Compresión de imágenes de DNA microarray basada en autosimilaridad

Manuel Ruibal Montoro, Universitat Autònoma de Barcelona

***Abstract*—Los experimentos DNA microarray generan miles de datos en forma de imágenes alrededor de todo mundo. Esta información tiene que ser eficientemente almacenada ya que es necesario guardarla hasta que las técnicas de investigación estén cien por ciento desarrolladas y estandarizadas. A partir del estudio de las características de las imágenes DNA microarray, este artículo describe un algoritmo de compresión de imagen sin pérdida basado en la autosimilaridad de las mismas. Se han generado tablas de datos a partir de experimentar con la codificación definida en este artículo. De los resultados obtenidos se observa que en determinados casos se logra obtener una mejoría de hasta casi 2 bpp en la compresión de una imagen codificada respecto a una sin codificar por el algoritmo basado en la autosimilaridad. En otros casos esta codificación significa hasta 3 bpp de empeoramiento en los resultados. En conclusión, con una técnica aún por mejorar se ha demostrado que utilizando la autosimilaridad de las imágenes DNA microarray para realizar una codificación es posible obtener mejores resultados de compresión.**

Keywords*—*Compresión de imágenes microarray; imágenes microarray DNA; codificación basada en autosimilaridad.

I. Introducción

A. DNA Microarrays

DNA microarray es efectiva herramienta, utilizada en el campo de la biomedicina, que permite analizar miles de genes en un único experimento. Un microarray es una matriz de miles de fragmentos de ADN ordenados en una superficie sólida siguiendo un determinado patrón. Gracias a esta tecnología podemos obtener información del genoma humano para la investigación de diferentes enfermedades. Esto es posible gracias a que cada experimento microarray compara el nivel de expresión de cada gen en dos células en un determinado momento.

La técnica se basa en marcar dos muestras de tejido (uno sano y otro canceroso por ejemplo), respectivamente, con dos marcadores fluorescentes llamados Cy3 y Cy5. Una máquina robótica ordena miles de secuencias de genes de estas muestras en un único portaobjeto. A continuación se utiliza un escáner especial para medir la intensidad fluorescente de cada spot. El resultado obtenido son dos imágenes en escala de grises donde, como vemos en la Figura 1, las diferentes intensidades entre el blanco y el negro reflejan la actividad de estos genes. Comparando la intensidad de los spots la imagen generada a partir de la muestra sana con la imagen generada a partir de la muestra con una enfermedad, se puede determinar que genes se expresan con distinta intensidad en cada caso. Estos resultados son utilizados para realizar diferentes hipótesis sobre la función de un gen en concreto bajo diferentes circunstancias.

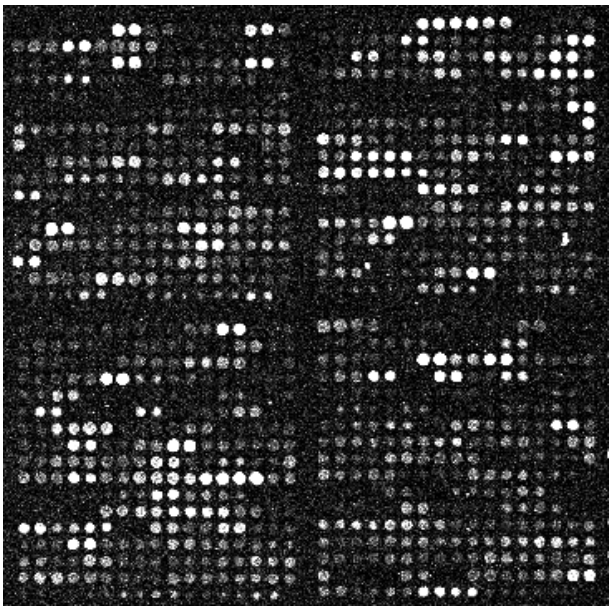

Figura 1: Recorte de una Imagen DNA microarray.

Las imágenes DNA microarray son analizadas por ordenador. Sin embargo, no solamente se requiere guardar esta información. Debido a que las técnicas no están desarrolladas ni estandarizadas completamente y debido a que repetir uno de estos experimentos puede ser demasiado costoso o imposible de realizar por el hecho de no poder repetir las características de una muestras, es necesario almacenar las imágenes DNA microarray. Las imágenes generadas varían en tamaños de 1000 x 1000 a más de 4000 x 13000, además ya que la expresión del gen puede variar en un amplio rango, la intensidad de cada píxel tiene un valor de 16 bits por píxel. Por lo tanto dependiendo del tamaño del microarray y la resolución especial del escáner, el tamaño de estas imágenes puede alcanzar los 100 Mbytes. La necesidad de almacenaje y compartición de este tipo de imágenes requiere una forma eficiente de compresión. Es importante tener en cuenta que es necesario guardar la totalidad de la información, por lo tanto es conveniente una compresión sin perdida, ya que cualquier diferencia podría alterar el resultado del análisis biológico.

La disposición de los spots en este tipo de imágenes depende del tipo de equipamiento utilizado para escanear el array. La Figura 1 es un ejemplo de una imagen DNA microarray, del corpus IBB, el cual se detallara más adelante en la sección

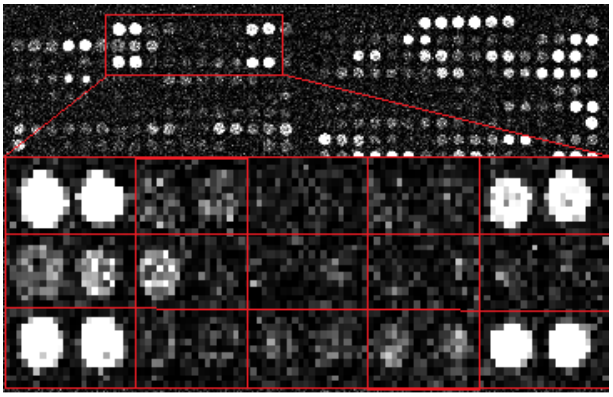

Figura 2: Porción de la Figura 1 remarcando diferentes parejas.

II. Se puede apreciar la estructura definida que tiene y se diferencian cuatro bloques de 17 spots de ancho y de alto.f En la Figura 2 se ha realizado un zoom remarcando diferentes parejas de spots para poder apreciar con mayor facilidad la similitud que presentan cada uno de los spots con su pareja correspondiente. Se observan 15 diferentes parejas de las cuales cada par de spot tiene una intensidad relativamente parecida. Este ejemplo ha sido seleccionado por la similitud presentada por las parejas, pero cabe la posibilidad de que los spots no tengan esta caracterásitica.

B. Compresión de imágenes de DNA microarrays

1) Estado del arte:: Igual que en la compresión de imágenes, la referente a imágenes DNA microarray se suele basar en 5 fases: preprocesamiento, transformación, cuantización, codificación por entropía y post-procesamiento. A continuación se detallara las diferentes técnicas utilizadas para abordar el problema en cada una de estas fases explicando en qué consiste la fase concretamente.

En el preprocesamiento de imágenes DNA microarray se pueden destacar principalmente dos técnicas aplicadas: la segmentación y la supresión de ruido. En cuanto a la segmentación, se han realizado diferentes propuestas. La segmentación, también llamada búsqueda de spots, consiste en determinar qué segmento de la imagen pertenece a la información de los spots representados en nuestras imágenes DNA microarray y qué parte es el fondo de la imagen. Los resultados de esta fase afectan directamente los resultados de las siguientes. En 2003 Faramarzpour et al. propusieron un codificador sin pérdida donde la segmentación del mismo consistía en localizar la región de los spots estudiando el periodo de la señal obtenida por la suma de las intensidades por filas y columnas y estudiando su mínima. A continuación estimando los centros de los spots en base al centroide de la región [6]. Luego en 2004, Lonardi y Luo presentaron su software de compresión MicroZip. Aquí utilizaban una variante de la idea de Faramarzpour, considerando la existencia de bloques de spots, las cuales se definen antes que las regiones de los spots [12]. También en 2006 Bierman et al. describieron un esquema de compresión sin perdida, con un simple método para dividir imágenes microarray en bajas y altas intensidades. Consistía en determinar los valores más bajos del umbral de 28, 29, 210 o 211 así aproximadamente el 90 por ciento de los píxeles caen dentro de él mismo [4].Finalmente en 2009, Battio y Rundo publicaron un propuesta basada en redes neuronales celulares (CNNs). Definieron dos capas para su sistema sin pérdida, cada una con la misma cantidad de células como píxeles tiene la imagen. La entrada y estado de la primera capa son los píxeles de la imagen original. El resultado de ésta es la entrada de la segunda capa. La segunda capa da como resultado una imagen binaria donde los píxeles de los spots tienden a 1 y los píxeles del fondo tienden a 0 [2].

En la fase de transformación se cambia el dominio de la imagen desde el dominio espacial a otro que pueda ser más eficientemente codificado. Los ejemplos más comunes son la aplicación de la DCT para obtener una representación de la frecuencia o usar una transformada wavelet para cambiar al dominio espacial-frecuencia. En el caso específico de las imágenes DNA microarray, la utilización de wavelet no consigue mejorar los resultados obtenidos con la aplicación de la DCT, por lo que no es utilizada frecuentemente. En 2004, Hua et al. publicaron una modificación del algoritmo EBCOT que incluía una transformación adaptada impar-simétrica para la propuesta de su esquema de compresión [8]. En 2004, Lonardi y Luo, utilizaron la transformada de Burrows-Wheeler [5] para su compresión en su software MicroZip. Hay varios casos de utilización de transformadas para a la supresión de ruido, los cuales no se detallaran debido a la perdida de información que significa para la imagen.

La cuantización consiste en dividir un conjunto de valores o vectores en grupos para reducir el total de símbolos necesarios para representarlos, incrementado la tasa de compresión, a expensas de introducir perdida de información. No hay muchas contribuciones a esta fase en la compresión de imágenes DNA microarray debido a la no acceptación de perdida de información. Igual que en la fase de supresión de ruido, no se entrará en más detalle ya que no se generaría compresión sin pérdida.

En cuanto a la codificación, en esta fase se quiere expresar la información obtenida anteriormente de una forma eficiente para generar un flujo de bits más compacto. Teniendo en cuenta que las imágenes DNA microarray tienen una gran regularidad espacial, muchas técnicas separan las imágenes en fondo y spots para codificarlos por separado. Otras técnicas tratan de predecir la intensidad del siguiente píxel basándose en el anterior. Este proyecto se centrará en esta fase para obtener mejores resultados de compresión a partir de una predicción específica.

Por último la fase de post-procesamiento consiste, generalmente, en tratar las imágenes comprimidas para mejorar su calidad o añadir nuevos atributos. Pero en el caso de las imágenes DNA microarray, normalmente no son tratadas sino analizadas para extraer su información genética y estudiar su contenido. Por esta razón, algunos investigadores han propuesto métricas de calidad específicas para las imágenes microarray DNA, las cuales se aplican después de segmentar la imagen en spots y fondo. En 2001 Wang et al. propusieron un índice de calidad

combinada (qcom), que consideraba los tamaños de los spots, relación señal-ruido, variabilidad del fondo y un fondo local excesivamente alto [15]. En 2004, Sauer et al. analizaron esta métrica y propusieron extenderla a dos nuevas métricas qcom1 y qcom2 [14]. Más tarde, en 2005, Battiato et al. definieron una métrica de calidad de segmentación de imágenes basada en la medida qcom2 [3]. También Pan-Gyu et al. definieron otra métrica de calidad considerando la señal, el ruido del fondo, la escala de invariación, la regularidad de los spots y el alineamiento de los spots [10].

Una vez realizado un rápido repaso del estado del arte en la compresión de imágenes DNA microarray, se puede remarcar que la utilización de la autosimilaridad no es un campo que se haya desarrollado. El método propuesto en este artículo exclusivo y desarrollado completamente a partir de los conocimientos de su autor y del tutor de este proyecto.

2) Objetivos: El objetivo principal de la compresión de imágenes digitales es reducir la cantidad de bytes necesarios para su representación. En este proyecto se realizará una aplicación específica para la compresión sin perdida de imágenes DNA microarray, con el objetivo de obtener una mejora en la tasa de compresión. Esta mejoría está basada en la autosimilaridad de las imágenes DNA microarray, ya que estas imágenes presentan una estructura definida y repetida en los diferentes ejemplares. Y es a partir de esta similitud que se busca realizar una comparación o predicción para reducir entropía de la imagen, facilitando a su vez la compresión. Se comparará los resultados de compresión, antes y después de realizar la codificación propuesta, a partir de tres diferentes códecs: JPEG2000[1], JPEG-LS [9] y Neves [13].

C. Estructura del artículo

A continuación, la sección II especificará mejor qué son las imágenes DNA microarray y sus características . En tercer lugar en la sección III, se expondrá, justificadamente, el trabajo realizado explicando detalladamente la hipótesis desarrollada (III-A), la codificación y decodificación (III-B), la compresión y descompresión (III-C) realizadas. Seguidamente en la Sección IV se mostrarán y analizarán los resultados experimentales obtenidos a partir de la aplicación desarrollada en este proyecto. La sección V detallará las conclusiones pertinentes obtenidas a partir de los resultados expresados en la sección anterior.

II. Imágenes de DNA microarrays

De cada experimento DNA microarray se obtiene como resultado dos imágenes TIFF que contienen la información que nos proporcionan los colorantes fluorescentes rojo y verde correspondiente a cada una de dos muestras utilizadas. Cada imagen DNA microarray es una representación del microarray escaneado y se pueden apreciar varios bloques de spots, con más o menos intensidad, ordenados en filas y columnas. La profundidad de cada píxel de estas imágenes es de 16 bits basándose en el rango de intensidad que puede adoptar cada uno de los genes. Por tanto el valor analógico pasado a digital de cada píxel está en un rango de 0 a $2^{16} - 1$ (65535).

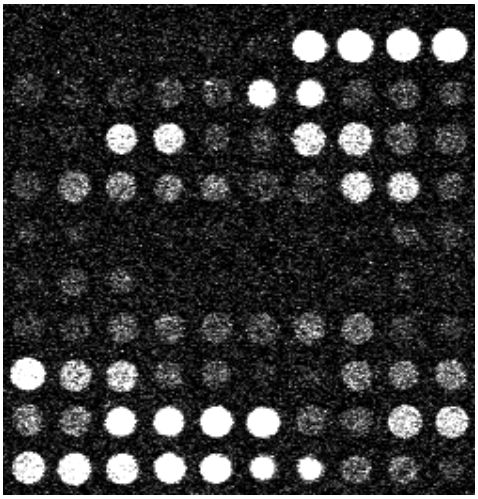

Figura 3: Porción de imagen DNA microarray 10x10 spots

En la Figura 3 se puede apreciar de forma clara la estructuración del microarray dentro de la imagen. Como se observa, la intensidad de los diferentes spots siguen un patrón o diseño ya que al pertenecer a un mismo gen tienden a tener el mismo grado de intensidad. Los spots de un mismo gen están dispuestos secuencialmente uno al lado del otro de dos en dos.

Diferentes instituciones alrededor del mundo están investigando y obtienen diferentes resultados en el campo de los DNA microarray. Actualmente se pueden distinguir una serie de conjuntos de imágenes DNA microarray de diferentes características provenientes de diferentes centros de investigación. Cada conjunto de imágenes se adapta diferentemente a los distintos métodos de compresión. En la Tabla I se describen las diferentes propiedades de los conjunto de imágenes más importantes y el utilizado en este proyecto. La aplicación desarrollada trabajará con el corpus de imágenes IBB (Institut de Biotecnologia i Biomedicina), proveniente del Servei de Genòmica de la Universidad Autónoma de Barcelona. Esto es debido a la regularidad de ordenación que presentan los bloques de este conjunto de imágenes y sobretodo, a que de este corpus se poseen los archivo GPR para cada pareja de imágenes generada por los experimentos DNA microarray. Estos archivos GPR serán utilizados durante el desarrollo de este proyecto ya que contienen diferente información sobre todos y cada uno de los spots representados en las imágenes. Información necesaria para saber donde se encuentran los spots dentro de la imagen. Por tanto no es posible aplicar la técnica desarrollada en este artículo para otros conjuntos de imágenes DNA microarrray. Estos archivos GPR contienen información como el número de bloque en la imagen al que pertenece cada spot, el número de fila y de columna, el nombre por el cual se identifica ese gen, la posición en la imagen, el diámetro, y más información biológica sobre el gen concreto que representa el spot. Los archivos GPR son valorados especialmente debido a que no están disponibles para otros corpus y la información

Tabla I: Conjuntos de imágenes ordenados por año

Propiedad	**Yeast**	**Stanford**	**ApoA1**	**ISREC**	**MicroZip**	**IBB**	**Arizona**
Año	1998	2001	2001	2001	2004	2010	2011
Imágenes	109	20	32	14	3	44	6
Tamaño	1024x1024	>2000x2000	1044x1041	1000x1000	>1800x1900	2019x6235	4400x13800
Disposición de spots	cuadrado	cuadrado	cuadrado	cuadrado	cuadrado	cuadrado	hexagonal
Cantidad de spots	9.10^3	4.10^3	$\sim 6.10^3$	$\sim 2.10^2$	$\sim 9.10^3$	$\sim 1,4.10^4$	2.10^5
Intensidad media	$5,39\%$	$28,83\%$	$39,51\%$	$33,34\%$	$37,71\%$	$6,09\%$	$82,82\%$
Entropía media	6,628	8,293	11,033	10,435	9,831	8,50	9,306

que contienen no se puede obtener sin acceso a los scanners originales.

III. Compresión basada en autosimilaridad

A. Hipótesis

Teniendo en cuenta el objetivo principal del proyecto de realizar una aplicación capaz de comprimir las imágenes DNA microarray y mejorar la tasa de compresión, se ha formulado la hipótesis de que codificando estas imágenes teniendo en cuenta su autosimilaridad se podría conseguir buenos resultados en cuanto a tasa de compresión. Esto está basado en que, las imágenes DNA microarray presentan una estructura bastante definida, dependiendo del corpus que utilicemos, en bloques, filas y columnas y se aprecia en las imágenes que los spots son una repetición con mayor o menor variación de intensidad.

Hipótesis: La tasa de compresión de las imágenes DNA microarray puede ser mejorada codificando estas imágenes a partir de una predicción basada en la autosimilaridad de las mismas.

En la Figura 4 se puede ver, ya que los spots pertenecen al mismo gen, que de dos en dos secuencialmente estos son significativamente parecidos. Un spot se asemeja a su pareja en forma, intensidad y posición. Por lo tanto, los spots de un mismo gen tienden a tener un diámetro del mismo tamaño y una intensidad parecida.

La hipótesis plantea que a la hora de codificar, si utilizamos el primer spots de una pareja de un mismo gen para predecir el segundo, se acertará o se aproximará lo suficiente al resultado como para obtener una reducción de la variación de la imagen y por lo tanto un flujo de bits mejor adaptado a la compresión. Se quiere explotar la autosimilaridad de las imágenes DNA microarray para que los resultados de la compresión este tipo de imágenes se más eficiente.

B. Codificación

Para comenzar se dividió el proceso de compresión en varias fases. Después de cargar la imagen en una matriz con los valores de cada píxel de la imagen, la aplicación, a partir del archivo GPR, procede a segmentar la imagen guardando la información de la posición del centro de los spots en la imagen, con las variables “X” y “Y”, el diámetro de cada spot y la variable ID que identifica el gen al que pertenece ese spot. Será a partir de esta información que se podrá localizar los diferentes spots y decidir cuales interesa ser comparados.

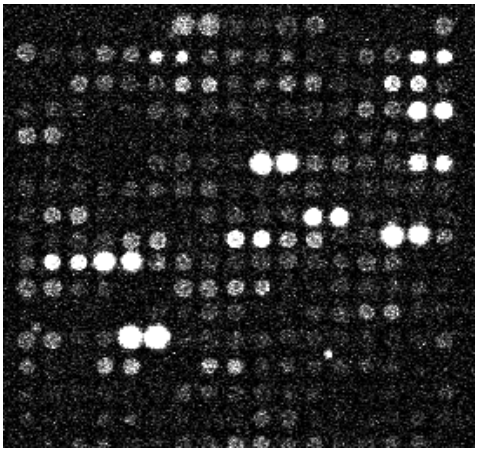

Figura 4: Porción de imagen DNA microarray original 1 del corpus IBB.

De cada centro de spot localizado se define un rectángulo del tamaño del diámetro máximo de todos los spots de la imagen. Como se ha investigado y estudiado la autosimilaridad, se pretende que los spots de dos en dos secuencialmente, pertenecientes al mismo gen, sean lo más parecidos posible y así, a partir del primero, predecir el segundo de una forma acertada. La predicción se realiza píxel a píxel en la misma posición de los rectángulos definidos para cada spot. Por lo tanto, a continuación, recorriendo la imagen de bloque en bloque y verificando si cada pareja de spot pertenecen a un mismo gen, se codifica cada rectángulo de arriba a abajo, de izquierda a derecha y píxel a píxel de forma que el segundo rectángulo de cada pareja de spots, es igual al resultado de la resta del primer rectángulo menos el segundo. Este valor, en la mayoría de los casos, debería ser bajo y por lo tanto generar un flujo de bits más fácilmente compresible. Teniendo en cuenta que cada bloque tiene un número impar de spots, exactamente 17, el último spot, como todos los primeros de cada pareja, se deja intacto en la imagen (ver Figura 5). El resultado de este proceso de codificación es una imagen basada en la imagen original con las modificaciones realizadas según el algoritmo.

Además al realizar esta resta entre los valores de los píxeles, se debe tener en cuenta la posibilidad de obtener un resultado entre -65535 y 65535, por lo tanto no puede ser representado en 16 bits. Debido a que según el formato de

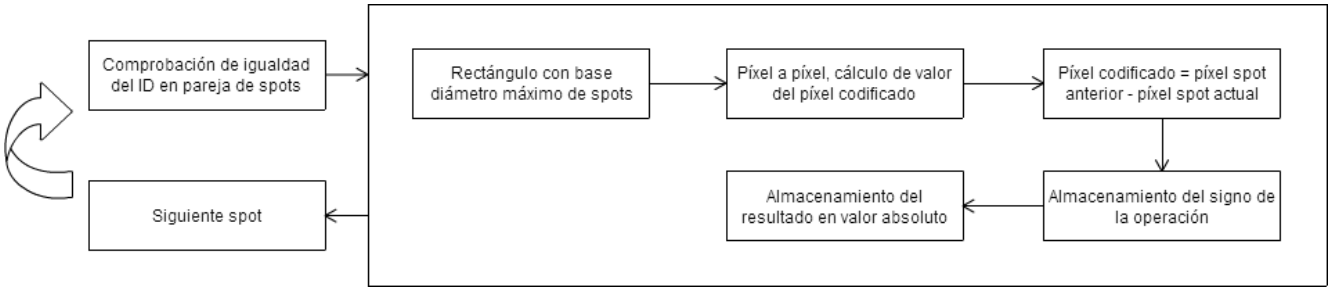

Figura 6: Diagrama de codificación

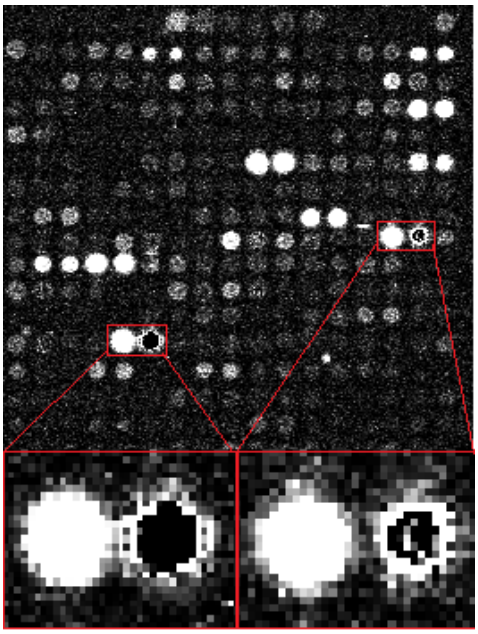

Figura 5: Porción de imagen DNA microarray códificada 1 del corpus IBB remarcando codificación realizada.

la imagen, no se permite almacenar números negativos, se ha decidido guardar los resultados en valores absolutos dentro de la matriz correspondiente a la imagen y, en un archivo externo, cada signo del resultado de la operación, obteniendo así una cadena de 0 (negativo) y 1 (positivo). Esto se lleva a cabo para poder recuperar exactamente el valor de cada píxel a la hora de realizar la decodificación. En total este proceso genera dos archivos, la imagen codificada y el archivo de signos, además es necesario almacenar el archivo GPR para el proceso de decodificación. Estas imágenes codificadas serán las comprimidas por los diferentes métodos de compresión en la siguientes fases de la aplicación.

En la Figura 4 y la Figura 5 se puede apreciar, a a gran escala, el resultado de la codificación respecto a una imagen original. Como se puede observar, en las parejas de spots que pertenecen al mismo gen, el segundo spot es el resultado de la resta del primero menos el segundo, por lo tanto al ser similares queda en un valor aproximado a 0 (color negro). En la Figura 5 se ha realizado un zoom en los spots donde se puede apreciar la diferencia con mayor nitidez.

Para mayor entendimiento del proceso seguido por la codificación de la aplicación se ha añadido la Figura 6 con un diagrama donde se ilustra paso a paso el algoritmo anteriormente descrito y utilizado por la implementación. Se inicia el proceso comprobando si dos spots pertenecen a un mismo gen y se finaliza al acabarse los spots en la imagen. Una vez finalizado se procede a guardar en el disco la imagen codificada junto a su archivo de signos, dando por supuesto que el archivo GPR ya esta guardado ya que se ha sacado de este la información para la codificación..

Este proceso de codificación se puede revertir completamente obteniendo la imagen original sin ningún tipos de perdida de información. Lo que hace la decodificación es invertir los pasos realizados durante la codificación para recuperar la imagen original a partir de los datos guardados. A partir de la imagen codificada, se carga en memoria una matriz con los valores de los píxeles de la misma. Y se recorre, inversamente que en la codificación, bloque a bloque buscando, gracias al archivo GPR, cada rectángulo y verificando que su pareja sea de la misma ID. Cada vez que coinciden una pareja de spots, el segundo spot será un spot anteriormente codificado, por lo tanto se transforma el valor absoluto almacenado al valor original según el signo guardado en el archivo externo de signos. Es por esto que se tienen en cuenta los 3 archivos (imagen, GPR y signos) a la hora de calcular el peso total necesario para la recuperación de la imagen original.

Se han tenido en cuenta otras formas de codificación pero que finalmente fueron descartadas. Por ejemplo, utilizando otras técnicas se podría realizar un emparejamiento para que la similitud de los spots en una pareja sea máxima y así mejorar la predicción o incluso se podrían realizar diferentes formas de agrupación, cambiando la cantidad de spot agrupados.

C. Compresión

Los compresores utilizados son: Kakadu 7.4 [1], una implementación completa de JPEG2000 estándar fuertemente optimizada; LOCO-I [9], un software que implementa el estándar JPEG-LS y Neves [13], un método de compresión sin perdida basado en una codificación aritmética a partir de los planos de

bits, el cual ha demostrado ser actualmente el mejor compresor sin pérdida para las imágenes DNA microarray. En el caso de JPEG2000 se han probado dos configuraciones de parámetros diferentes. La primera es una configuración estándar de la aplicación Kakadu con 5 niveles de decomposición wavelet y la segunda corresponde al método HST/JPEG2000 [7] con 0 niveles de decomposición wavelet. Se han elegido los códec estándar de compresión JPEG2000 y JPEG-LS debido a que son diferentes algoritmos de compresión sin perdida y tienen gran reconocimiento internacional, por lo tanto los resultados obtenidos en este artículo son fácilmente comparables con otros resultados en la investigación de la compresión de imagen. Por otro lado, se ha decidido utilizar la implementación Neves por el hecho de ser un algoritmo de compresión sin perdida de información específicamente diseñado para la compresión de imágenes DNA microarray y por ser el algoritmo que obtiene mejores resultados de compresión con las imágenes originales del corpus utilizado en este proyecto.

A su vez, para el ahorro de memoria y por lo tanto una mayor tasa de compresión total, también se comprimen los archivos externos de signos y los archivos GPR con un algoritmo genérico de compresión de datos LZMA [11]. Se ha decidido realizar la compresión de estos dos archivos con el algoritmo estándar LZMA, ya que se adapta con buenos resultados a muchos formatos de archivos y este proyecto está más enfocado a la compresión de imágenes y no tanto los archivos externos. El resultado de esta compresión son dos de archivos en formato Zipx por cada imágen. A continuación en los resultados experimentales se tendrá en cuenta estos archivos externos como parte del total de bytes necesarios para la total compresión y descompresión sin perdida de información de una imagen DNA microarray. A la hora de descomprimir la aplicación realiza una llamada al descompresor correspondiente del Kakadu, LOCO-I o Neves y guarda la imagen aún codificada en el disco. El siguiente paso es realizar la decodificación que devolverá la imagen original sin perdida de información.

IV. Resultados experimentales

Una vez finalizada la aplicación se ha procedido a realizar diferentes pruebas de compresión, para comprobar cuál era el resultado obtenido a partir de la codificación basada en la autosimilaridad descrita en este artículo. Además, era necesario saber si el principal objetivo de este proyecto de desarrollar una implementación de compresión sin perdida que superarse las tasas de compresión actuales, se había cumplido.

Las pruebas realizadas han consistido, primero, en comprimir todas las imágenes DNA microarray originales del corpus IBB con una serie de algoritmos y en segundo lugar, comprimirlas una vez codificadas con la técnica propuesta en este artículo. La compresión ha sido realizada a partir de dos estándares de compresión, JPEG-LS y JPEG2000 y a partir del algoritmo de compresión Neves. Una vez acabadas las pruebas, se ha presentado en una tabla una comparación de de los bits por píxel (bpp) resultantes en cada caso. El cálculo de este valor se realiza dividiendo la cantidad de bits que representa la imagen entre la cantidad de píxeles que tiene la imagen. Por lo tanto, cuanto más bajo el valor bpp mejor es el resultado en cuanto a compresión. En el caso de las imágenes codificadas se han realizado dos columnas en la tabla. La primera contiene los bpp que representa la imagen codificada por separado y la segunda los bpp teniendo en cuenta todos los bits necesarios para recuperar el 100 por ciento de la imagen original, por lo tanto teniendo en cuenta los archivos adicionales almacenados (archivo GPR y archivo de signos). Se ha dispuesto de esta forma debido a que este proyecto se ha centrado en mejorar la compresión de imágenes concretamente y no toda la información almacenada en estos archivos externos es necesaria para la decodificación. Estos archivos fueron comprimidos por un algoritmo estándar (LZMA), pero esta compresión no está optimizada.

A su vez se ha realizado una tabla con un estudio del porcentaje que representa cada uno de los tres archivos dentro del total de bits por píxel de las imágenes codificadas y comprimidas. Para poder comparar el valor que representan cada uno de los tres archivos necesarios para la decodificación.

En la Tabla II se presentan los resultados de bpp obtenidos a partir de la compresión de las imágenes originales, las imágenes codificadas y las imágenes codificadas con sus archivos externos. Para poder realizar una comparación se ha decidido generar los resultados obtenidos por los algoritmos antes de realizar la codificación descrita, obteniendo una media de 9,84 bpp con el algoritmo JPEG-LS, 9,07 bpp con el algoritmo JPEG2000, 10,44 bpp con el algoritmo HST/JPEG2000 y 7,86 bpp con el algoritmo Neves en las imágenes originales.

Analizando la tabla, se puede ver que los resultados son positivos para un total de 17 imágenes de las 44 iniciales sin tener en cuenta los archivos externos y un total, según el algoritmo, de entre 8 y 10 imágenes teniendo en cuenta los archivos externos. Este es el caso de los algoritmos de compresión JPEG-LS, JPEG2000 y HST/JPEG2000 pero no para el algoritmo Neves, donde no se ha podido mejorar los resultados en ninguna de las imágenes. De las 17 imágenes que se logra mejorar los resultados hay un máximo de 2,88 bpp de mejoría en la imagen 25.

Por otra parte, la media total de bpp en caso de JPEG-LS es de 10,03 sin tener en cuenta los archivos externos y 10,85 teniéndolos en cuenta. En el caso de JPEG2000 la media total de bpp es de 9,49 sin tener en cuenta los archivos externos y un 10,29 teniéndolos en cuenta. Por último en el caso de HST/JPEG2000 la media total de bpp es de 10,45 sin tener en cuenta y 11,26 teniendo en cuenta los archivos externos. Globalmente no se mejoran los resultados a partir del algoritmo basado en la autosimilaridad, de hecho empeora los resultados en algunos de los casos en más de 3 bpp totales. En el caso de Neves, los resultados obteniendo incrementan entre 0,03 y 0,36 bpp de las imágenes codificadas comprimidas por los códecs de compresión, obteniendo una media de 8,06 sin tener en cuenta los archivos externos y 8,87 teniéndolos en cuenta.

Analizando estos resultados se puede decir que siendo Neves un algoritmo específicamente diseñado para la compresión de este tipo de imágenes, es más complicado que una modificación de las imágenes DNA microarray pueda mejorar los resultados. En el caso de los otros tres métodos, teniendo en cuenta que están diseñados para cualquier tipo de imagen, la

Tabla II: Resultados en bpp de la compresión de las imágenes antes (a) y después (b) de la codificación y bpp de las imágenes codificadas teniendo en cuenta los archivos externos (c). Se plantean tres columnas para cada uno de los códecs utilizados, JPEG-LS, JPEG2000, HST/JPEG2000 y Neves.

	JPEG-LS			JPEG2000			HST/JPEG2000			Neves		
Imagen	**(a)**	**(b)**	**(c)**	**(a)**	**(b)**	**(c)**	**(a)**	**(b)**	**(c)**	**(a)**	**(b)**	**(c)**
1	9,72	9,97	10,77	8,44	9,13	9,93	10,94	10,97	11,77	6,82	7,13	7,94
2	10,88	10,76	11,57	10,35	10,05	10,86	11,12	11,00	11,80	9,48	9,55	10,36
3	9,95	10,45	11,23	8,72	9,86	10,63	10,57	10,69	11,47	7,36	7,63	8,40
4	9,90	10,74	11,53	8,68	10,28	11,06	10,56	10,89	11,67	7,30	7,58	8,36
5	9,80	10,59	11,38	8,53	10,07	10,85	10,51	10,75	11,53	7,12	7,40	8,18
6	9,91	10,32	11,11	8,69	9,73	10,51	10,56	10,72	11,50	7,31	7,56	8,35
7	11,40	9,17	10,00	11,14	8,66	9,49	11,64	10,72	11,54	10,40	10,42	11,25
8	9,78	8,45	9,27	8,96	8,29	9,11	10,52	10,52	11,34	7,51	7,71	8,53
9	11,05	10,07	10,87	10,53	9,72	10,51	11,32	10,32	11,11	9,63	9,71	10,51
10	10,61	9,75	10,53	9,76	9,29	10,07	11,00	10,04	10,82	8,62	8,79	9,58
11	10,26	9,83	10,69	9,55	9,39	10,24	10,66	10,10	10,95	8,51	8,68	9,53
12	10,57	11,05	11,88	10,19	10,43	11,25	10,90	11,12	11,94	9,25	9,33	10,15
13	10,10	9,06	9,88	9,34	8,66	9,48	10,68	9,70	10,52	8,95	9,05	9,87
14	10,43	11,08	11,91	9,92	10,63	11,45	10,75	11,24	12,06	8,04	8,26	9,07
15	8,86	9,96	10,75	7,74	9,59	10,38	10,67	10,37	11,15	6,04	6,40	7,19
16	8,08	10,95	11,72	7,21	10,48	11,25	10,47	11,14	11,92	5,46	5,81	6,59
17	9,85	10,47	11,33	9,53	10,04	10,89	10,28	10,73	11,57	8,53	8,66	9,51
18	9,58	10,95	11,77	9,04	10,48	11,30	10,04	11,13	11,95	7,96	8,07	8,88
19	8,86	11,11	11,94	8,00	10,61	11,43	9,65	11,31	12,13	8,14	8,24	9,06
20	9,67	10,11	10,93	9,20	9,39	10,21	10,11	10,36	11,17	6,82	7,07	7,89
21	10,83	10,40	11,29	10,51	9,79	10,66	11,21	10,56	11,44	9,43	9,54	10,42
22	9,72	9,40	10,28	9,25	8,85	9,72	10,31	9,68	10,55	7,98	8,18	9,05
23	10,75	10,13	10,98	10,35	9,27	10,11	11,13	10,59	11,43	9,27	9,38	10,22
24	10,26	9,48	10,33	9,87	8,94	9,78	10,71	9,80	10,64	8,71	8,83	9,66
25	10,71	7,83	8,75	10,31	7,68	8,59	11,09	9,42	10,32	9,25	9,39	10,29
26	10,85	9,30	10,21	10,45	8,71	9,61	11,27	9,97	10,87	9,28	9,42	10,31
27	10,84	9,34	10,13	10,55	8,79	9,58	11,19	9,56	10,35	7,91	8,11	8,90
28	10,93	9,58	10,38	10,71	9,09	9,88	11,29	9,78	10,57	8,66	8,77	9,56
29	9,96	10,13	10,90	9,03	9,43	10,20	10,35	10,41	11,18	7,45	7,63	8,40
30	10,25	10,43	11,21	9,58	9,82	10,59	10,56	10,62	11,39	7,49	7,66	8,43
31	9,27	9,76	10,52	8,48	9,30	10,06	9,67	9,97	10,72	4,89	5,22	5,97
32	9,34	9,65	10,41	8,57	9,14	9,90	9,79	9,92	10,68	6,50	6,82	7,58
33	7,54	9,96	10,69	6,39	9,25	9,99	9,36	10,19	10,92	7,51	7,67	8,40
34	9,11	10,08	10,81	7,89	9,23	9,96	9,95	10,58	11,31	7,96	8,09	8,82
35	9,23	10,23	11,03	8,45	9,62	10,42	9,55	10,38	11,18	7,86	8,06	8,86
36	9,45	9,28	10,08	8,86	8,74	9,53	9,78	9,51	10,31	8,58	8,71	9,51
37	9,97	9,17	9,96	9,04	8,61	9,40	10,40	9,48	10,27	8,19	8,30	9,08
38	10,27	11,05	11,84	9,58	10,62	11,40	10,61	11,19	11,98	7,80	7,95	8,73
39	9,61	11,14	11,90	9,13	10,74	11,50	9,97	11,29	12,05	7,71	7,92	8,67
40	9,50	9,98	10,74	8,85	9,12	9,88	9,92	10,53	11,29	8,43	8,57	9,33
41	9,82	10,09	10,84	8,83	9,22	9,97	10,18	10,58	11,32	7,35	7,53	8,28
42	10,09	11,58	12,33	9,37	11,15	11,89	10,38	11,63	12,38	7,02	7,24	7,98
43	9,17	9,99	10,84	8,34	9,38	10,22	9,51	10,53	11,38	9,57	9,65	10,49
44	9,06	11,22	12,07	8,11	10,64	11,48	9,47	11,32	12,16	9,69	9,74	10,59
Media	9,84	10,03	10,85	9,07	9,49	10,29	10,44	10,45	11,26	7,86	8,06	8,87

modificación de algunas de las imágenes DNA microarray ha concluido con una mejora de los bpp debido a que en algunos casos la similitud de los spots de las imágenes era mayor que en otros, resultando en un flujo de bits más compacto para los diferentes compresores.

En la Tabla III se puede observar los diferentes porcentajes que ocupan cada uno de los tres archivos en el total de bytes ocupados en memoria para la recuperación de la imagen original, a partir de la imagen codificada y los archivos externos necesarios. Se puede apreciar, comparando los resultados obtenidos por los 4 métodos de compresión utilizados, que el porcentaje del espacio ocupado por una imagen tiene un máximo de 93.97, valor alcanzado por el compresor que utiliza el algoritmo HST/JPEG2000. En este caso el archivo de signos representa un 1.06 por ciento y el archivo GPR representa un 4.97 por ciento. A su vez la media del espacio ocupado por la imagen, en este método de compresión, es también el más elevado con un 92.96 por ciento. Esto significa que algoritmo HST/JPEG2000 es el que peor se adapta a las imágenes DNA microarray codificadas, ya que estas representan un mayor volumen proporcional respecto al resto de archivos. En cuanto al mínimo porcentaje de peso que significa una imagen DNA microarray codificada y comprimida es de 87.38, este valor es alcanzado por el compresor Neves. En este caso el porcentaje que representa el archivo GPR es de 10.61 y el archivo de signos es de 2.02 por ciento. Este método de compresión tiene la media más baja de lo que representa la imagen codificada con un 91.12 por ciento. Esto significa que este compresor es, como se esperaba, el que alcanza una mayor tasa de compresión con las imágenes DNA microarray. Analizando

Tabla III: Porcentajes correspondientes a al peso total generado al comprimir las imágenes DNA microarray codificadas por los códecs JPEG-LS, JPEG2000, HST/JPEG2000 y Neves. Los porcentajes corresponden a la imagen (a), el archivo de signos (b) y el archivo GPR (c).

	JPEG-LS			JPEG2000			HST/JPEG2000			Neves		
	Imagen	**Signos**	**GPR**	**Imagen**	**Signos**	**GPR**	**imagen**	**Signos**	**GPR**	**Imagen**	**Signos**	**GPR**
1	92, 52	1, 20	6, 28	91, 91	1, 27	6, 81	93, 18	1, 07	5, 75	89, 8	1, 59	8, 53
2	92, 97	1, 18	5, 85	92, 56	1, 21	6, 23	93, 15	1, 11	5, 73	92, 20	1, 27	6, 53
3	93, 03	1, 19	5, 78	92, 68	1, 21	6, 11	93, 21	1, 12	5, 66	90, 74	1, 53	7, 73
4	93, 21	1, 16	5, 63	92, 97	1, 16	5, 87	93, 33	1, 10	5, 57	90, 70	1, 53	7, 77
5	93, 10	1, 17	5, 73	92, 81	1, 19	6, 00	93, 23	1, 12	5, 65	90, 46	1, 58	7, 96
6	92, 93	1, 20	5, 86	92, 58	1, 23	6, 20	93, 21	1, 12	5, 66	90, 65	1, 54	7, 81
7	91, 70	1, 37	6, 93	91, 31	1, 39	7, 31	92, 86	1, 14	6, 00	92, 67	1, 17	6, 16
8	91, 08	1, 45	7, 47	91, 00	1, 39	7, 61	92, 77	1, 12	6, 11	90, 39	1, 49	8, 12
9	92, 65	1, 25	6, 10	92, 45	1, 24	6, 31	92, 85	1, 18	5, 97	92, 44	1, 24	6, 31
10	92, 51	1, 28	6, 21	92, 20	1, 31	6, 49	92, 74	1, 22	6, 04	91, 79	1, 38	6, 83
11	91, 99	1, 27	6, 75	91, 73	1, 23	7, 04	92, 26	1, 15	6, 58	91, 11	1, 33	7, 57
12	93, 06	1, 15	5, 79	92, 72	1, 17	6, 11	93, 14	1, 11	5, 76	91, 93	1, 30	6, 77
13	91, 64	1, 38	6, 98	91, 33	1, 39	7, 27	92, 19	1, 26	6, 56	91, 68	1, 34	6, 99
14	93, 03	1, 13	5, 84	92, 85	1, 07	6, 08	93, 22	1, 01	5, 77	90, 99	1, 35	7, 67
15	92, 67	1, 16	6, 17	92, 43	1, 18	6, 39	92, 96	1, 10	5, 95	89, 07	1, 70	9, 22
16	93, 39	1, 03	5, 58	93, 13	1, 06	5, 81	93, 51	1, 00	5, 49	88, 26	1, 81	9, 93
17	92, 44	1, 20	6, 36	92, 26	1, 12	6, 62	92, 72	1, 05	6, 23	91, 14	1, 28	7, 58
18	93, 01	1, 16	5, 84	92, 77	1, 15	6, 08	93, 16	1, 09	5, 75	90, 80	1, 47	7, 74
19	93, 08	1, 14	5, 78	92, 82	1, 14	6, 04	93, 24	1, 07	5, 69	90, 95	1, 44	7, 61
20	92, 42	1, 21	6, 36	91, 98	1, 20	6, 81	92, 67	1, 10	6, 23	89, 63	1, 56	8, 81
21	92, 15	1, 20	6, 65	91, 80	1, 16	7, 04	92, 36	1, 08	6, 56	91, 61	1, 19	7, 20
22	91, 39	1, 31	7, 30	91, 05	1, 23	7, 72	91, 75	1, 13	7, 11	90, 39	1, 32	8, 29
23	92, 26	1, 24	6, 50	91, 68	1, 26	7, 06	92, 64	1, 11	6, 25	91, 77	1, 24	6, 99
24	91, 77	1, 32	6, 91	91, 44	1, 26	7, 30	92, 13	1, 16	6, 71	91, 33	1, 28	7, 39
25	89, 55	1, 55	8, 90	89, 48	1, 45	9, 06	91, 25	1, 21	7, 54	91, 22	1, 21	7, 56
26	91, 05	1, 33	7, 62	90, 66	1, 24	8, 10	91, 75	1, 10	7, 16	91, 30	1, 15	7, 55
27	92, 17	1, 34	6, 49	91, 76	1, 37	6, 86	92, 37	1, 27	6, 36	91, 13	1, 48	7, 39
28	92, 35	1, 31	6, 34	92, 01	1, 34	6, 66	92, 53	1, 25	6, 22	91, 74	1, 38	6, 88
29	92, 89	1, 24	5, 87	92, 44	1, 29	6, 27	93, 10	1, 18	5, 72	90, 82	1, 57	7, 62
30	93, 08	1, 21	5, 71	92, 73	1, 23	6, 04	93, 24	1, 14	5, 62	90, 87	1, 54	7, 59
31	92, 81	1, 16	6, 03	92, 50	1, 20	6, 30	92, 96	1, 12	5, 91	87, 38	2, 02	10, 61
32	92, 66	1, 26	6, 09	92, 31	1, 29	6, 40	92, 87	1, 19	5, 93	89, 96	1, 68	8, 36
33	93, 14	1, 27	5, 59	92, 68	1, 34	5, 98	93, 31	1, 22	5, 47	91, 30	1, 59	7, 11
34	93, 22	1, 26	5, 52	92, 67	1, 33	6, 00	93, 54	1, 17	5, 28	91, 72	1, 51	6, 78
35	92, 75	1, 22	6, 03	92, 37	1, 26	6, 38	92, 89	1, 17	5, 94	91, 02	1, 48	7, 50
36	92, 06	1, 35	6, 59	91, 65	1, 38	6, 97	92, 28	1, 28	6, 45	91, 63	1, 38	6, 99
37	92, 06	1, 37	6, 57	91, 63	1, 41	6, 96	92, 33	1, 29	6, 38	91, 33	1, 46	7, 21
38	93, 33	1, 14	5, 53	93, 10	1, 16	5, 74	93, 43	1, 10	5, 47	90, 99	1, 51	7, 50
39	93, 62	1, 13	5, 25	93, 42	1, 15	5, 43	93, 72	1, 10	5, 18	91, 28	1, 52	7, 20
40	92, 92	1, 27	5, 82	92, 33	1, 34	6, 32	93, 29	1, 18	5, 53	91, 88	1, 42	6, 69
41	93, 08	1, 25	5, 67	92, 50	1, 33	6, 17	93, 40	1, 17	5, 43	90, 97	1, 60	7, 42
42	93, 92	1, 09	4, 99	93, 72	1, 11	5, 17	93, 97	1, 06	4, 97	90, 65	1, 65	7, 70
43	92, 12	1, 26	6, 62	91, 72	1, 27	7, 02	92, 56	1, 14	6, 30	91, 93	1, 23	6, 83
44	92, 93	1, 13	5, 94	92, 64	1, 11	6, 25	93, 05	1, 05	5, 90	92, 02	1, 20	6, 77
Media total	92, 71	1, 23	6, 06	92, 40	1, 24	6, 35	92, 96	1, 12	5, 92	91, 12	1, 46	7, 52

estos resultados se puede observar que ,a diferencia de los archivos de signos almacenados que ocupan entre un 1.06 y un 2.02 por ciento del total, los valores de los archivos GPR son bastante significativos ya que oscilan entre 4.97 y 10.61. Esto se debe a que este proyecto se ha centrado en buscar una mejoría en la compresión concreta de las imágenes y estos archivos externos no se han llegado a optimizar. Teniendo en cuenta que hay información no utilizada que se esta comprimiendo en estos dos archivos externos, la optimización de la compresión de estos significaría una diferencia importante en la tasa de compresión. Aunque para mejorar los resultados de los codificadores originales, haría falta mejorar la codificación y la compresión de las imágenes codificadas.

V. CONCLUSIONES

En una industria en pleno crecimiento como es la biomedicina, y especialmente la relacionada con la investigación del genoma humano es necesario aportar las herramientas necesarias para que los diferentes profesionales del sector puedan tanto almacenar como compartir sus conocimientos. El análisis de las imágenes DNA microarray está aún en desarrollo y la repetición de algunos de estos experimentos es difícil o imposible. Es por esto que es necesario tener una forma eficiente de almacenar las imágenes resultantes de cada experimento.

En este artículo se presenta un algoritmo de codificación y compresión sin perdida basado en la autosimilaridad de las imágenes DNA microarray. Este algoritmo utiliza los archivos GPR del corpus IBB para facilitar la búsqueda y comparación de spots. Este archivo GPR nos da la posición del centro de

cada spot y el diámetro de cada spot. Partiendo de que los spots pertenecen a un mismo gen de dos en dos, la aplicación utiliza esta posición y diámetro máximo de todos los spots de la imagen para realizar una predicción píxel a píxel del segundo spot de una pareja a partir del primero.

A partir de las pruebas realizadas se ha demostrado que en algunos casos es posible mejorar los resultados de compresión obtenidos por diferentes métodos. Por lo tanto, la hipótesis planteada de que la tasa de compresión de las imágenes DNA microarray puede ser mejorada codificando estas imágenes a partir de una predicción basada en la autosimilaridad de las mismas ha quedado demostrada pero no en la totalidad de los casos. Los resultados experimentales indican que con el algoritmo aquí explicado se llega a mejorar los resultados de algunas de las imágenes en hasta 2,88 bpp sin tener en cuenta los archivos externos. Con unos archivos externos que ocupan hasta un 8% del tamaño total para la decodificación sin perdida de toda la información de la imagen, se puede mejorar notablemente los resultados optimizando la compresión de estos.

El sistema utilizado para la codificación puede ser mejorado. Como trabajo futuro, además de almacenar exclusivamente la información necesaria del archivo GPR para la recuperación de la imagen, se podría indagar más en el sistema de predicción. Por ejemplo, se podría realizar un emparejamiento más elaborado para que la similitud de los spots en una pareja sea máxima.

También se podría agrupar los spots en diferentes cantidades para optimizar la información almacenada en la compresión. Y por último habría que considerar el impacto que tendría no utilizar un tamaño máximo de diámetro para todos los spots de la imagen sino que utilizar un tamaño máximo para cada pareja de spots. Esto supondría guardar la mitad de los valores de diámetros, en vez de uno solo máximo, pero podría resultar en un beneficio general respecto a las comparaciones de spots.

Este proyecto abre la puerta a un posible camino a seguir, basándose en la autosimilaridad para obtener mejores resultados de compresión en las imágenes DNA microarray.

AGRADECIMIENTOS

El autor de este documento agradece por la ayuda prestada al tutor de este proyecto, Miguel Hernández, de la Universidat Autònoma de Barcelona.

REFERENCIAS

[1] Kakadu JPEG2000 (http://www.kakadusoftware.com).

[2] S. Battiato and F. Rundo. A bio-inspired CNN with reindexing engine for lossless DNA microarray compression and segmentation. In *Proceedings of the International Conference on Image Processing, ICIP*, volume 1-6, pages 1717–1720. IEEE, 2009.

[3] S. Battiato, F. Rundo, and F. Stanco. Self organizing motor maps for color-mapped image re-indexing. *Image Processing, IEEE Transactions on*, 16(12):2905 –2915, December 2007.

[4] 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*, pages 167–172, 2006.

[5] M. Burrows and D. J. Wheeler. A block-sorting lossless data compression algorithm. Technical Report 124, HP, 1994.

[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*, volume 2, pages 1501–1504, November 2003.

[7] 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*, pages 32–41, 2012.

[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*, 2004(1):92–107, January 2004.

[9] JPEG-LS with LOCO-I algorithm. http://www.hpl.hp.com/research/info_theory/loco/locodown.htm, 2010.

[10] K. C.-H. Kim, P.G.; Park. A quality measure model for microarray images. In *Int. J. Inf. Technol.*, 2005, 11, 117–124.

[11] J. Lempel, A.; Ziv. A universal algorithm for data compression. In *IEEE Trans. Inform Theory*, 1977, 337-343.

[12] S. Lonardi and Y. Luo. Gridding and compression of microarray images. In *Proceedings of the Computational Systems Bioinformatics Conference*, pages 122–130. IEEE, 2004.

[13] A. J. R. Neves and A. J. Pinho. Lossless compression of microarray images using image-dependent finite-context models. *IEEE Transactions on Medical Imaging*, 28(2):194–201, February 2009.

[14] C. H.-S. R. Sauer, U.; Preininger. Quick and simple: Quality control of microarray data. In *Bioinformatics*, 2004, 21, 1572–1578.

[15] S. G.-S. Wang, X.; Ghosh. Quantitative quality control in microarray image processing and data acquisition. In *Nucleic Acids Res*, 2001, 29, e75.