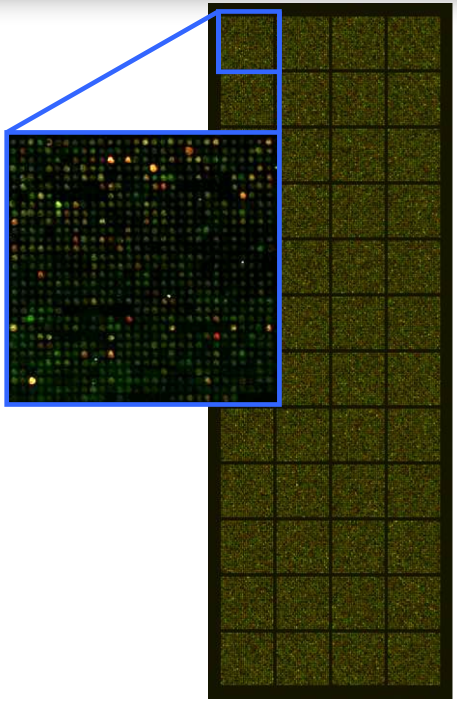
Valutazione di software di analisi di microarray basato su simulazioni di immagini da dati reali



Ignazio Infantino

Tutorial su "**Metodi e strumenti per l'analisi dei dati di espressione genica**"
18 Dicembre 2007 ICAR-CNR Napoli

Overview

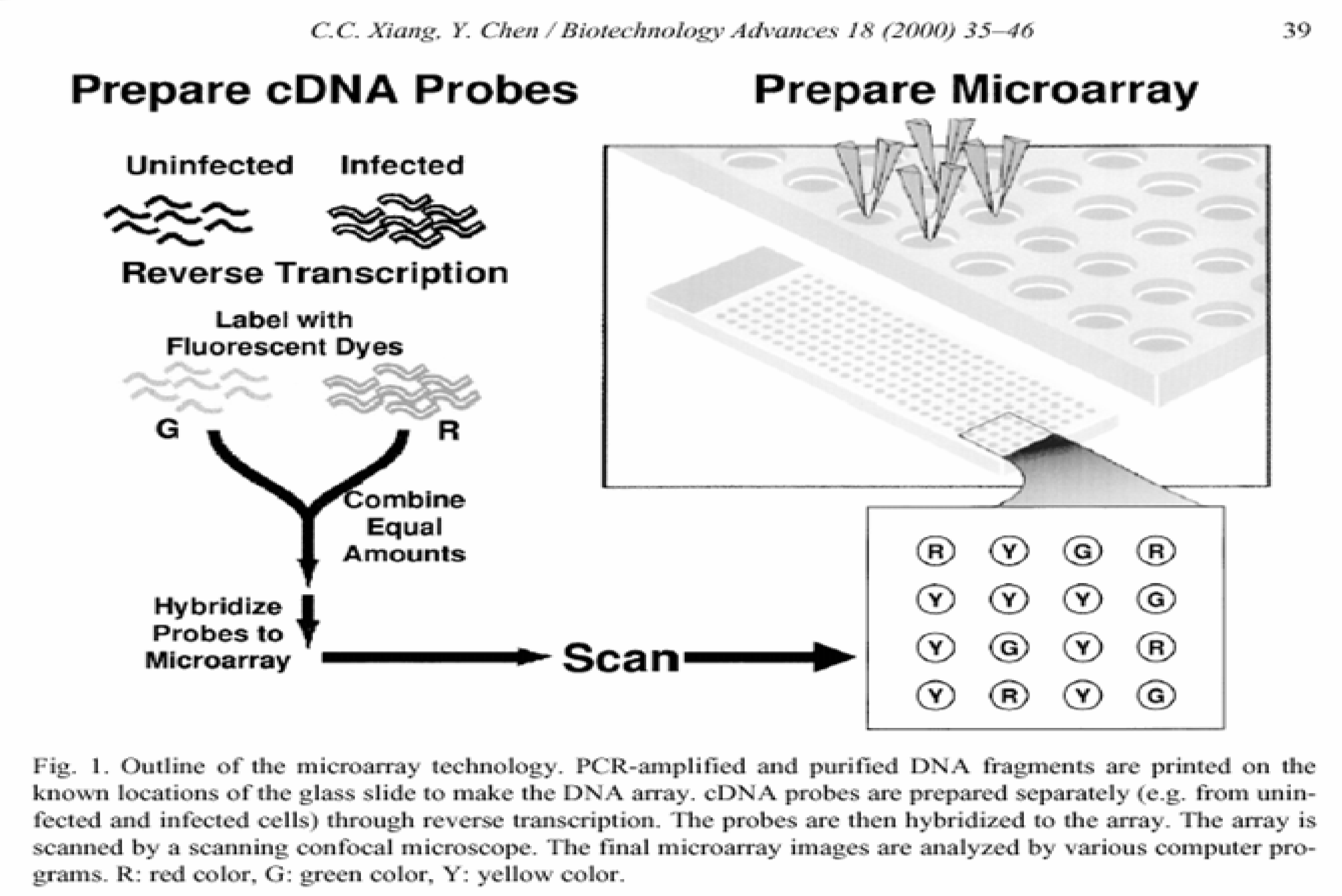
- Introduction
  - Analysis of Microarray images
    - Addressing or gridding
    - Segmentation
    - Intensity extraction
- Microarray results: how accurate are they?
- Evaluation by simulation
  - Statistical analysis of real data
  - Simulation Model
  - Testing
  - An example of application
- Conclusions
- References

Introduction

- cDNA microarray technologies have large diffusion in biological research field
  - For studying gene expression in many different organism
  - To large-scale gene discovery
  - For polymorphism screening and mapping of genomic DNA clones
- Fully automated and reliable software analysis systems are required to process the large amount of data produced
- Effective testing and evaluation of employed computational approaches is still an open problem

C. C. Xiang, and Y. Chen, "**cDNA microarray technology and its applications**", in Biotechnology Advances, vol. 18, 2000, pp. 35–46.

Microarray technology

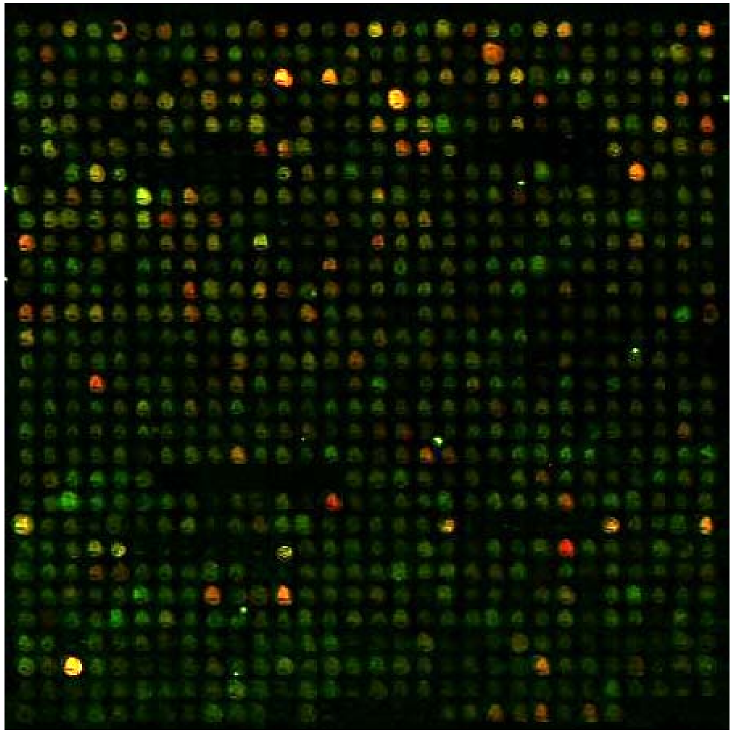


C.C. Xiang, Y. Chen / Biotechnology Advances 18 (2000) 35–46 39

Fig. 1. Outline of the microarray technology. PCR-amplified and purified DNA fragments are printed on the known locations of the glass slide to make the DNA array. cDNA probes are prepared separately (e.g. from uninfected and infected cells) through reverse transcription. The probes are then hybridized to the array. The array is scanned by a scanning confocal microscope. The final microarray images are analyzed by various computer programs. R: red color, G: green color, Y: yellow color.

Analysis of microarray images

- A laser scanner detects sample fluorescence
  - Cy3 probe at 532 nm
  - Cy5 probe at 633 nm
- Combined RGB color image represents differentially expressed genes
  - The fluorescence intensities are stored as 16-bit images which we view as “raw” data
- Image processing software analyzes red and green hybridization intensities and red/green ratio for every gene on the array, $\log_2(R/G)$



Analysis of microarray images (cntd)

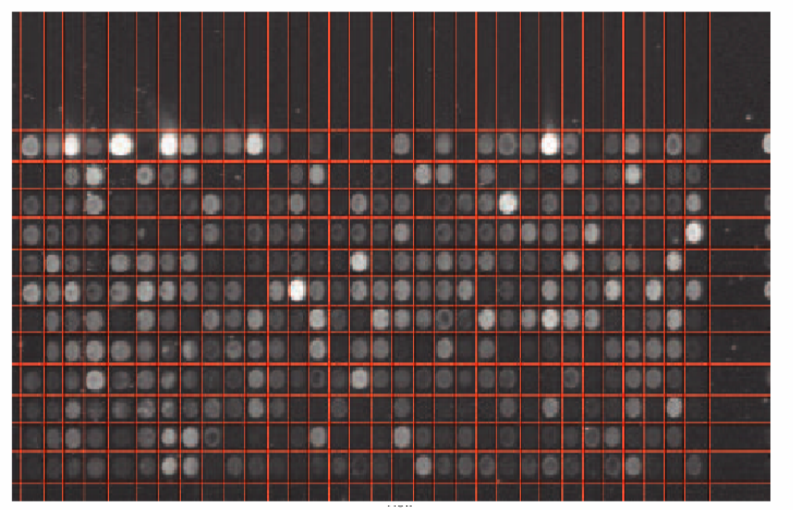
- The digitalization process produces, for each pixel, a signal that represents the total fluorescence in the region corresponding to that pixel
- When properly processed, this signal should correlate to the area density of dye molecules

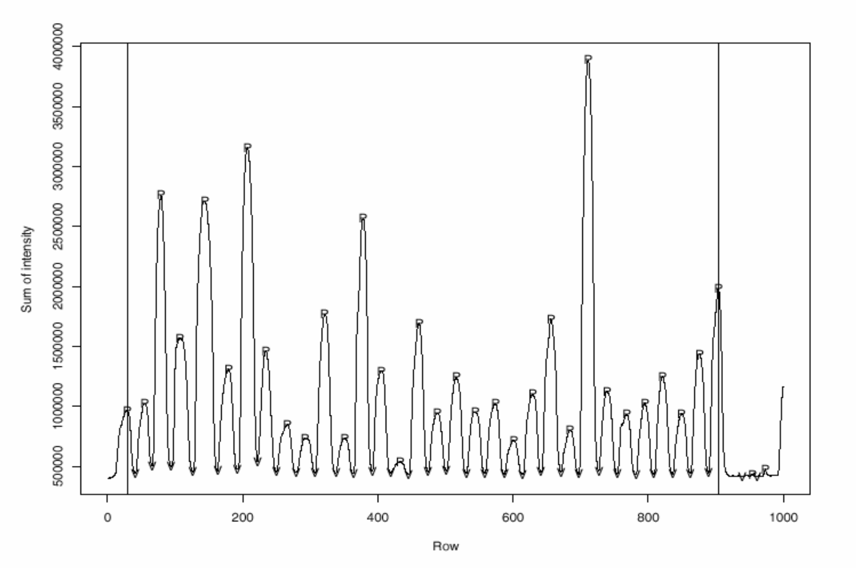
- Addressing or gridding
- Segmentation
- Intensity extraction

Y. H. Yang, M. J. Buckley M, et al., "**Comparison of methods for image analysis on cDNA microarray data**", in J. Comp. Graph. Stat., 11, 2002, 108-136

Addressing or gridding

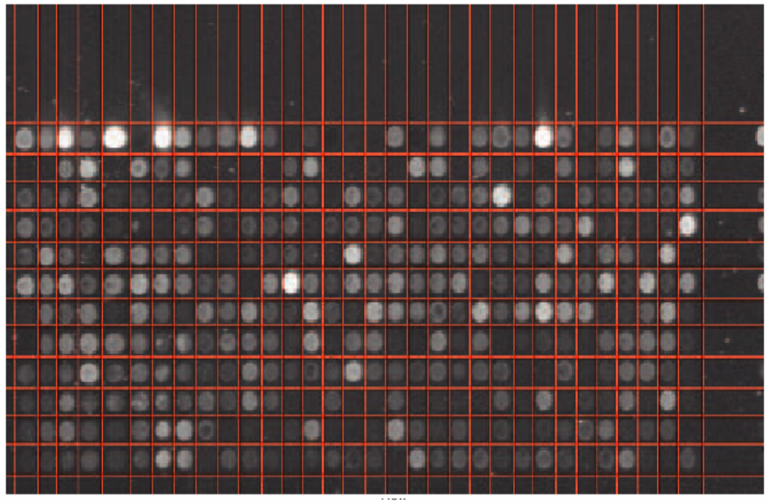
- It is the process of assigning coordinates to each of the spots
- Automating this part of the procedure permits high throughput analysis

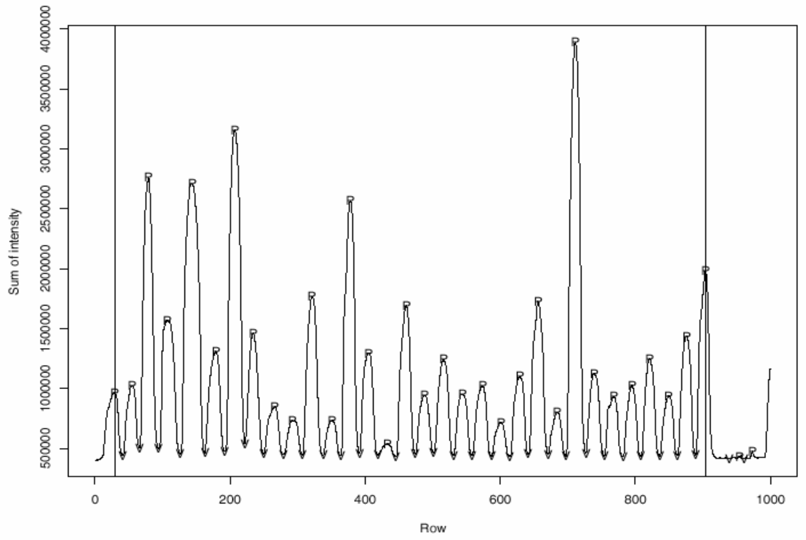




Addressing or gridding (cntd.)

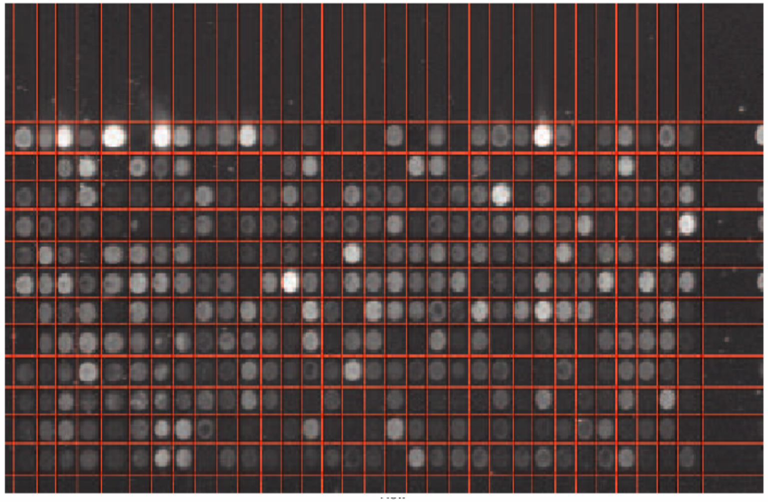
- Separation between rows and columns
- Individual translation of grids
  - caused by slight variations in print-tip positions
- Separation between rows and columns of spots
- Overall position of the array in the image

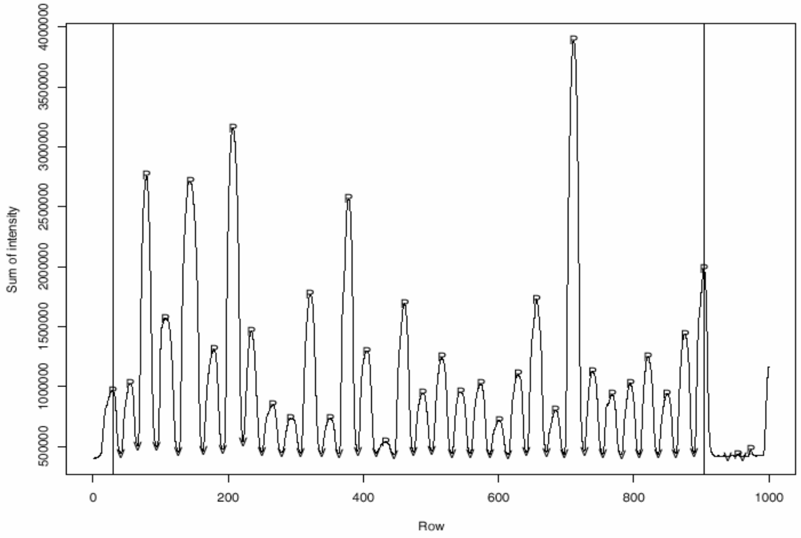




Addressing or gridding (cntd.)

- Misregistration of the red and green channels
- Rotation of the array in the image
- Skew in the array





Segmentation

- It allows the classification of pixels either as foreground (i.e. the spot of interest) or as background

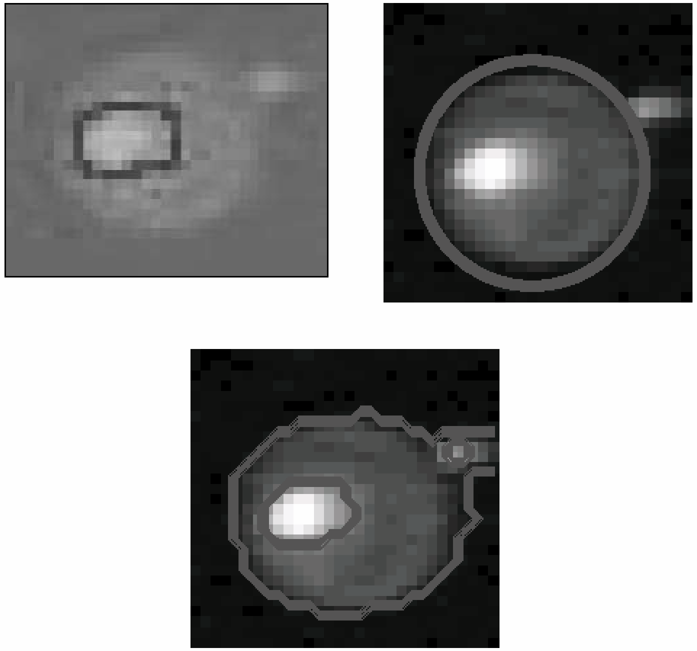


Fig. 7. Segmentation results for another donut-shaped spot. Top left: SPOT; top right: spot-on; and bottom: model-based segmentation.

Used methods

- Fixed circle segmentation
  - ScanAnalyze
- Adaptive circle segmentation
  - GenePix
- Adaptive shape segmentation (watershed, seeded region growing SRG)
  - Spot
- Histogram segmentation (Otsu)
  - QuantArray

Segmentation (cntd.)

- In microarray image analysis, we are in the rather unusual situation where the number of features (spots) is known exactly *a priori*
- the approximate locations of the spot centers are determined at the addressing stage

Q. Li, C. Fraley, R. E. Bumgarner, K. Y. Yeung, and A. E. Raftery, "**Donuts, scratches and blanks: robust model-based segmentation of microarray images**", in Bioinformatics, 21:12, 2005, pp. 2875–2882

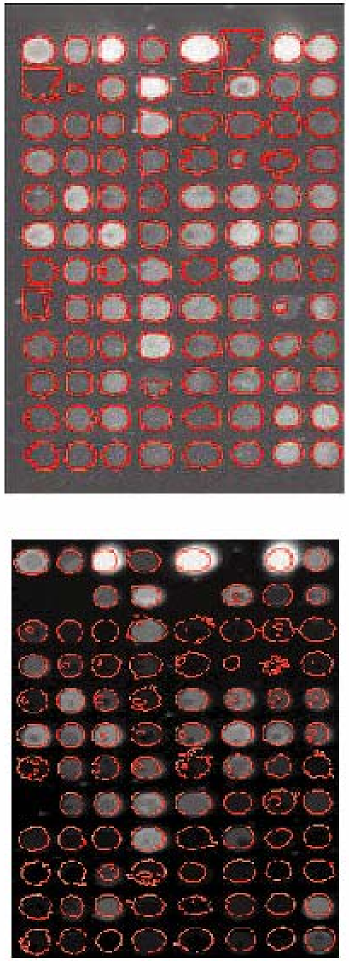


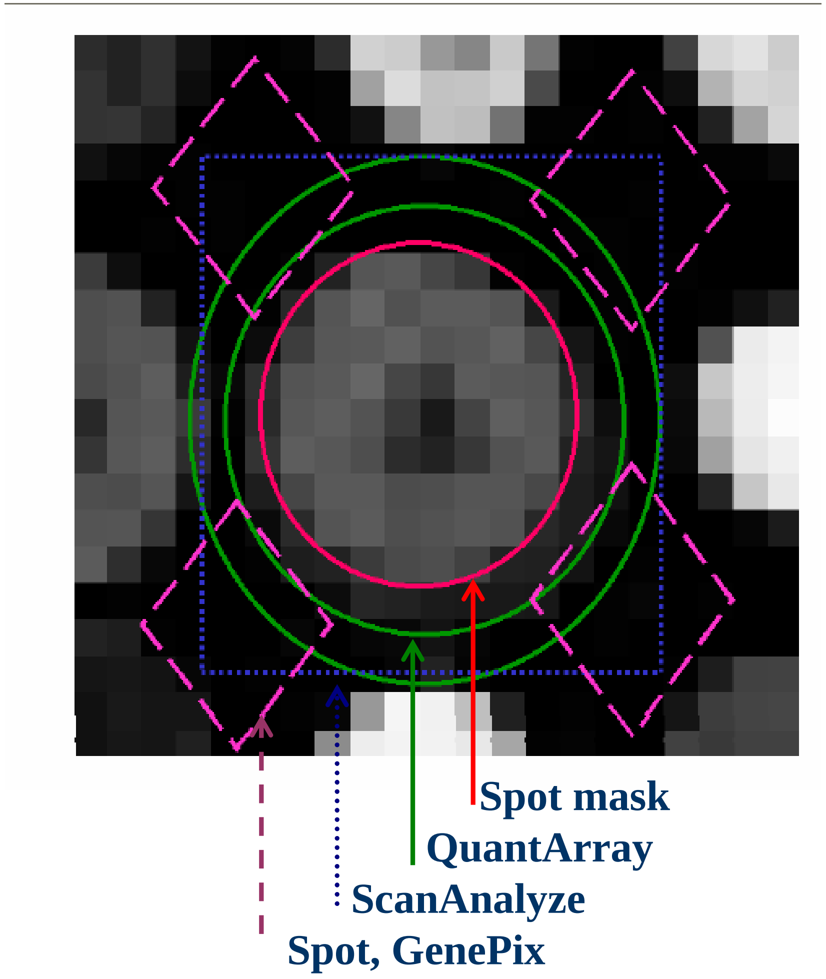
Fig. 11. Segmentation results for a 12 × 8 subset of the array. SPOT; and lower panel, model-based segmentation.

Intensity extraction

- This step includes calculating for each spot on the array
  - red and green foreground fluorescence intensity pairs (R,G)
  - background intensities
    - Measured fluorescence intensity includes a contribution which is not specifically due to the hybridization process (background correction)
  - quality measures
    - Measures of
      - Spot size
      - Spot shape
      - Background intensity vs foreground intensity

Intensity extraction (cntd.)

- Most microarray analysis packages define the foreground intensity as the mean or median of pixel values within the segmented spot mask
- More variation exists in the choice of background calculation method
  - Example: taking the median of values in selected regions surrounding the spot mask



Microarray results: how accurate are they?

- Several inconsistencies from different commercially available systems
  - Inconsistent sequence fidelity of the spotted microarrays
  - Variability of differential expression
  - Low specificity of cDNA microarray probes
  - …
- Conclusions: In view of the pitfalls, data from microarray analysis need to be interpreted cautiously

R. Kothapalli, S. J Yoder, S. Mane, and T. P. Loughran Jr, “**Microarray results: how accurate are they?**”, in BMC Bioinformatics, 3:22, 2002

J. Quackenbush, “**Computational analysis of microarray data**”, in Nat. Rev. Genet., vol. 2, 2001,418-427.

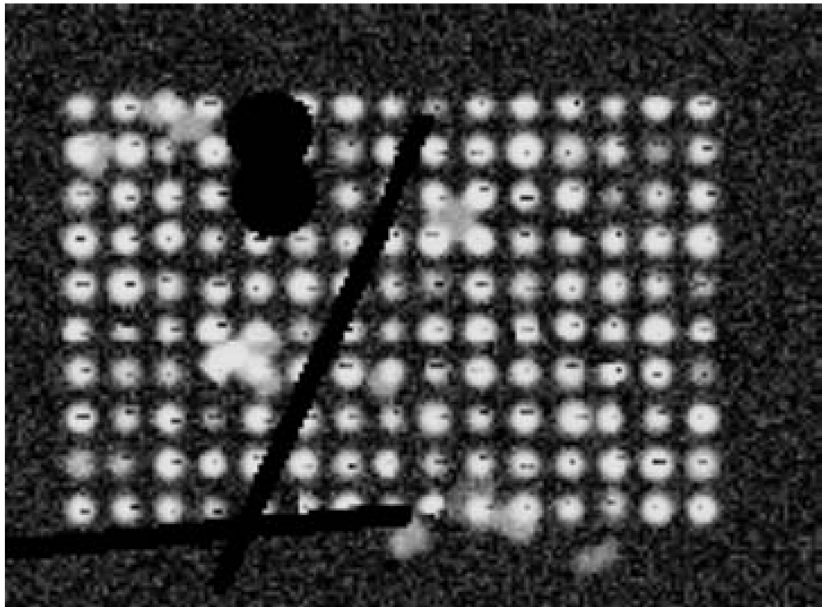
Modeling microarray experiments

- A microarray simulation model can be used to validate different kinds of data analysis algorithms
  - Obtaining realistic biological measurement data
  - Referring to valid ground truth data
- The simplest approach to generate the ground truth data is to sample data randomly from a specific distribution.
  - the distribution and its parameters can be estimated from real measurements.
  - the ground truth data can be obtained by sampling a simulated ideal distribution with estimated parameters

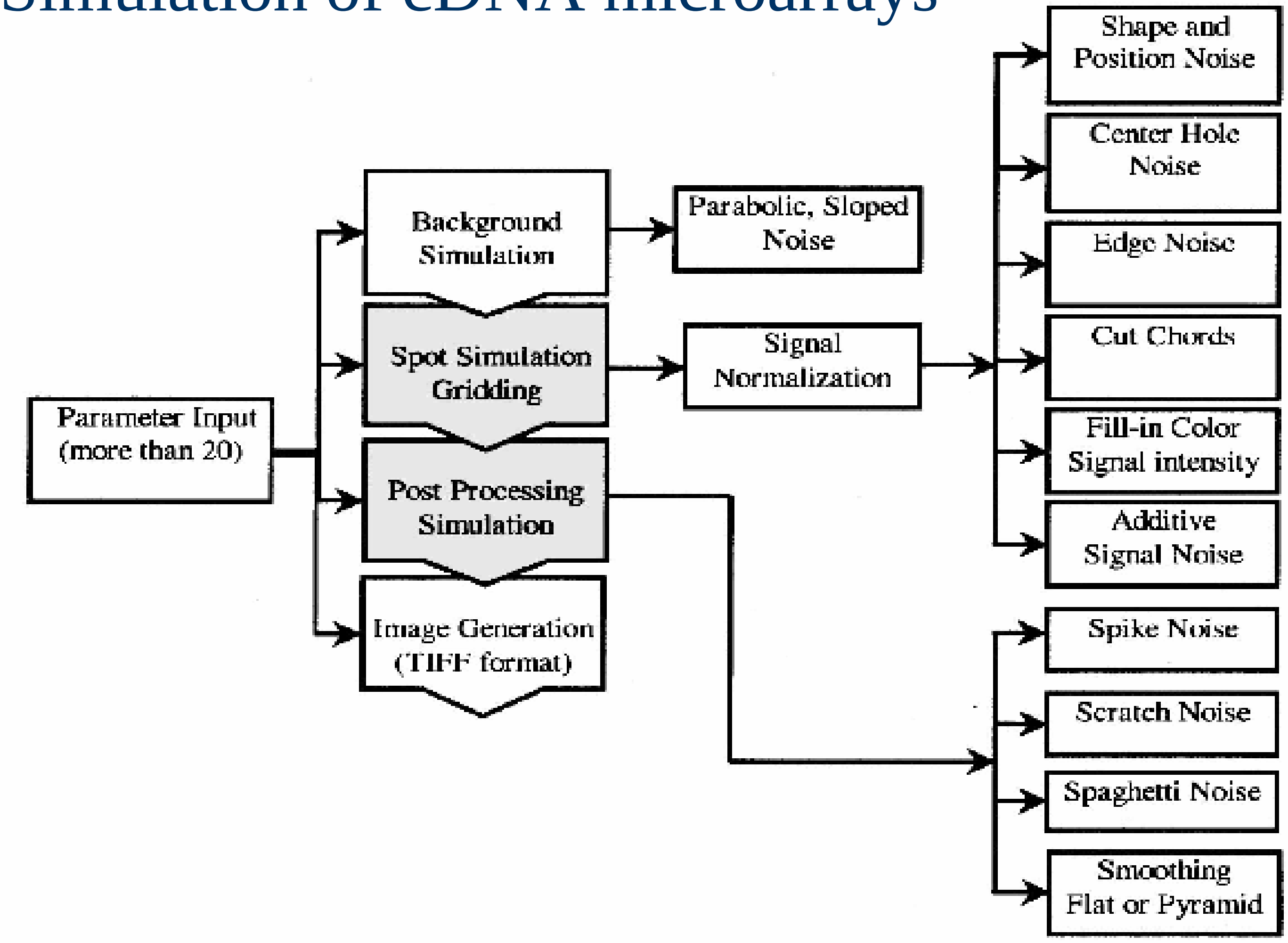
M. Nykter, T. Aho, M. Ahdesmäki, P. Ruusuvuori, A. Lehmussola, and O. Yli-Harja, "**Simulation of microarray data with realistic characteristics**", in BMC Bionformatics, 7:349, 2006

Errors and noises

- Slide manufacturing
  - subarray drifting from ideal rectangular layout.
- Slide hybridization
  - Spot blending
  - Scratches
  - Air bubbles
  - Background noise
- Slide scanning
  - Spot saturation
  - Channels misalignment
  - Translation, rotation, skew



Simulation of cDNA microarrays



Y. Balagurunathan, E. R. Dougherty, Y. Chen, and M. L. Bittner, J. M. Trent, "**Simulation of cDNA microarrays via a parameterized random signal model**", in Journal Biomed Opt, 7(3), 2002, pp. 507-523

Our approach

- To provide researchers with a tool for testing and evaluating the performance of analysis software

- Characteristics of a given microarray experiment are captured from public databases
  - Extracted data are then used for generating a synthetic microarray image and the corresponding ground truth data (i.e. the gene expression values).
  - Given a particular experiment, and/or a specific hardware, and/or a software tool for microarray image analysis and so on, we use raw data obtained in identical or similar conditions to model simulated images with known values of gene activation.

- Using the simulated images as benchmark, one can estimate expected errors and choose the most suitable analysis software for the real experiment.

Statistical analysis of real data

- In order to simulate realistic microarray images, public databases could be used to choose parameters of the model of spot and grid generation
  - Stanford MicroArray Database
    - http://genome-www5.stanford.edu/
- Grid geometry, spot locations, and other details are recovered from results file (gpr format)

I. Infantino, C. Lodato, S. Lopes, “**Testing and evaluation of microarray image analysis software**”, 2nd Intl. Conf. on Complex, Intelligent and Software Intensive Systems (CISIS 2008), Intl. Workshop on Intelligent Informatics in Biology and Medicine (IIBM 2008), Barcelona, Spain, March 4th – 7th, 2008

Spot model: feature positions

- Experiment #28370 in Stanford Microarray Database
  - Related to the normal tissue of hyperinsulinemic clamp in human muscle in diabetes
  - Tiff images have dimension 5556 x 1952
  - 43200 spots
  - Results saved as gpr file
- To transform the coordinate system from nm to pixels
  - image origin $O_1=(x_{01,}y_{01})$ is (920,7720)
  - pixel size S_{pixel} is 10 nm
  - coordinates $(x_f,\ y_f)$ and diameter diaf in pixels of features are obtained as:

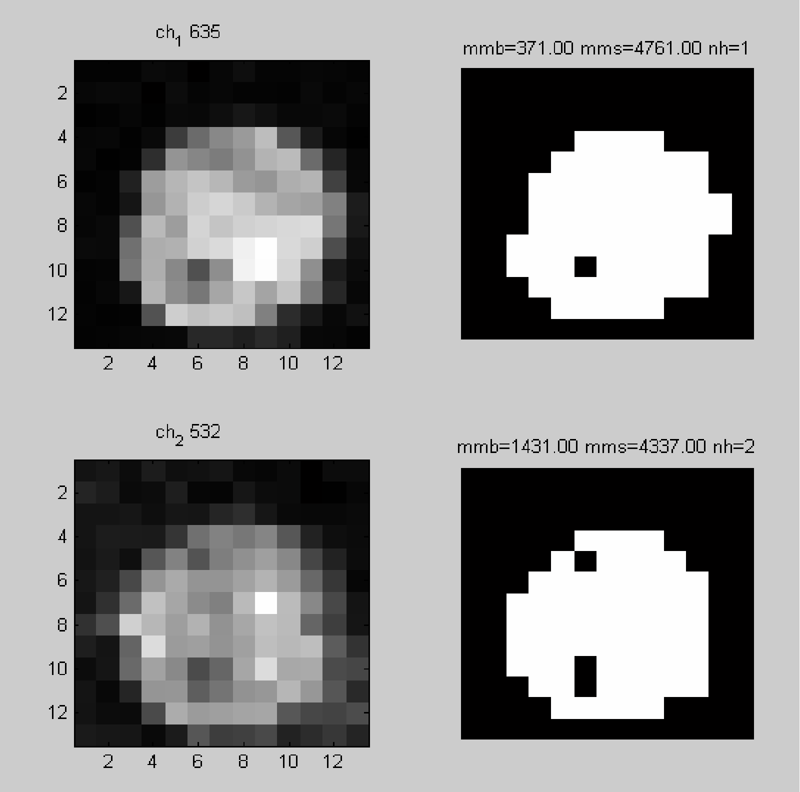
$$x_f = x - x_{01} / s_{pixel}$$

$$y_f = y - y_{01} / s_{pixel}$$

$$dia_f = dia / s_{pixel}$$

Spot model: background and foreground

- Background and spot pixels of considered area are separated by K-means algorithm
  - looking for two intensities clusters grouping by squared Euclidean distance
  - K-means initial seed points are the first pixel of the region (upper-left corner) and the central one
- Holes are found by labeling connected regions using 4-connection



Spot model: feature intensities

- median of spot pixel intensities medians
  - MM_{spot_ch1}, MM_{spot_ch2}
- mean of spot pixel intensities variances
  - Var_{spot_ch1}, Var_{spot_ch2}
- mean of correlation of spot pixel intensities between channels
  - $Corr_{spot\_ch1}$, $Corr_{spot\_ch2}$
- variance of correlation of spot pixel intensities between channels
  - $VarCorr_{spot\_ch1}$, $VarCorr_{spot\_ch2}$

Noise and defects

Background noise

- median of background pixel intensities medians
  - MM_{b_ch1}, MM_{b_ch2}
- mean of background pixel intensities variances
  - Var_{b_ch1}, Var_{b_ch2}

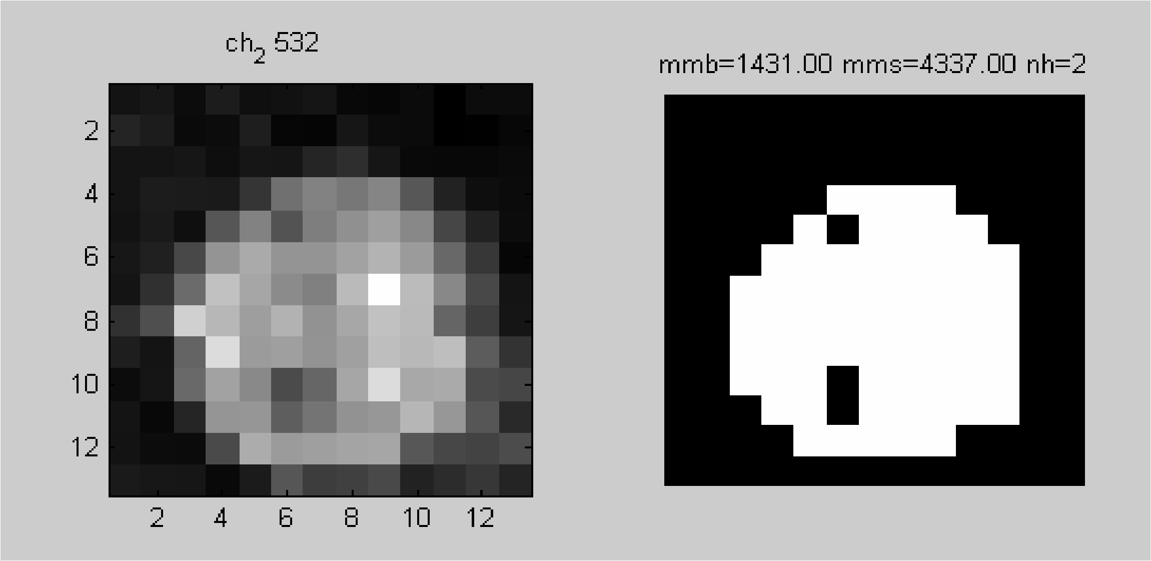
Spot alignment

- mean of x misalignment between barycentre of spot and centre of sub-region which includes it
  - $M_{x\_mis\_ch1}$, $M_{x\_mis\_ch2}$
- mean and variance of y misalignment
  - $M_{y\_mis\_ch1}$, $M_{y\_mis\_ch2}$

Noise and defects (cntd.)

Holes within spot

- mean of number of holes
  - $M_{n\_hole\_ch1}$, $M_{n\_hole\_ch2}$
- mean of distances of holes from spot barycentre
  - $M_{d\_hole\_ch1}$, and $M_{d\_hole\_ch2}$



Values

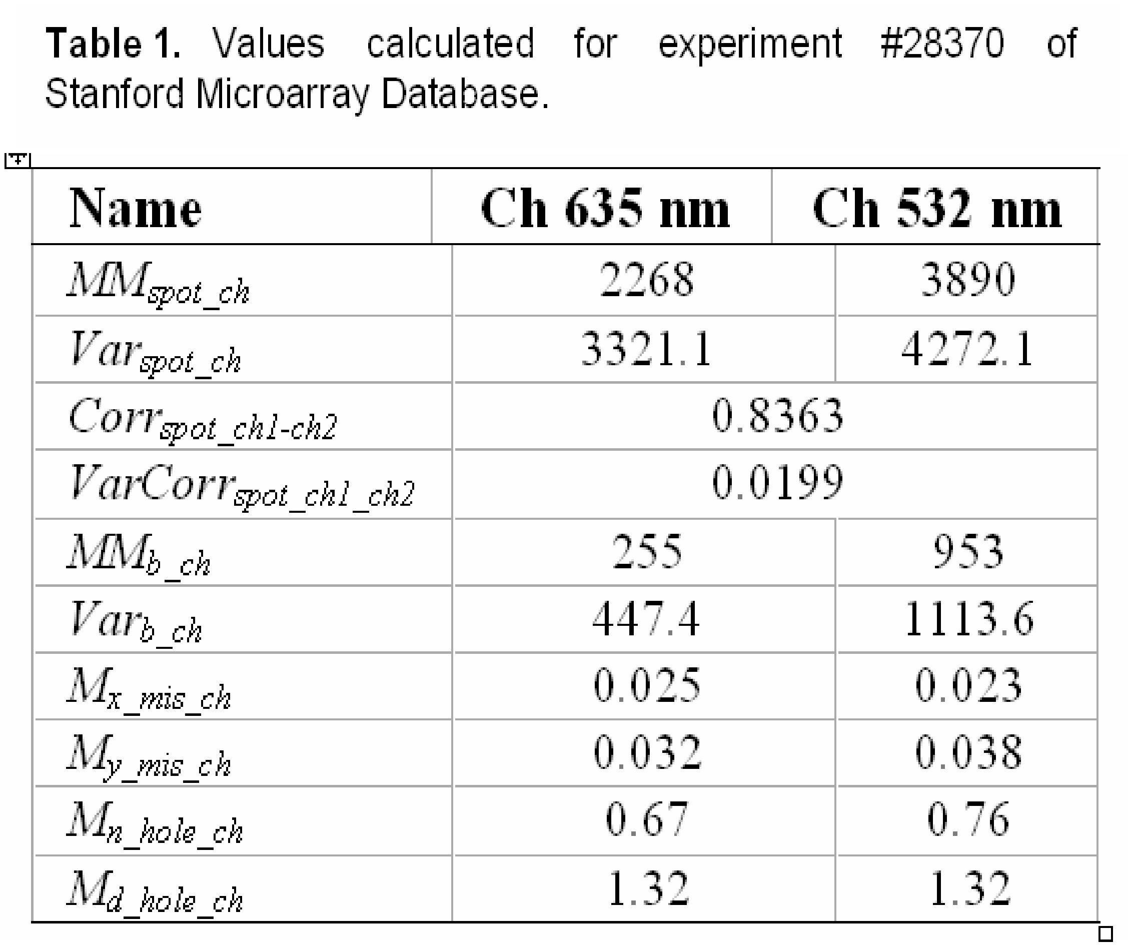


Table 1. Values calculated for experiment #28370 of Stanford Microarray Database.

Name	Ch 635 nm	Ch 532 nm
MM_{spot_ch}	2268	3890
Var_{spot_ch}	3321.1	4272.1
$Corr_{spot\_ch1-ch2}$	0.8363	
$VarCorr_{spot\_ch1\_ch2}$	0.0199	
MM_{b_ch}	255	953
Var_{b_ch}	447.4	1113.6
$M_{x\_mis\_ch}$	0.025	0.023
$M_{y\_mis\_ch}$	0.032	0.038
$M_{n\_hole\_ch}$	0.67	0.76
$M_{d\_hole\_ch}$	1.32	1.32

Simulation Model

- spots generally have a non-regular shape
  - some morphological distortion instead of being perfectly circular
  - spots with doughnut shape are frequently observed in real microarray images
- Shape can be effectively modeled by a linear combination of two bivariate Gaussian distributions
  - Given a bivariate Gaussian distribution *G(x,y)*
  - and a set of *n* secant planes $\{P_n(x,y)\}$
  - each surface can be expressed by the last element of a sequence $\{C_n(x,y)\}$ of linear combination between functions describing a surface and the corresponding secant plane

Simulation Model (cntd.)

- denoting with *m* the generic surface of the envelope, the model can be expressed by the following equations:

$C_{m0}(x,y)=G_m(x,y)$

$C_{mn}(x,y)=C_{mn-1}(x,y)$

for each x,y in $D|\ C_{mn-1}(x,y) \leq P_{mn}(x,y)$ (eq. 1)

$C_{mn}(x,y)=a_{mn}C_{mn-1}(x,y)+b_{mn}P_{mn}(x,y)$

for each x,y in $D|\ C_{mn-1}(x,y)>P_{mn}(x,y)$,

with a_{mn},b_{mn} in {-1,0,1,2} (eq. 2)

where $m=1,2,\ldots M$ (M no. envelope surfaces),

$n=1,2,\ldots N_m$ (N_m no. section planes of m-th surface).

Spot shapes

- The values of coefficients a_{mn} and b_{mn}, affect the specific type of shape: convex, plane, concave
- Equations (1) and (2) describe how the surface is modified by the application of the corresponding secant plane
- the surface is generally cut in two regions
  - The one under the secant plane remain unchanged
  - whereas the other is modified in accord to the equation (2)
- There are three different combination modes allowed, namely
  - *clip* $a_{mn} = 0$, $b_{mn} = 1$
  - *dig* $a_{mn} = -1$, $b_{mn} = 2$
  - *lift* $a_{mn} = 2$, $b_{mn} = -1$

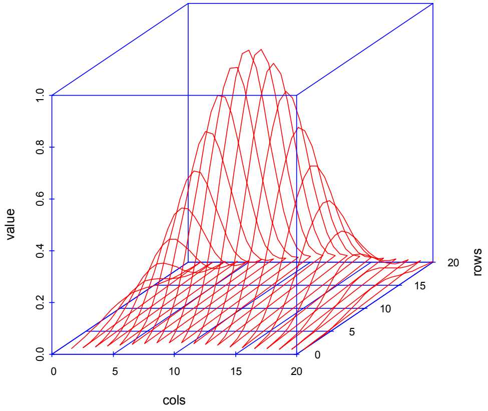
Spot shapes (cntd.)

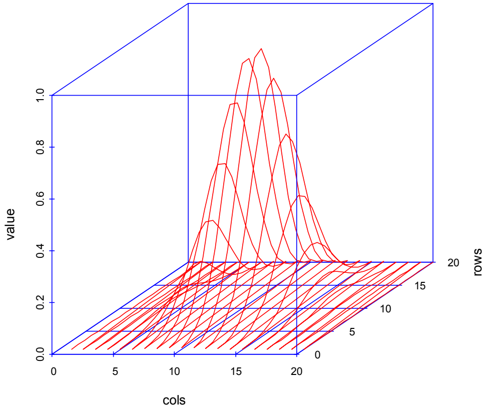
- The previous described procedure is repeated until all the secant planes have been applied, thus producing a single element of the envelope
- The succeeding elements will be defined repeating the same procedure as many times as many are the number of specified Gaussian
- Hence, the final surface is obtained enveloping all the single surfaces:

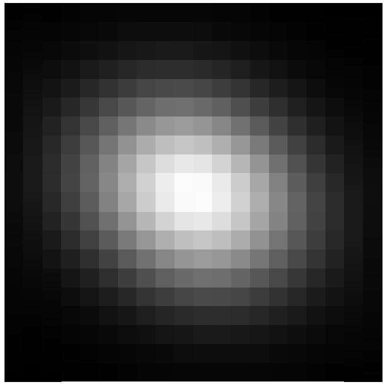
$$S(x,y)=\max(C_m N_m(x,y))$$

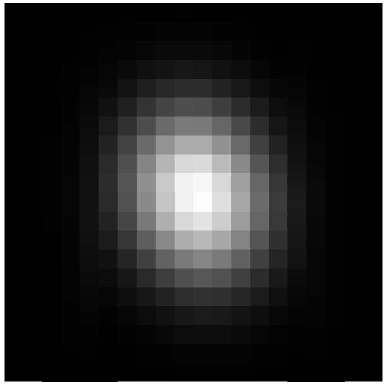
- The resulting function S(x,y) represents the spatial distribution of the simulated spot brightness.

Spot example

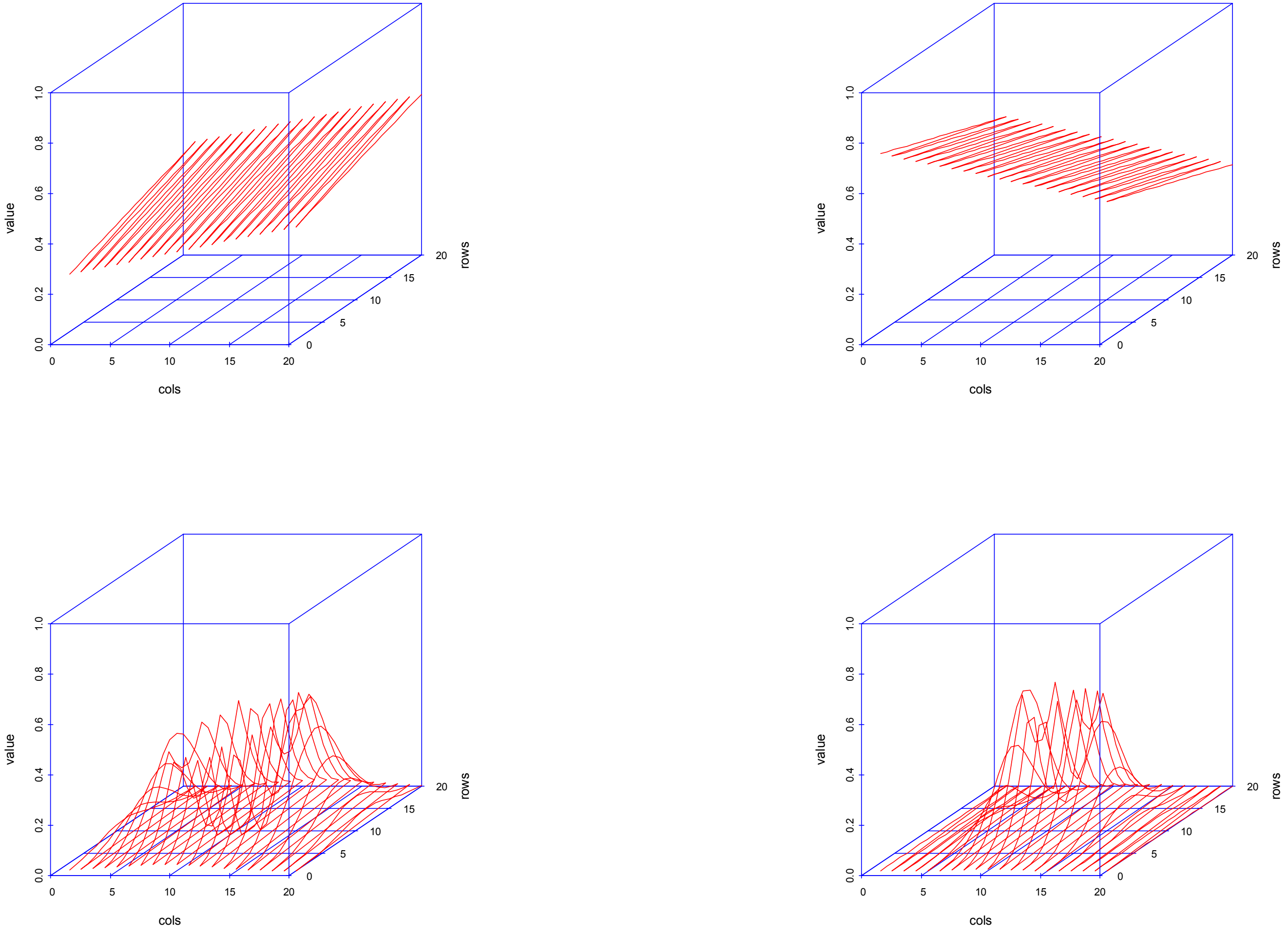




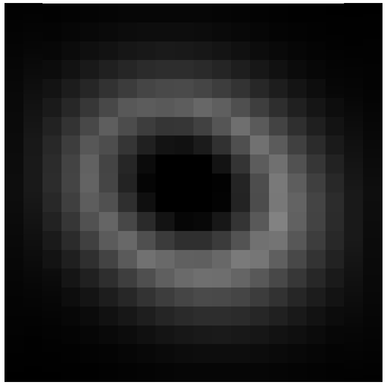


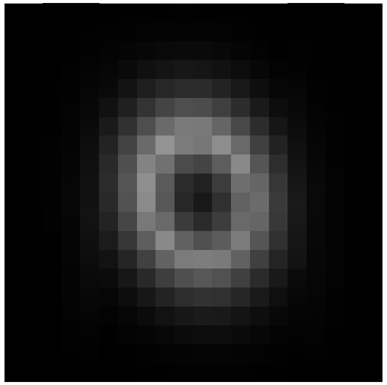


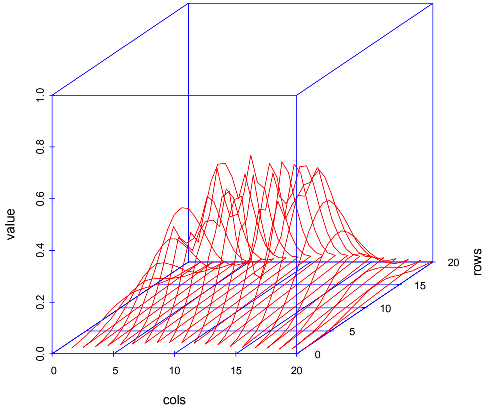
Spot example

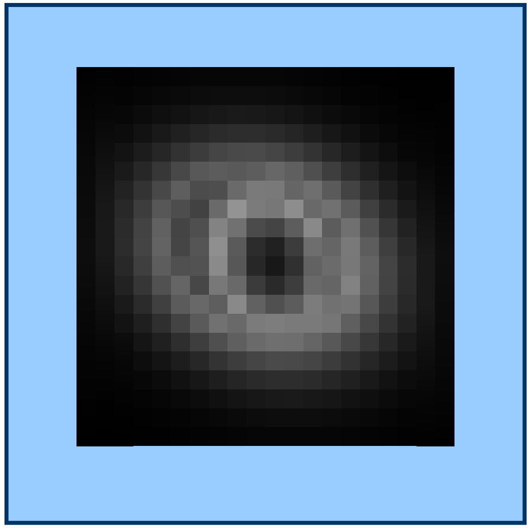


Spot example



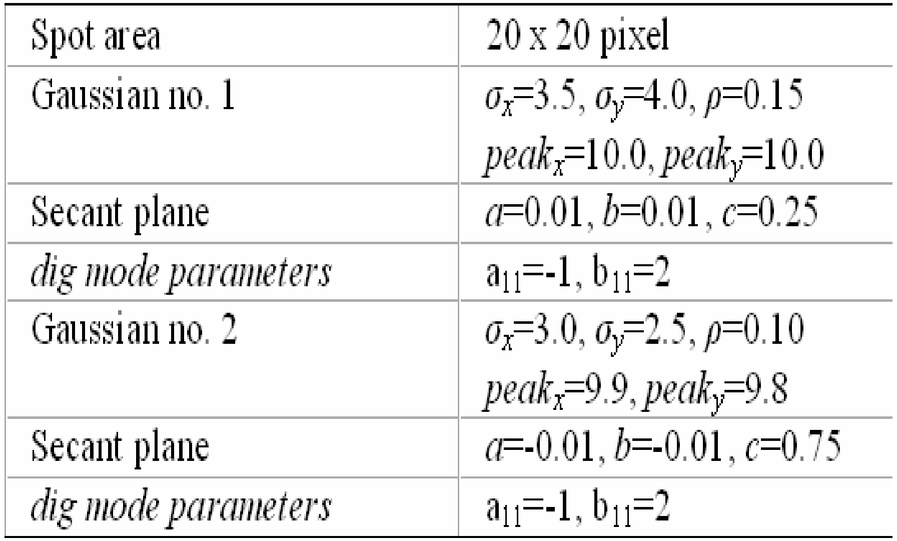






Shape parameters

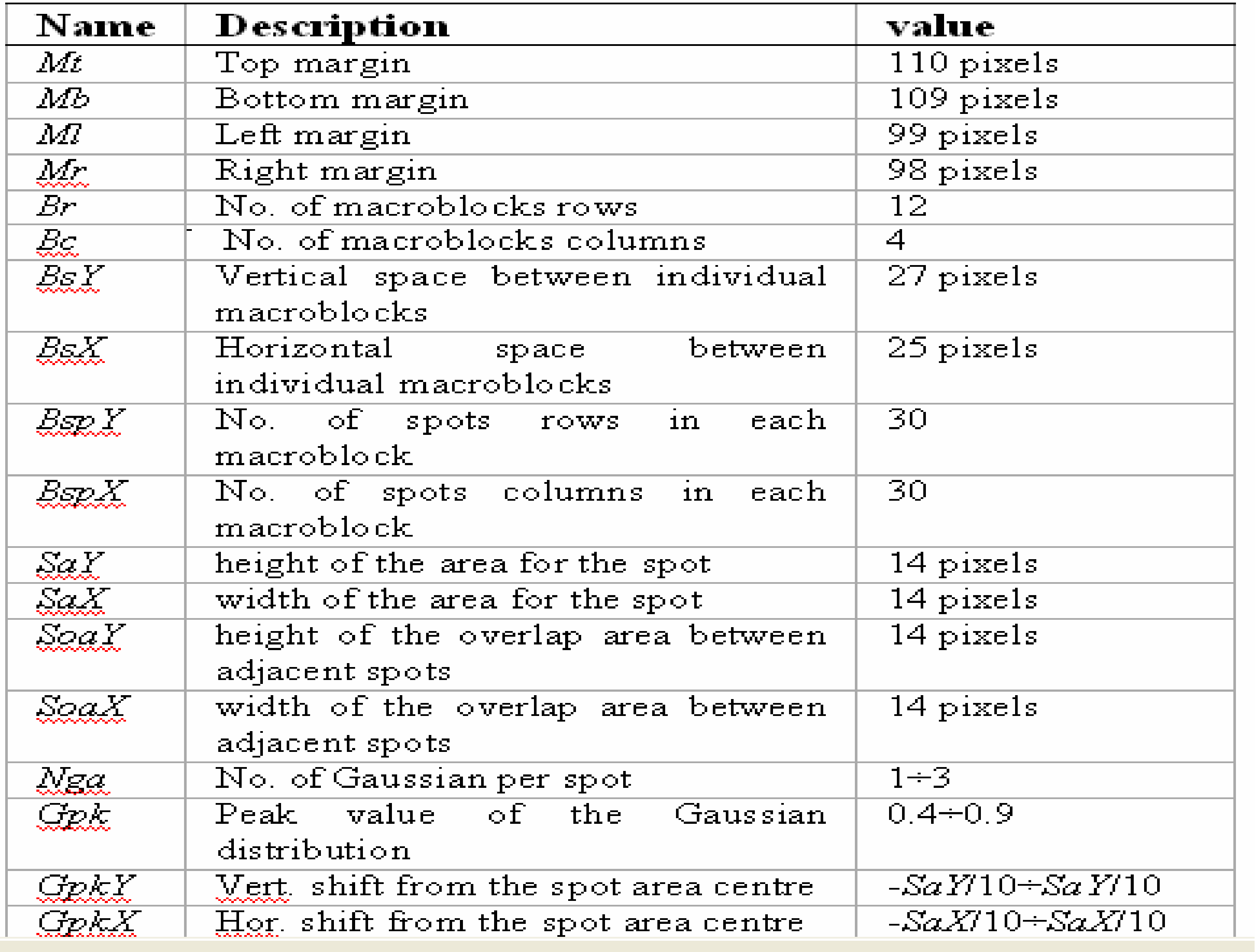
- The parameters characterizing each Gaussian distribution are
  - the standard deviation along *x*,*y*-direction (σ_x,σ_y)
  - the correlation factor *r*
  - the peak coordinates ($peak_x$, $peak_y$)
- Three coefficients (*a*, *b*, *c*) are needed in order to specify each correlated secant plane.



Spot area	20 x 20 pixel
Gaussian no. 1	σ_x=3.5, σ_y=4.0, ρ=0.15 $peak_x$=10.0, $peak_y$=10.0
Secant plane	a=0.01, b=0.01 , c=0.25
dig mode parameters	a_{11}=-1, b_{11}=2
Gaussian no. 2	σ_x=3.0, σ_y=2.5, ρ=0.10 $peak_x$=9.9, $peak_y$=9.8
Secant plane	a=-0.01, b=-0.01 , c=0.75
dig mode parameters	a_{11}=-1, b_{11}=2

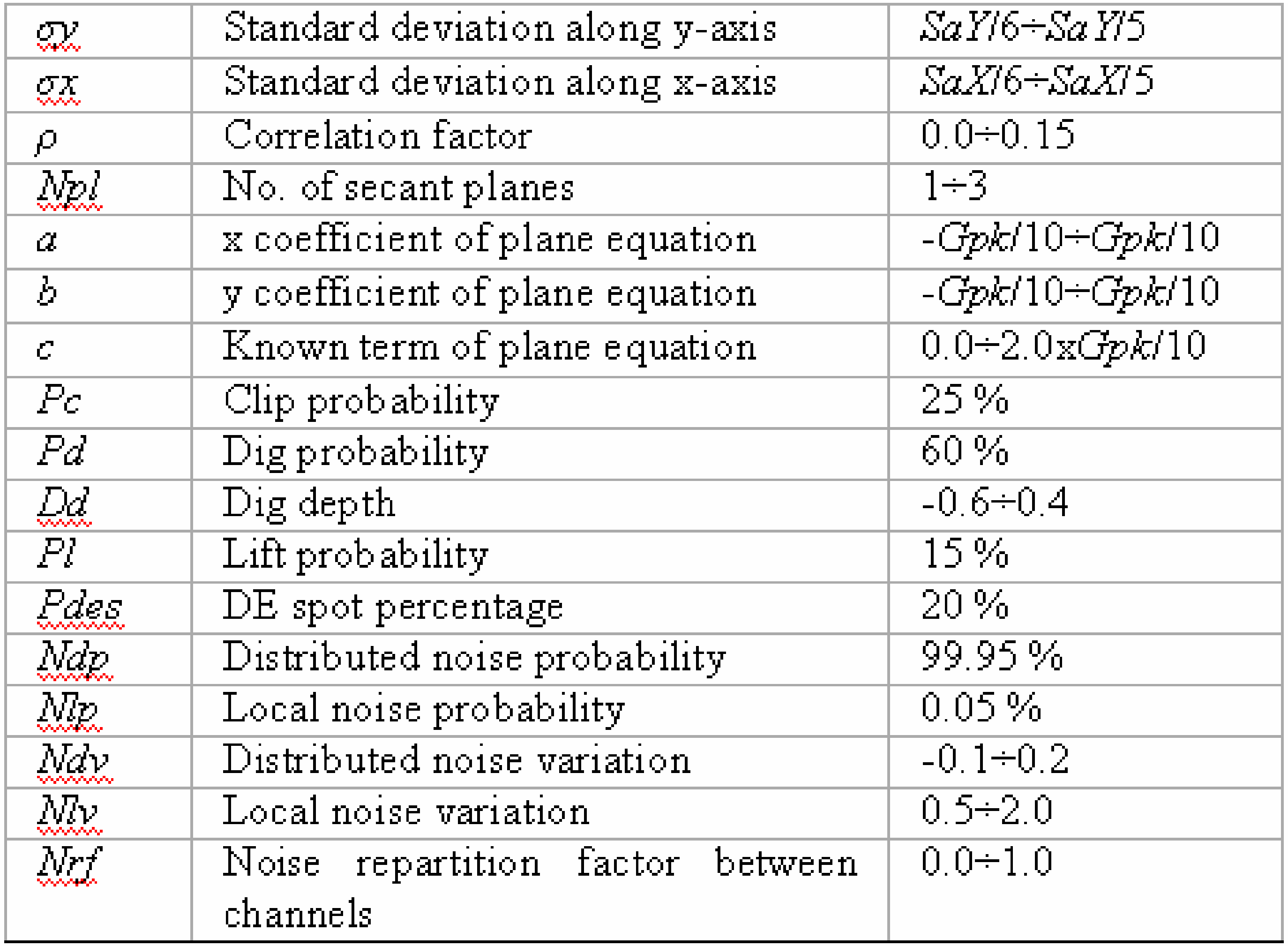
Table 2. Parameters values for the spot generation.

Simulation Parameters



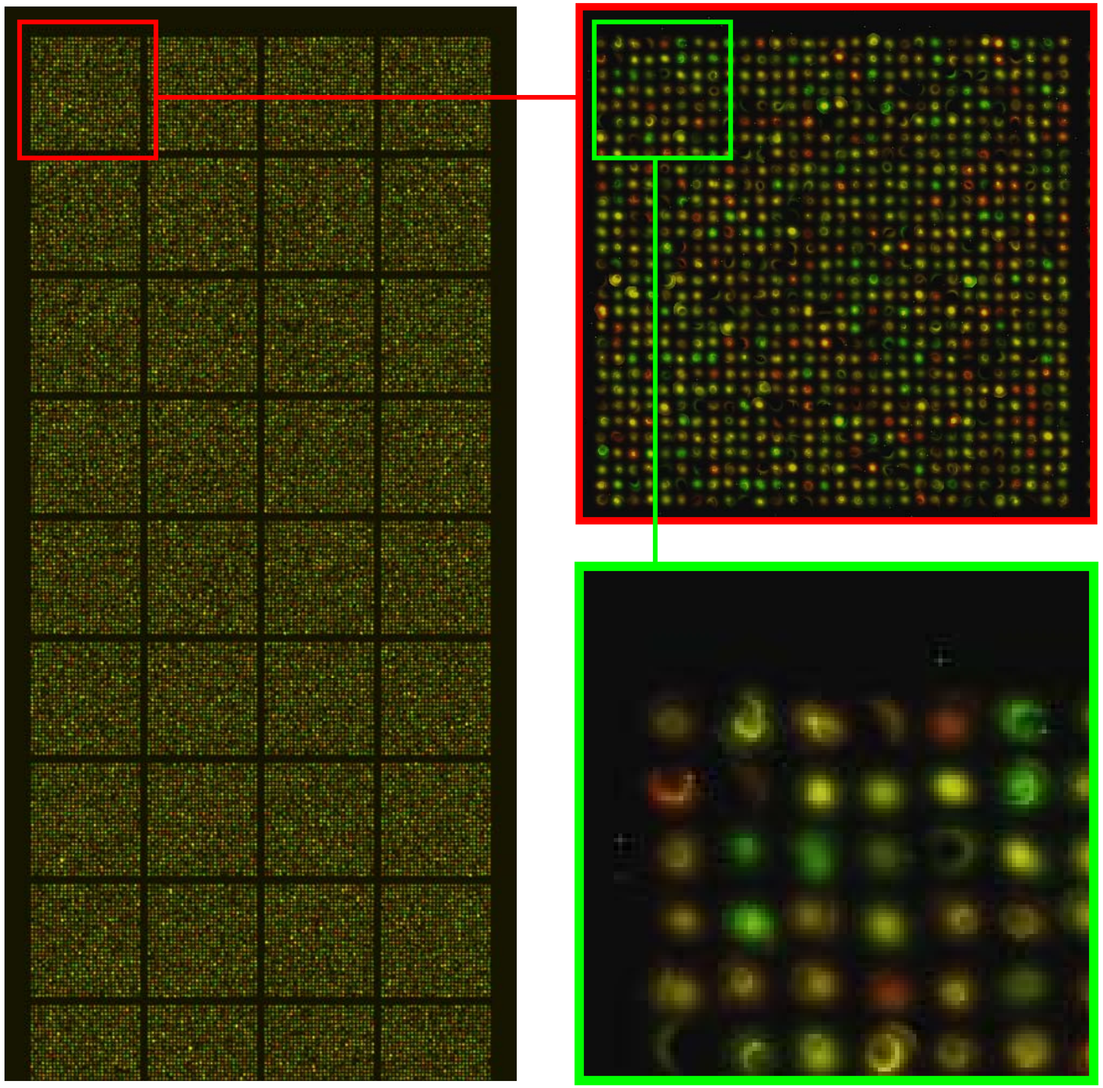
Name	Description	value
Mt	Top margin	110 pixels
Mb	Bottom margin	109 pixels
Ml	Left margin	99 pixels
Mr	Right margin	98 pixels
Br	No. of macroblocks rows	12
Bc	No. of macroblocks columns	4
BsY	Vertical space between individual macroblocks	27 pixels
BsX	Horizontal space between individual macroblocks	25 pixels
BspY	No. of spots rows in each macroblock	30
BspX	No. of spots columns in each macroblock	30
SaY	height of the area for the spot	14 pixels
SaX	width of the area for the spot	14 pixels
SoaY	height of the overlap area between adjacent spots	14 pixels
SoaX	width of the overlap area between adjacent spots	14 pixels
Nga	No. of Gaussian per spot	1÷3
Gpk	Peak value of the Gaussian distribution	0.4÷0.9
GpkY	Vert. shift from the spot area centre	-*SaY*/10÷*SaY*/10
GpkX	Hor. shift from the spot area centre	-*SaX*/10÷*SaX*/10

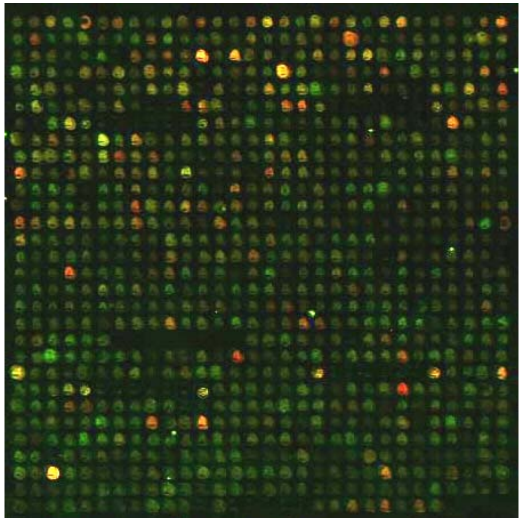
Simulation Parameters



σy	Standard deviation along y-axis	$SaY/6 \div SaY/5$
σx	Standard deviation along x-axis	$SaX/6 \div SaX/5$
ρ	Correlation factor	0.0÷0.15
Npl	No. of secant planes	1÷3
a	x coefficient of plane equation	-Gpk/10÷Gpk/10
b	y coefficient of plane equation	-Gpk/10÷Gpk/10
c	Known term of plane equation	0.0÷2.0xGpk/10
Pc	Clip probability	25 %
Pd	Dig probability	60 %
Dd	Dig depth	-0.6÷0.4
Pl	Lift probability	15 %
$Pdes$	DE spot percentage	20 %
Ndp	Distributed noise probability	99.95 %
Nlp	Local noise probability	0.05 %
Ndv	Distributed noise variation	-0.1÷0.2
Nlv	Local noise variation	0.5÷2.0
Nrf	Noise repartition factor between channels	0.0÷1.0

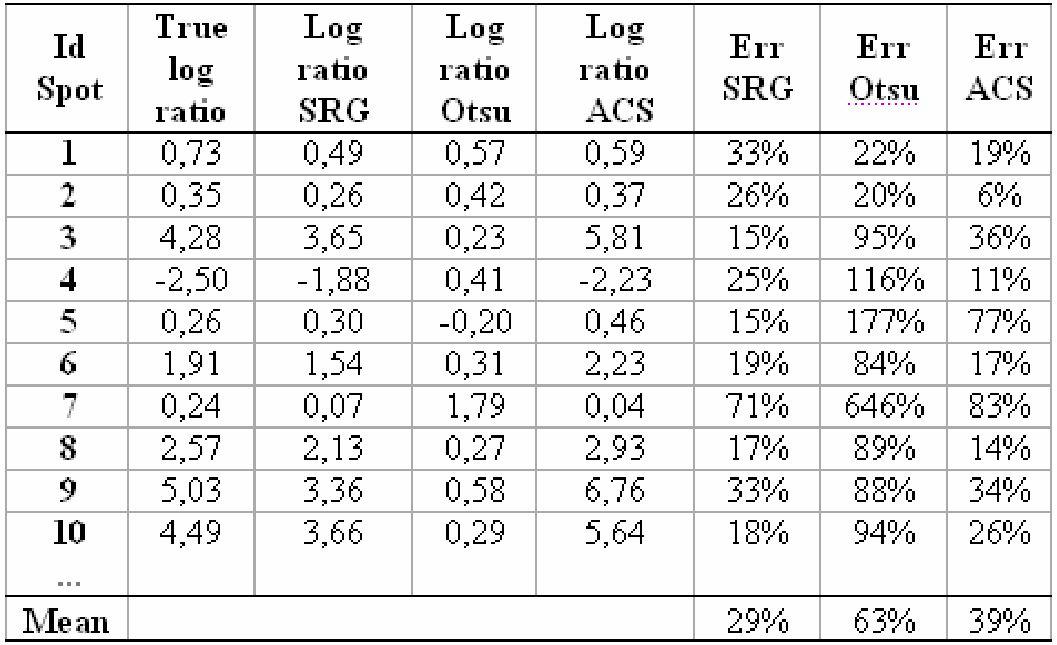
Simulated Image





Testing

- Simulated microarray image #28370 generated has been processed by three different software tools:
  - one based on seeded region growing algorithm (SRG)
  - one based on Otsu algorithm
  - one based on adaptive circle segmentation (ACS)



Id Spot	True log ratio	Log ratio SRG	Log ratio Otsu	Log ratio ACS	Err SRG	Err Otsu	Err ACS
1	0,73	0,49	0,57	0,59	33%	22%	19%
2	0,35	0,26	0,42	0,37	26%	20%	6%
3	4,28	3,65	0,23	5,81	15%	95%	36%
4	-2,50	-1,88	0,41	-2,23	25%	116%	11%
5	0,26	0,30	-0,20	0,46	15%	177%	77%
6	1,91	1,54	0,31	2,23	19%	84%	17%
7	0,24	0,07	1,79	0,04	71%	646%	83%
8	2,57	2,13	0,27	2,93	17%	89%	14%
9	5,03	3,36	0,58	6,76	33%	88%	34%
10	4,49	3,66	0,29	5,64	18%	94%	26%
...							
Mean					29%	63%	39%

Conclusions

- The proposed methodology allows to capture specificities of a given microarray experiment and to create simulated images to evaluate errors of analysis software tools
- The proposed model is simple and includes only parameters that can be determined in fast and robust way starting from real data
- Our idea is that does not exist a segmentation algorithm or software definitely better than all the others, but it is possible to choose the one assuring the best precision for given experiment and technology.