Automatic Data Extraction from Raw Microarray Images

Juan C. Borrás
University of Helsinki
Gustaf Hällströmin katu 2b1
Helsinki, Finland
borras@cs.helsinki.fi

ABSTRACT

Even though protocols have been published to carry out typically *wet lab* procedures, results are inherently noisy with significant variance among the same batch of experiments. Therefore there is a gap between completing a hybridization reaction on the printed spots of a microarray substrate and analyzing the obtained data. As a consequence robust methods must be applied in order to extract the actual hybridization results to be analyzed from the experiment.

We describe automated data extraction strategies based on off-the-shelf algorithms for performing such tasks. The starting point assumes very few of problem condition and we develop some solutions to the situation where some assumptions can be made or where the inherent variability of the cDNA technology creates situations difficult to solve.

Categories and Subject Descriptors

I.4.6 [**COMPUTING METHODOLOGIES**]: IMAGE PROCESSING AND COMPUTER VISION.Segmentation; J.3 [**COMPUTER APPLICATIONS**]: LIFE AND MEDICAL SCIENCES

General Terms

Image processing.

Keywords

Image Processing, Microarray analysis, Fourier Transform, Mathematical Morphology.

1. INTRODUCTION

cDNA microarray technology [1] allows scientists to obtain a *snapshot* of a cell protein expression levels by converting mRNA from expressed genes back to cDNA and comparing hybridization values against a control sample on a wafer where known nucleotides sequences have been printed. Though proprietary technology exist, cDNA based microarray chips can be manufactured by any research laboratory in the world with the proper means, allowing a high degree of flexibility when carrying out experiments, hence its popularity.

The actual protein expression levels however are obtained first by performing image processing procedures over digitalized light responses from two different laser frequencies. These images are quantized using 16 bits per pixel, and no further information (i.e. the area where the hybridization is expected to happen) is included. Therefore segmenting the regions of interest must be carried out prior to any data extraction [4].

Segmentation can be performed by using two different approaches: by locating the hybridization spots in the image and using this information to infer higher level structures like experiment blocks or replicates (*bottom-up*), or by locating high level entities first and locating the spots comprised within them afterward(*top-down*). In any case *gridding* or locating the overall regions of interest where the spots are located, and *addressing* or mapping the individual spots to Cartesian coordinates in the hybridization plate must be carried out.

Due to the popularity of cDNA technology a considerable number of laboratories are carrying out experiments with it using standardized working protocols. As a consequence wide systematic variation in the results arises when comparing images produced from different labs. Therefore a system that minimizes human subjectiviy is highly desired to carry out data extraction from microarray images [9].

The goal of this work is to explore the capabilities of out of shelf image processing algorithms to perform top-down or bottom up gridding and addressing with few assumptions (non-parametric type of estimation) on the underlying data. Early results however will show that an accurate segmentation requires extensive background domain knowledge preventing on some cases a top-down approach.

2. MATERIALS AND METHODS

We have access to two sources of raw microarray images: previous research on image quality assessment [6] and images from the Institute of Biotechnology a the University of Helsinki. Images belonging to the first group are a good starting point since available structures are clearly laid. Images from the second group on the other hand reflect more day to day operations in a typical lab.

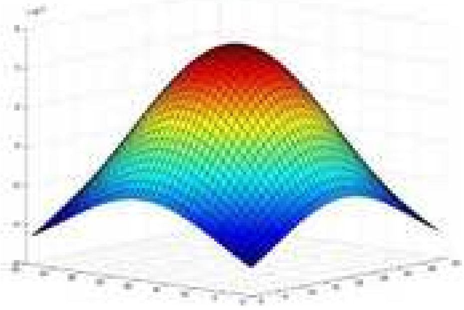

Figure 1: Continuous representation of a matched filter modeling a hybridization response.

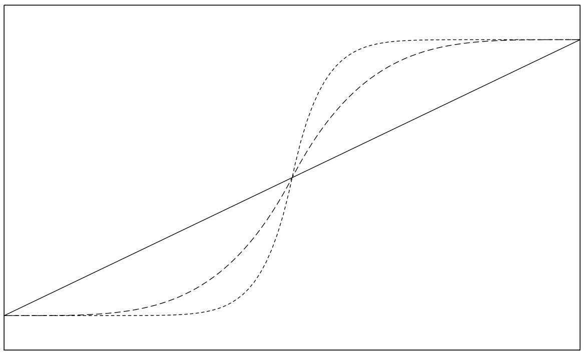

Figure 2: Contrast enhancement function for $d = 1$ (linear localized contrast enhancement), $d = 10$ (almost Heaviside step function or binary quantization) and an intermediate $d = 5$

2.1 Image processing related normalization

The obtained protein expression levels may vary considerably between the different Cy5/Cy3 responses. As a consequence expression values are calculated as the logarithm of the ratio of both Cy5/Cy3 channel readings. Such high variation in the image responses (which are nothing else that the image intensity values) make it a challenging environment [2, 10]. However if we consider the segmentation problem a separated one, we can normalize the overall image intensity values, resulting in an image with false expression levels and therefore invalid for proteomic studies but very valid for tackling the spot detection problem.

Intensity normalization is done locally by applying a sliding window over the image that frames a single pixel and a square neighborhood around it. For each neighborhood the maximum and minimum intensity values are used to compute the new pixel value [8]. The mathematical definition of the contrast enhancement function is defined as follows:

$$u = \Gamma(t) = \begin{cases} \frac{\frac{1}{2}(u_{\max}-u_{\min})}{(\mu_f - t_{\min})^d} \cdot (t - t_{\min})^d + u_{\min} & \text{if } t \leq \mu_f \\ \\ \frac{-\frac{1}{2}(u_{\max}-u_{\min})}{(\mu_f - t_{\max})^d} \cdot (t - t_{\max})^d + u_{\max} & \text{if } t \geq \mu_f \end{cases}$$

t_{min} and t_{max} are respectively the minimum and the maximum intensity value of the neighborhood. These values vary as the window travels through the image. μ_f is the mean intensity of the neighborhood. u_{min} and u_{max} define the posterior normalized intensity range and are chosen to be respectively 0 and 255 in order to achieve high contrast for visualization by humans.

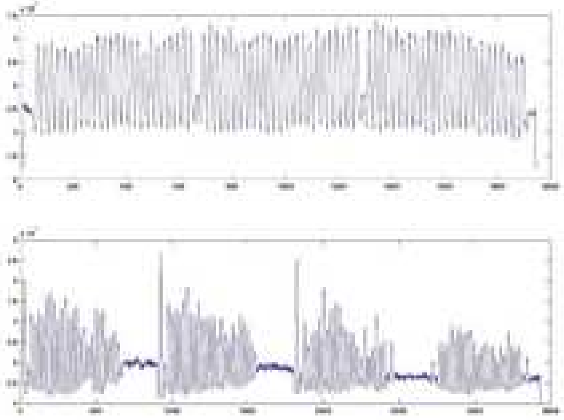

Figure 3: Side profiles over columns (above) and over rows (under).

The only free parameter is d which controls how dramatic the contrast stretches in neighborhoods with low variance. A plot of the function for different values of d can be seen in figure 2. Readers familiar with the sigmoid function may associate it but in this case the enhanced intensity values for I_{min} and I_{max} in the neighborhood are exactly the enhanced contrast values.

2.2 Regular Assumptions

Hybridization spots are printed using automated equipment that lay the samples on a point lattice[1]. The existence of this highly regular arrangement can greatly simplify the gridding process by constructing *side profiles*. Side Profiles, shown in figure 3 are sums over all rows and over all columns of the image intensity values. Local minima of the side profiles tell us where the lines dividing rows of spots lay, and local maxima of the side profiles are a rough estimation of where the spot centers are in the whole image.

Additionally, the periodic nature of the side profiles can be subjected to Fourier analysis, where the representation in the frequency domain allows us to estimate spots features like size or location. Also by computing the inverse transform using the components comprising most of the power, once can discard at the same time noise originating from the image while using the local maxima of the reconstructed signal as a probability estimator of spot locations.

2.3 Replicate blocks

Replicate blocks are exact copies of a given spot layout with the same contents ready for hybridization. Replicate blocks are very useful when dealing statistically with issues like dye

[1]think of a point lattice as a fancy name for a regular square grid

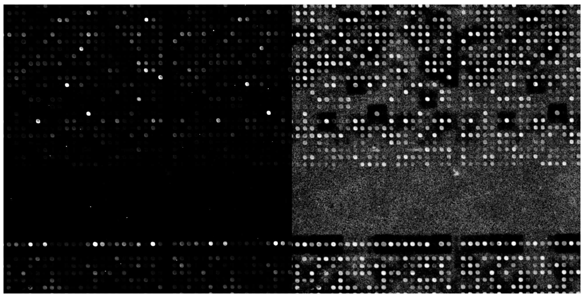

Figure 4: Comparison of the original microarray image from the QA study before (left) and after (right) contrast enhancement and normalization.

bias, or more simply allow you to have a large population of hybridization samples in the same experiment.

As a consequence, addressing must locate hybridized spots on the microarray glass and also map them to their corresponding replicate block. This task can be simplified if the individual point lattice dimension is known for each block, which is typically the case. Here however we try to identify the different replicate structures by detecting inter-block separation if present in the side profiles.

2.4 Matching filters

Matched filters are simple convolution masks, used in linear filtering, where its coefficients are chosen carefully so that they mimic the shape (pattern) we are interested to detect. Matching filters are local operators in scope. Results of the convolution are taken then as a similarity measure where higher filter responses corresponds to the shapes we are interested to detect. Local maxima are an estimate of the location of the hybridized spots.

Spots are modeled as a two-dimensional Gaussian functions (see figure 1) where the dimensions of the spot and the area around it are known and define the filter size. The remaining σ parameter in the Gaussian function can bet taken from the range from one half to one third of the filter size since the influence of the value within those limits can be safely discarded.

2.5 Mathematical Morphology

Mathematical Morphology [5] or simply *morphology* can be defined as a theory for the analysis of spatial structures. The name arises because it aims at analyzing the shape and form of objects. It is mathematical in the sense that the analysis is based on set theory, integral geometry and lattice algebra. Mathematical morphology is a powerful image analysis technique.

Morphological operators aim at extracting relevant structures of the image considered as a set. This is achieved by probing the image with another set of known shape called *structuring element* (SE). The shape of the SE is usually chosen according to some a priori knowledge about the geometry of the relevant and irrelevant (to be suppressed) image structures. Essentially, a SE is nothing but a small and well defined set used to probe the image under study.

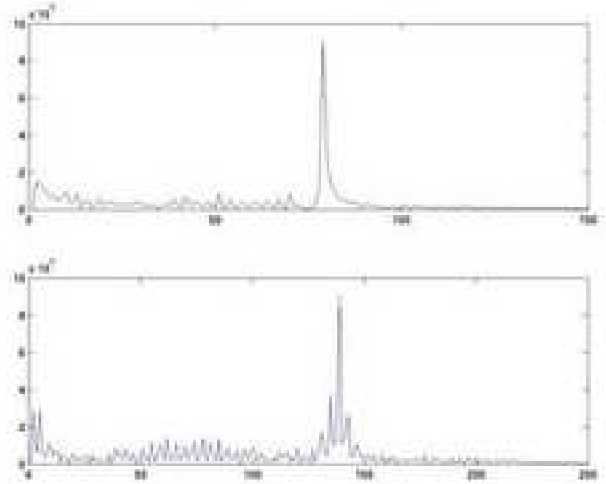

Figure 5: Highly regular side profiles show mostly low frequency signal components (only the lower components are shown).

This study uses two dimensional SEs (in practice a binary matrix), which receive the name *flat* SEs.

Fundamental morphological operators require the definition of an origin for each SE, the basic operations are defined as follows:

The *erosion* of an grayscale image f by a structuring element B is denoted by $\varepsilon_{\mathbf{b}}(f)$ and is defined as the minimum of the translations of f by the vectors $-\mathbf{b} \in B$. Hence, the eroded value at a given pixel $\mathbf{x}$ is the minimum value of the image in the window defined by the structuring element when its origin is at $\mathbf{x}$:

$$[\varepsilon_{\mathbf{b}}(f)](\mathbf{x}) = \min_{\mathbf{b} \in B} f(\mathbf{x} + \mathbf{b}) \tag{1}$$

The *dilation* of a grayscale image f by a structuring element B is denoted by $\delta_{\mathbf{b}}(f)$ and is defined as the maximum of the translations of f by the vectors $-\mathbf{b} \in B$. Hence, the dilated value at a given pixel $\mathbf{x}$ is the maximum value of the image in the window defined by the structuring element when its origin is at $\mathbf{x}$:

$$[\delta_{\mathbf{b}}(f)](\mathbf{x}) = \max_{\mathbf{b} \in B} f(\mathbf{x} + \mathbf{b}) \tag{2}$$

The *opening* of an grayscape image f by a structuring element B is denoted by $\gamma_B(f)$ and is defined as the erosion of f by B followed by the dilation with the *reflected* SE $\check{B}$:

$$\gamma_B(f) = \delta_{\check{B}}[\varepsilon_B(f)] \tag{3}$$

Finally, the *closing* of an grayscale image f by a structuring element B is denoted by $\phi_B(f)$ and is defined as the dilation of f with a structuring element B followed by the erosion with the reflecting structuring element $\check{B}$:

$$\phi_B(f) = \varepsilon_{\check{B}}[\delta_B(f)] \tag{4}$$

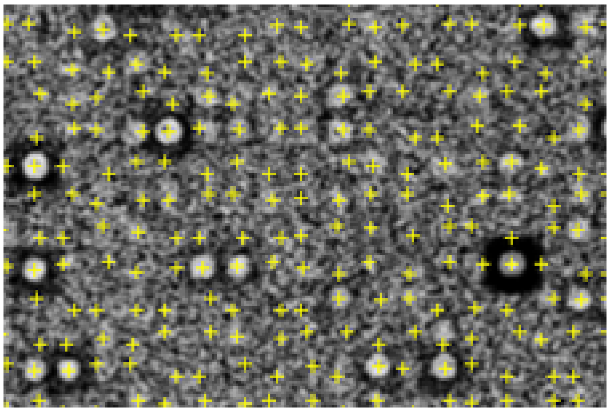

Figure 6: Spot locations using the inverse FFT local maxima with local corrections.

3. EXPERIMENTS

All images from the QA study have been subjected to contrast enhancement and normalization. Figure 4 allows the reader to compare the effect. As noted, this processing stage invalidates hybridization values the image but makes gridding and addressing easier so we can extract location coordinates from the enhanced image and apply them to the original one.

Enhanced images show highly regular oscillating side profiles that can be subjected to Fourier analysis. Representation in the frequency domain shows that the signal power is highly concentrated in the lower part of the spectrum as show in figure 5.

Therefore we can take the frequency component with the highest value and use its local maxima in the spatial domain to estimate centers. Notice however that some spots may be located outside their theoretical positions in the lattice, therefore we must account for some local correction using a matched filters in order to improve accuracy. Figure 6 shows the result.

The images coming from the Institute of Biotechnology however are a more realist representation of how daily microarray images look like. Expression levels show extreme values, with saturation reading on some samples while the overall expression results are quite low. So in order to increase image contrast we add together the laser responses corresponding to both Cy3 and Cy5 channels, prior to apply contrast enhancement.

Also, they are usually *slanted* because the lattice orientation is not parallel with the boundaries of the microarray glass which turn does determine the image orientation.

Slanting is important since a straight line drawn from one of the topmost spots in the image and continued to the last spot at the bottom *following the direction of the image*, we end up crossing the boundaries of the spot alignment on the lattice. Manual slant calculation is not time consuming, and its value is required to properly calculate side profiles.

Side profiles in this set of image show periodic oscillations

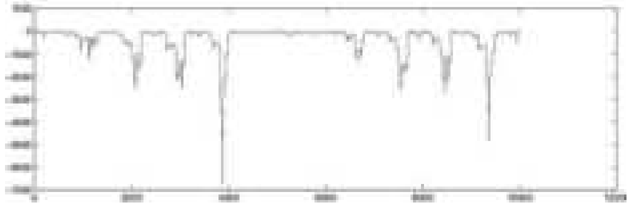

Figure 7: Closing of the side profile over columns. Blocks a delimited by the values that sink under the uniform values in the side profile

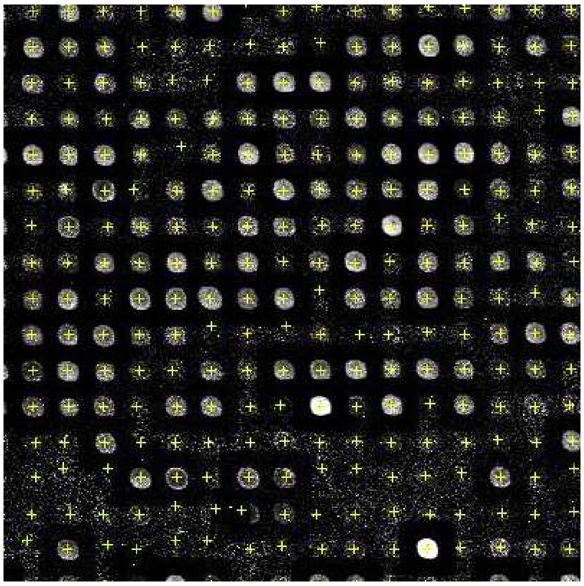

Figure 8: Spot location results using matched filters.

but they are much more localized in the space.A direct consequence is that the frequency spectrum is not concentrated around a single component and therefore we cannot use for estimating spot centers using this approach.

The significant separation that appears between block in the intensity sums over columns is due to the organization of the experiments as replicate blocks. This information must be returned as part of the addressing problem, and we can extract it by applying the morphological closing over the side profile [3, 7] over columns as shown in figure 7.

The good news is that spot locations can be extracted by applying a matched filter as described above and look for local maxima in the image using the spot size as a criteria for discarding candidates too close to each other. The results of this approach can be see in figure 8. The drawback is however the high cost in terms if computing power and the need of extra information in order to avoid detecting local maxima in the spacing between spots.

4. DISCUSSION

Non-parametric (assumption-free) microarray spot detection is highly desirable since a solution of this kind can be applied over a wide range of images classes produced from many different labs. However background domain knowledge (spot size, lattice geometry, high level replicate structuring) can greatly simplify both gridding and addressing. Actually the best results are achieved by minimizing the overall intervention of automated higher level structures analysis. Ideally a matched filter approach where all other information related to the experiment is know is the most promising situation.

Normalization and local contrast enhancement of the original microarray image corrupts completely the Cy3/Cy5 response signals but it is definitely of help when segmenting the image. Specially when the overall hybridization results are generally low.

Up-down analysis makes sense when the high level structures from the experiment design can be inferred from the image, or when that information is known. The results of addressing are then complete, and not only the spot coordinates are offered to the user but also the replicate characteristics from its block number.

Bottom-up analysis performed as convolution of matched filters can be always carried out and it's results are quite independent from the underlying contrast enhancement approach applied on the image.

Simple unlocalized Fourier analysis is of great use for estimating parameters of the microarray experiment like the spot and inter spot spacing values due to the periodic nature of the lattice. Notice however that if the separation of the replicates is unknown this technique looses effectively and required other approaches that localize in the space the oscillating features inherent in any microarray analysis.

Slating is a reminder of the many factors that contribute to the large variability of cDNA experiments, mostly as a results of a technology that basically allows any practitioner to develop their own experimental setup which is particular for each laboratory in the world.

5. CONCLUSIONS

The problem of gridding and addressing in microarray images has been presented and out of the shelf algorithms have been applied to solve them.

The assumption free approach to tackle both problems is definitely too ambitious, specially when a tailor made solution for the products of a single laboratory can use all background knowledge available simplifying the overall data extraction procedure. Notice however that it is possible to achieve quite acceptable results in the latter situation.

6. ACKNOWLEDGMENTS

The author would like to thank Salla Ruosaari and Rashi Gupta for supplying raw microarray images without which this work would have not been possible.

7. REFERENCES

[1] A. Brazma and J. Vilo. Gene expression data analysis. *FEBS Letters*, 480:17–24, 2000.

[2] S. Dudoit, Y. H. Yang, M. Callow, and T. Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. Technical report, University of California at Berkeley. Department of Statistics, 2002.

[3] R. Hirata, J. Barrera, R. F. Hashimoto, D. O. Dantas, and G. H. Esteves. Segmentation of microarray images by mathematical morphology. *Real-Time Imaging*, 8(6):491–505, 2002.

[4] A. Jain, T. Tokuyasu, A. Snijders, R. Segraves, D. Albertson, and D. Pinkel. Fully automatic quantification of microarray image data. *Genome Research*, 2(12):325–332, February 2002.

[5] P.Soille. *Morphological Image Analysis. Principles and Applications*. Springer-Verlag, New York, 2 edition, 2003.

[6] S. Ruosaari and J. Hollmén. Image analysis for detecting faulty spots from microarray images. In *Proceedings of the 5th International Conference on Discovery Science*, pages 259–266. Springer-Verlag, 2002.

[7] K. I. Siddiqui, A. O. Hero, and M. M. Siddiqui. Mathematical morphology applied to spot segmentation and quantification of gene microarray images. In *Asilomar Conference on Signals and Systems*, volume 1, pages 926–930, November 2002.

[8] T.Walter and J-C.Klein. A computational approach to diagnosis of diabetic retinopathy. In *6th Conference on Systemics, Cybernetics and Informatics (SCI)*, Orlando, Florida, 2002.

[9] C. Uehara and I. A. Kakadiaris. Towards automatic analysis of dna microarrays. In *Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision*, pages 57–62, 2002.

[10] Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed. Comparison of methods for image analysis on cdna microarray data. Technical report, University of California at Berkeley. Department of Statistics, 2000.