict the impact on subsequent image analysis results and assess the acceptability of a lossy compression process. Furthermore, this metric does not require the application of any image analysis algorithm, facilitating its incorporation in any compression/decompression pipeline. Simulations reveal that the MDM is able to differentiate important and unimportant modifications of the image and to quantify the amount of distortion introduced.

Even though several lossy compression techniques have been applied to DNA microarray images in the literature, most of them consist in the application of generic image compression methods (or adaptations thereof) to these images. An original microarray-specific compression algorithm –the Relative Quantizer (RQ) coder– was first introduced in [31]. By using specially crafted quantization intervals, the RQ is able to preserve the features that are more relevant to subsequent analysis processes. With the RQ, the compression performance of DNA microarray images is greatly improved while introducing only acceptable changes in the analysis results. Besides, the amount of introduced distortion can be traded of for compression performance by changing the value of an integer parameter k. For adequate values of k, average compression ratios exceeding 4.5:1 are obtained without introducing significant analysis results distortion. This figure is to be compared with the 2:1 ratio that is usually regarded as a practical limit to the lossless compression of DNA microarray images.

The proposed RQ method is purely lossless and the quality of the reconstructed image is determined by the selected value of k. Hence, in order to recover image versions of different quality (including the original data), several compressed versions need to be stored. In [32], a progressive lossy to lossless scheme –the Progressive RQ (PRQ) coder– was presented to solve this problem. With the PRQ coder, images can be coded once and recovered at any quality (including lossless reconstruction) with compression performance slightly better than the best lossless algorithm of the state of the art. By prioritizing a region of the image that contains most information relevant to subsequent analysis algorithms, rate-distortion results significantly better than those of the non-progressive RQ coder are obtained. Moreover, the progressive lossy-to-lossless coders proposed in [32] do not introduce any time complexity overhead. Instead, execution time reductions of up to 2.62% are observed, as compared to the best-performing lossless coding algorithm, on which the RQ and the PRQ coders are based.

5.2 Future Work

In light of all previous discussions, it is possible to for a big picture of DNA microarray image compression and speculate about the future of this field.

The lossless coding of this type of images appears to be close to the practical performance limit with current compression technology. Existing microarray compressors already improve upon the first-order and conditional entropy of all tested corpora and achieve performance only 0.8 bpp or less worse than the conditional entropy after segmentation for 7 of the 8 corpora. Since the entropy after segmentation does not take into account the overhead required to code the segmentation mask, it is reasonable to state that existing compressors are close to the optimal coding performance for the state-of-the-art decorrelation techniques. As discussed above, many bitplanes exhibit properties similar to those of white noise, which is generally considered to be incompressible. Of course, it is possible that a new highly effective decorrelation methods are found for DNA microarray images and more efficient compressors can be implemented. Notwithstanding, there are compelling reasons to think that significantly better lossless compression results cannot be obtained without a large conceptual breakthrough.

On the other hand, the lossy compression of DNA microarray images has been less thoroughly researched and several interesting questions remain open. One of the most intriguing is the relationship between changes in the images and distortion in subsequent analysis processes. Even though the microarray distortion metric (MDM) proposed as a part of this thesis is able to distinguish important and unimportant changes in the images, a fully quantitative method for predicting the impact on the analysis results is yet to be found. Thus, a very attractive research line is to further research the relationship between image changes and analysis distortion. Although stimulating, this research is likely to be arduous due to the highly non-linear behavior of the analysis algorithms, specially of the segmentation stage, when image pixel intensities are modified. A deeper knowledge of the aforementioned relationship would allow the definition of more precise distortion metrics, which, in turn, could be used to significantly improve the rate-distortion results of standard coders such as JPEG2000

via its post-compression rate-optimization (PCRD-op) mechanisms. Another appealing application of this knowledge is determining the extent to which an image can be modified so that only a certain amount of distortion is introduced in the analysis. This would help in the design of new lossy compression approaches and the improvement of existing algorithms such as the RQ and PRQ coders proposed in this thesis. For instance, the quantization scheme could be adapted to meet the maximum image distortion criteria while allowing a more efficient coding of the quantization indices.

Finally, another interesting continuation line for this thesis would be to incorporate added-value features in the compression algorithms. For instance, Jörnsten *et al.* proposed in 2003 a compression algorithm that allows the individual decompression of each spot in the image. This functionality could be used to retrieve only the spots corresponding to the genes of interest in future re-analysis processes. Analogous functionality could be included in more recent lossless and lossy coding algorithms. Another example of valuable functionality would be combining compression and visualization. Most modern scanners output a false-color representation of the two grayscale channels that some researchers visualize to obtain qualitative information about the experiment results. This extra output image could be avoided by compressing the images so that the false-color representation can be displayed without prior decompression. Standard tools such as JPEG2000-compliant viewers could be used for this purpose.

Appendix A

List of All Publications

The list of all publications produced for this thesis are provided next in chronological order:

1. **Miguel Hernández-Cabronero**, Ian Blanes, Michael W. Marcellin, Joan Serra-Sagristà, "A review of DNA microarray image compression," In Proceedings of the International Conference on Data Compression, Communication and Processing, CCP, pp 139-147, June 2011.

2. **Miguel Hernández-Cabronero**, Ian Blanes, Michael W. Marcellin, Joan Serra-Sagristà, "Standard and specific compression techniques for DNA microarray images," MDPI Algorithms, pp 30-49, 2012.

3. **Miguel Hernández-Cabronero**, Juan Muñoz-Gómez, Ian Blanes, Joan Serra-Sagristà, Michael W. Marcellin, "DNA microarray image coding," In proceedings of the IEEE Data Compression Conference, DCC, pp 32-41, 2012.

4. **Miguel Hernández-Cabronero**, Francesc Aulí-Llinàs, Joan Bartrina-Rapesta, Ian Blanes, Leandro Jiménez-Rodríguez, Michael W. Marcellin, Juan Muñoz-Gómez, Victor Sanchez, Joan Serra-Sagristà, Zhongwei Xu, "Multicomponent compression of DNA microarray images," In Proceedings of the CEDI Workshop on Multimedia Data Coding and Transmission, WMDCT, 2012.

5. **Miguel Hernández-Cabronero**, Victor Sanchez, Michael W. Marcellin, Joan Serra-Sagristà, ”A distortion metric for the lossy compression of DNA microarray images,“ In proceedings of the IEEE International Data Compression Conference, DCC, pp 171-180, 2013.

6. **Miguel Hernández-Cabronero**, Victor Sanchez, Michael W. Marcellin, Joan Serra-Sagristà, "Compression of DNA Microarray Images", In Book "Microarray Image and Data Analysis: Theory and Practice", CRC Press, pp 193-222, 2014.

7. **Miguel Hernández-Cabronero**, Ian Blanes, Armando J. Pinho, Michael W. Marcellin, Joan Serra-Sagristà, ”Analysis-Driven Lossy Compression of DNA Microarray Images,“ Submitted to IEEE Transactions on Medical Imaging.

8. **Miguel Hernández-Cabronero**, Ian Blanes, Armando J. Pinho, Michael W. Marcellin, Joan Serra-Sagristà, ”Progressive Lossy-to-Lossless Compression of DNA Microarray Images,“ Submitted to IEEE Signal Processing Letters.

Appendix B

Acronyms

AC Arithmetic coder

bpp Bits per pixel

CNN Cellular neural network

CT Computer tomography

DICOM Digital Imaging and Communications in Medicine standard

DNA Deoxyribonucleic acid

DWT Discrete wavelet transform

EOF End of file

HST Histogram Swap Transform

KLT Karhunen-Loève transform

MCT Multicomponent transform

PCRD-op Post-compression rate-distortion optimization

PRQ Progressive Relative Quantizer

RHaar Reversible Haar

ROI Region of Interest

RQ Relative Quantizer

Bibliography

[1] S. Moore, "Making chips to probe genes," *IEEE Spectrum*, vol. 38, no. 3, pp. 54–60, Mar. 2001.

[2] L. Rueda, Ed., *Microarray Image and Data Analysis: Theory and Practice*. CRC Press, 2014.

[3] J. DeRisi, L. Penland, P. Brown, M. Bittner, P. Meltzer, M. Ray, Y. Chen, Y. Su, and J. Trent, "Use of a cDNA microarray to analyse gene expression patterns in human cancer," *Nature Genetics*, vol. 14, no. 4, pp. 457–60, Dec. 1996.

[4] S. Satih, N. Chalabi, N. Rabiau, R. Bosviel, L. Fontana, Y.-J. Bignon, and D. J. Bernard-Gallon, "Gene Expression Profiling of Breast Cancer Cell Lines in Response to Soy Isoflavones Using a Pangenomic Microarray Approach," *Omics - A Journal of Integrative Biology*, vol. 14, no. 3, pp. 231–238, Jun. 2010.

[5] M. S. Giri, M. Nebozhyn, L. Showe, and L. J. Montaner, "Microarray data on gene modulation by HIV-1 in immune cells: 2000-2006," *Journal of Leukocyte Biology*, vol. 80, no. 5, pp. 1031–1043, Nov. 2006.

[6] Z. Bozdech, S. Mok, and A. Gupta, "DNA Microarray-Based Genome-Wide Analyses of Plasmodium Parasites," in *Malaria*, ser. Methods in Molecular Biology, R. Ménard, Ed. Humana Press, 2013, vol. 923, pp. 189–211.

[7] K. Blekas, N. Galatsanos, A. Likas, and I. Lagaris, "Mixture Model Analysis of DNA Microarray Images," *IEEE Transactions on Medical Imaging*, 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 Transactions on Medical Imaging*, 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 Transactions on Medical Imaging*, 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 Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 1, pp. 181–192, Jan. 2013.

[12] L. Srinivasan, Y. Rakvongthai, and S. Oraintara, "Microarray image denoising using complex Gaussian scale mixtures of complex wavelets," *Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1423–1430, 2014.

[13] N. Faramarzpour, S. Shirani, and J. Bondy, "Lossless DNA microarray image compression," in *Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers*, vol. 2, November 2003, pp. 1501–1504.

[14] R. Jornsten, 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, Apr. 2003.

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

[16] 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, Jan. 2004.

[17] Z. Yong, R. Parthe, and D. Adjeroh, "Lossless compression of DNA microarray images," in *Proceedings of the IEEE Computational Systems Bioinformatics Conference*, Aug. 2005, pp. 128 – 132.

[18] R. Bierman, N. Maniyar, C. Parsons, and R. Singh, "MACE: lossless compression and analysis of microarray images," in *Proceedings of the ACM Symposium on Applied Computing, SAC*, 2006, pp. 167–172.

[19] A. Neekabadi, S. Samavi, S. A. Razavi, N. Karimi, and S. Shirani, "Lossless microarray image compression using region based predictors," in *Proceedings of the International Conference on Image Processing, ICIP*, 2007, pp. 349–352.

[20] S. Battiato and F. Rundo, "A bio-inspired CNN with re-indexing engine for lossless DNA microarray compression and segmentation," in *Proceedings of the International Conference on Image Processing, ICIP*, vol. 1-6. IEEE, 2009, pp. 1717–1720.

[21] A. J. R. Neves and A. J. Pinho, "Lossless Compression of Microarray Images Using Image-Dependent Finite-Context Models," *IEEE Transactions on Medical Imaging*, vol. 28, no. 2, pp. 194–201, Feb. 2009.

[22] 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.

[23] 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.

[24] M. Hernández-Cabronero, I. Blanes, M. W. Marcellin, and J. Serra-Sagristà, "Standard and specific compression techniques for DNA microarray images," *MDPI Algorithms*, vol. 4, pp. 30–49, 2012.

[25] 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, Feb. 2009.

[26] M. Hernández-Cabronero, J. Muñoz-Gómez, I. Blanes, M. W. Marcellin, and J. Serra-Sagristà, "DNA microarray image coding," in *Proceedings of the IEEE International Data Compression Conference, DCC*, 2012, pp. 32–41.

[27] M. Hernández-Cabronero, F. Aulí-Llinàs, J. Bartrina-Rapesta, I. Blanes, L. Jiménez-Rodríguez, M. W. Marcellin, J. Muñoz-Gómez, V. Sanchez, J. Serra-Sagristà, and Z. Xu, "Multicomponent compression of DNA microarray images," in *Proceedings of the CEDI Workshop on Multimedia Data Coding and Transmission, WMDCT*, 2012.

[28] M. Hernández-Cabronero, M. W. Marcellin, and J. Serra-Sagristà, "Compression of DNA Microarray Images," in *Microarray Image and Data Analysis: Theory and Practice*, L. Rueda, Ed. CRC Press, 2014, ch. 8, pp. 195–225.

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

[30] M. Hernández-Cabronero, V. Sanchez, M. W. Marcellin, and J. Serra-Sagristà, "A distortion metric for the lossy compression of DNA microarray images," in *Proceedings of the IEEE International Data Compression Conference, DCC*, IEEE, Ed., 2013, pp. 171–180.

[31] M. Hernández-Cabronero, I. Blanes, A. J. Pinho, M. W. Marcellin, and J. Serra-Sagristà, "Analysis-Driven Lossy Compression of DNA Microarray Images," *Submitted to the IEEE Transactions on Medical Imaging*.

[32] ——, "Progressive Lossy-to-Lossless Compression of DNA Microarray Images," *Submitted to the IEEE Signal Processing Letters*.

[33] Yeast Cell Cycle Analysis Project, "Image set: Yeast [Online]. Available: http://deic.uab.es/∼mhernandez/media/imagesets/yeast.tar.bz2."

[34] Speed Berkeley Research Group, "Image set: ApoA1 [Online]. Available: http://deic.uab.es/∼mhernandez/media/imagesets/apoa1.tar.bz2."

[35] SIB Computational Genomic Group, "Image set: ISREC [Online]. Available: http://deic.uab.es/∼mhernandez/media/imagesets/isrec.tar.bz2."

[36] Y. Luo, "Image set: MicroZip [Online]. Available: http://deic.uab.es/∼mhernandez/media/imagesets/microzip.tar.bz2."

[37] Stanford Microarray Database, "Image set: Stanford [Online]. Available: http://deic.uab.es/∼mhernandez/media/imagesets/stanford.tar.bz2."

[38] NCBI Gene Expression Omnibus, "Image set: Omnibus [Online]. Available: deic.uab.es/∼mhernandez/media/imagesets/omnibus.txt."

[39] David Galbraith Laboratory, "Image set: Arizona [Online]. Available: http://deic.uab.es/∼mhernandez/media/imagesets/arizona.tar.bz2."

[40] Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, "Image set: IBB [Online]. Available: http://deic.uab.es/∼mhernandez/media/imagesets/ibb.tar.bz2."

[41] Digital Image and Communications in Medicine (http://medical.nema.org/), "DICOM."

[42] A. J. Pinho, A. R. C. Paiva, and A. J. R. Neves, "On the use of standards for microarray lossless image compression." *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 563–6, Mar. 2006.

[43] M. J. Weinberger, G. Seroussi, and G. Sapiro, "The LOCO-I lossless image compression algorithm: principles and standardization into JPEG-LS." *IEEE Transactions on Image Processing*, vol. 9, no. 8, pp. 1309–24, Jan. 2000.

[44] *JPEG-LS*, ISO/IEC Std. IS 14 495-1, 14 495-2, 1998.

[45] A. Said and W. A. Pearlman, "A new fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, pp. 243–250, 1996.