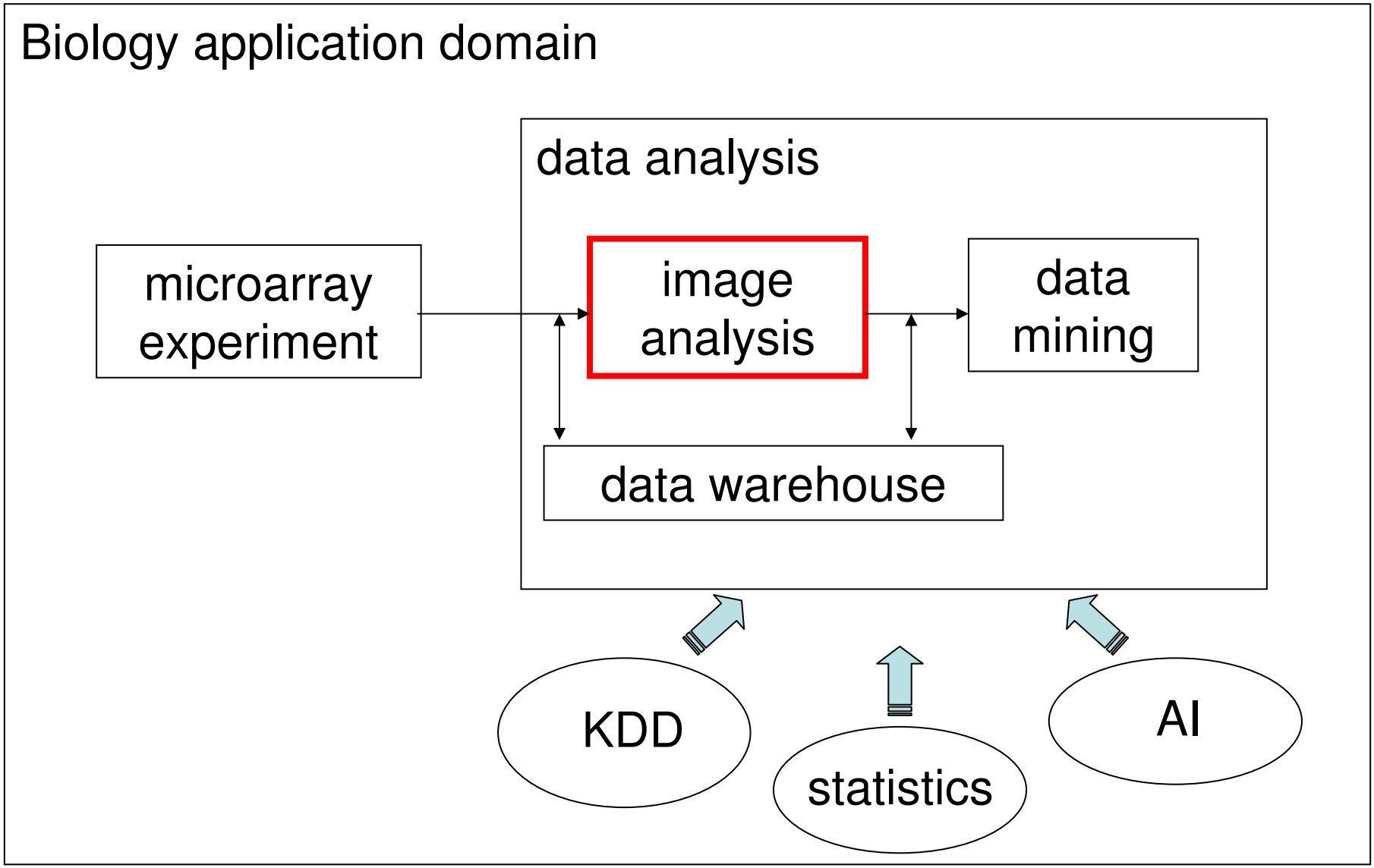
Microarray Image Analysis

Lucia Maddalena, ICAR-CNR

Alfredo Petrosino, University Parthenope of Naples

December 18, 2007

Main focus


Biology application domain
data analysis
microarray experiment
image analysis
data mining
data warehouse
KDD
statistics
AI

Microarray image analysis

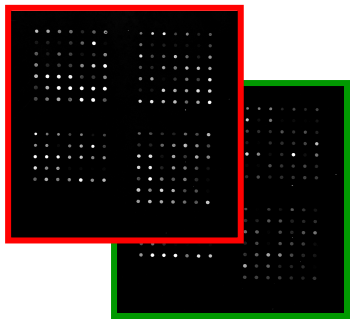
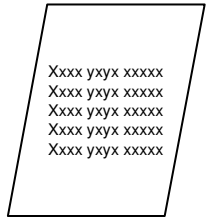


image analysis

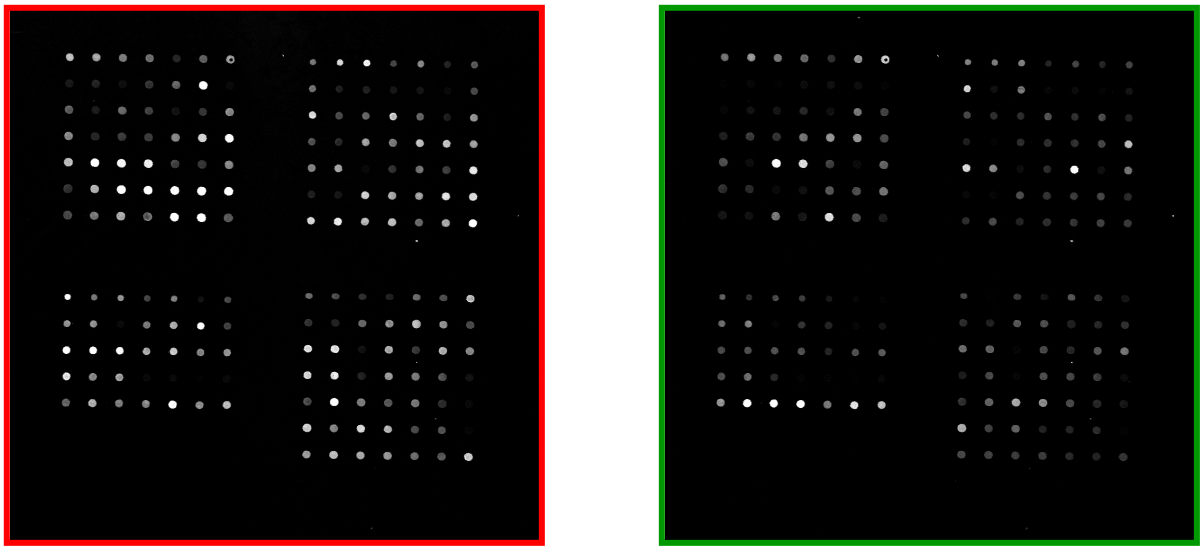


images from scanner

relative intensities for each spot

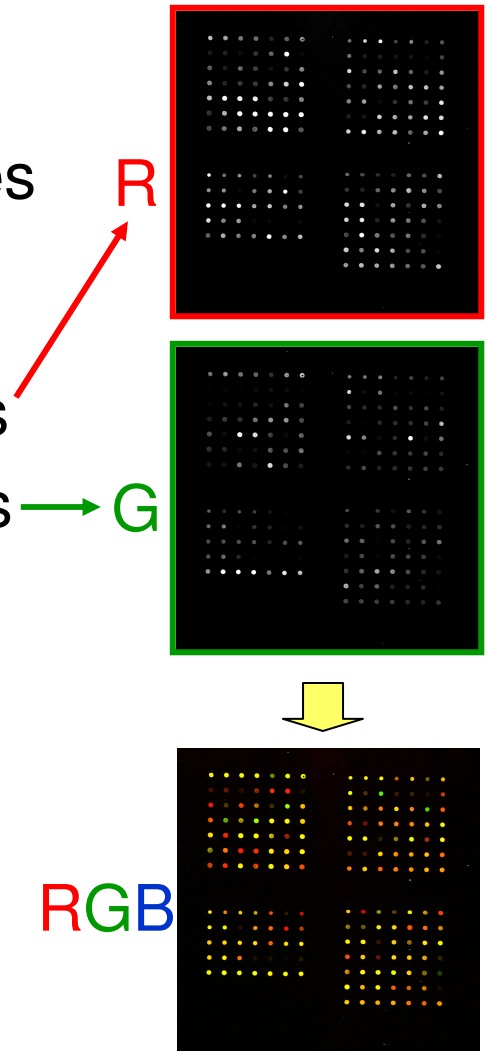
Images from scanner

- For a typical microarray experiment the scanner produces two TIFF (Tagged Image File Format) 16-bit (65’536 levels of grey) images, one for each fluorescent dye
- Commonly used dyes: Cy5 (red) and Cy3 (green)



Visualization of microarray images

- For visualization purposes (qualitative representation of results) a 24-bit RGB image displaying fluorescence intensities for both wavelengths is obtained by *pseudo-colour overlay*:
  - R: 8-bit (compressed) Cy5 intensities
  - G: 8-bit (compressed) Cy3 intensities
  - B: 8-bit zeros



Spot color	*Signal strength*	*Gene expression*
yellow	Control = perturbed	unchanged
red	Control < perturbed	induced
green	Control > perturbed	repressed

Visualization of microarray images (cont.)

- Sometimes, to enhance visual inspection, pseudo-colour overlay is modified
- e.g. *Scanalyze*:

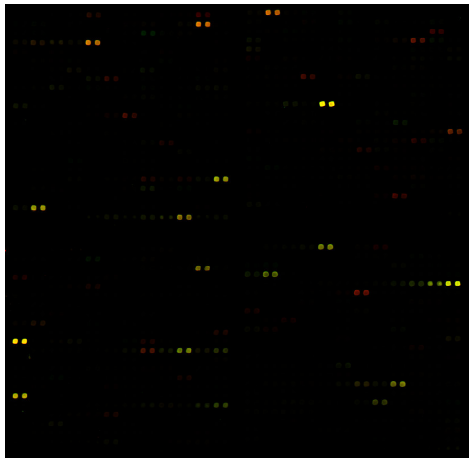
$$R = Gain * \frac{Cy5}{2^n} / Norm$$

$$G = Gain * \frac{Cy3}{2^n}$$

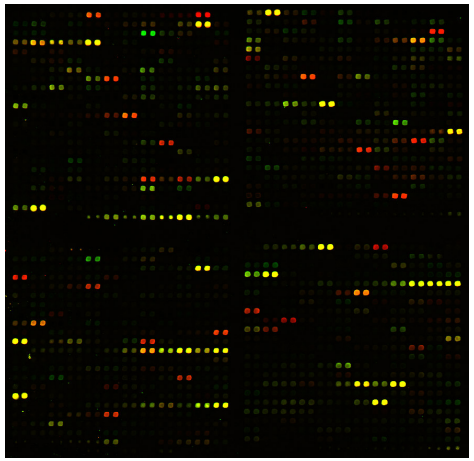
$$B = 0$$

- ➢ *Gain* controls brightness
- ➢ *Norm* controls balance between Cy3 and Cy5

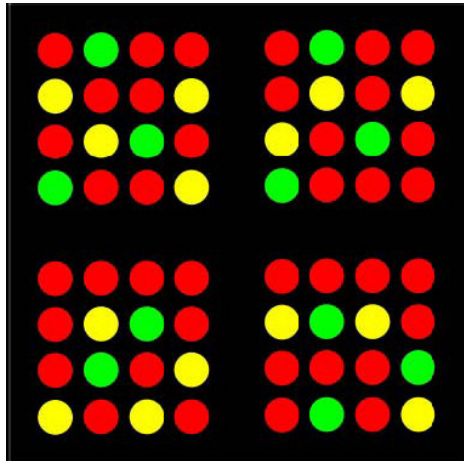
Gain =1
Norm =1



Gain =10
Norm =1



Ideal microarray image



Ideal microarray image
in terms of its image content:

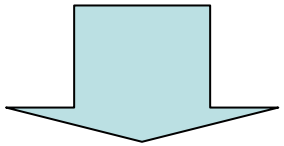
- deterministic grid geometry
- known background intensity with zero uncertainty
- pre-defined spot shape
- constant spot intensity that
  - is different from the background,
  - is directly proportional to the biological phenomenon, and
  - has zero uncertainty for all spots.
- for multi-channel microarray images same characteristics apply to each image channel, and channels are perfectly aligned

Alternatively, high statistical confidence in microarray measurements could be obtained acquiring images with really high resolution (very large number of pixels per spot). Problems: cost of microarray experiments, limited scanner resolution, cost of storage

Sources of microarray image variations

cDNA technology is a complex electrical-optical-chemical process

(spanning cDNA slide fabrication, mRNA preparation, fluorescence dye labeling, gene hybridization, robotic spotting, green and red fluorophores excitation by lasers, imaging using optics, slide scanning, analog to digital conversion, image storage and archiving)



Microarray image variations:

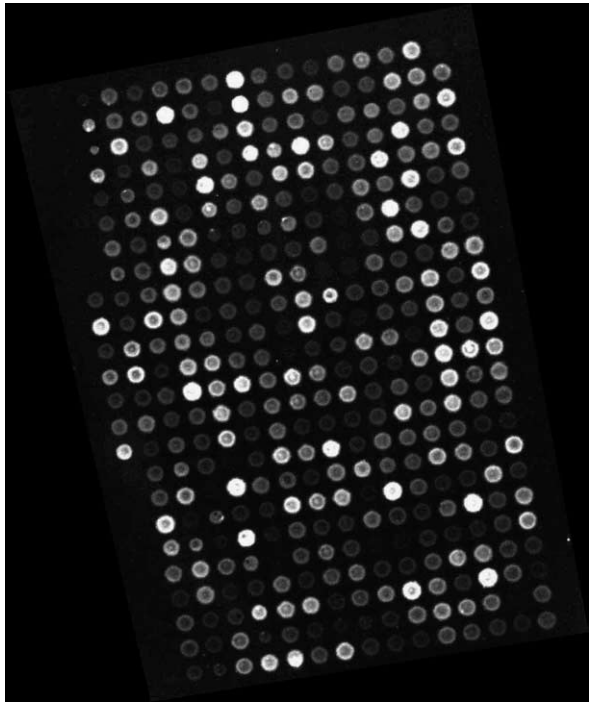
technology, grid geometry, background, spot morphology, foreground and background intensity

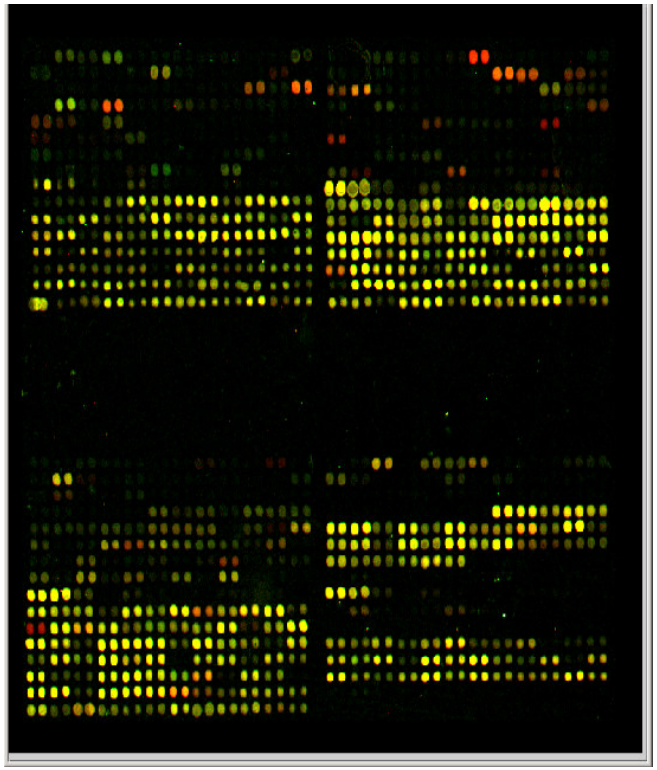
Variations due to technology

- **number of channels**: single-, double-, and multi-fluorescent microarray images
- **substrates**: glass, nylon membrane
- **labeling schemes**: radioactive, fluorescent
- **storage file format**: TIFF, SCN (Stanford Univ.)
- **data compression**: provided by file format
- **data accuracy**: number of bites per pixel
- …

Variations of grid geometry

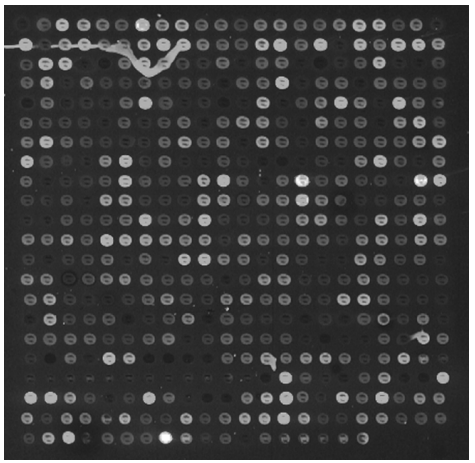
- multiple grids
- rotation (printing dipping pins or substrate)
- missing rows (low discrimination power for small surface areas in glass slides)

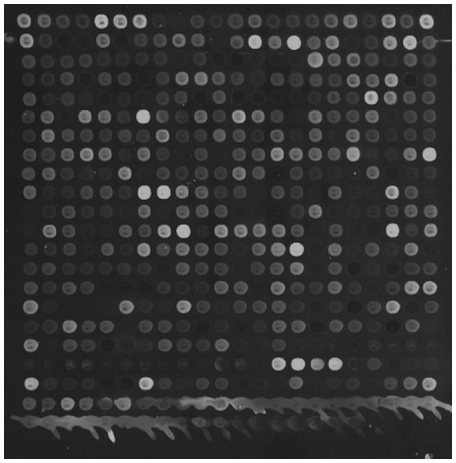




Variations of background

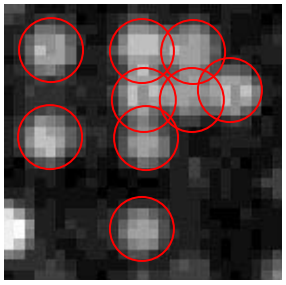
- microarray slide preparation (hybridization and spotting errors)
- inappropriate acquisition procedures (dust or dirt)
- image acquisition instruments (non-linearity of imaging components)
- …





Variations of spot location

- random deviation from ideal printing position (needles may vibrate slightly)
- warping (mechanical strain for nylon membrane)
- strong background signal (for fluorescent labeled probes)
- strong signal interference of neighboring spots (for radioactively labeled spots)
- …

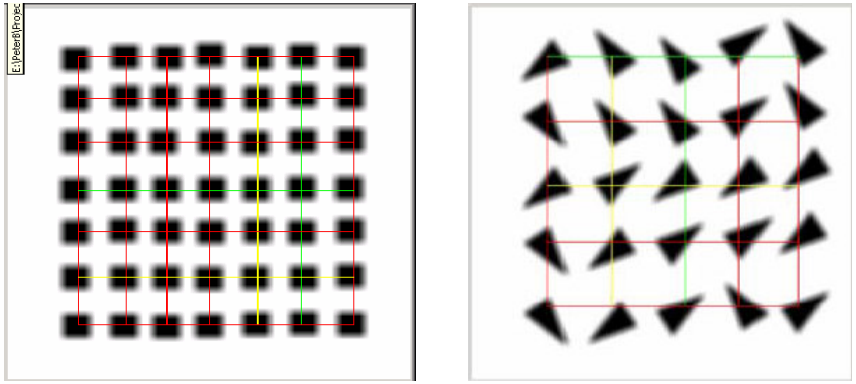


Spot overlap

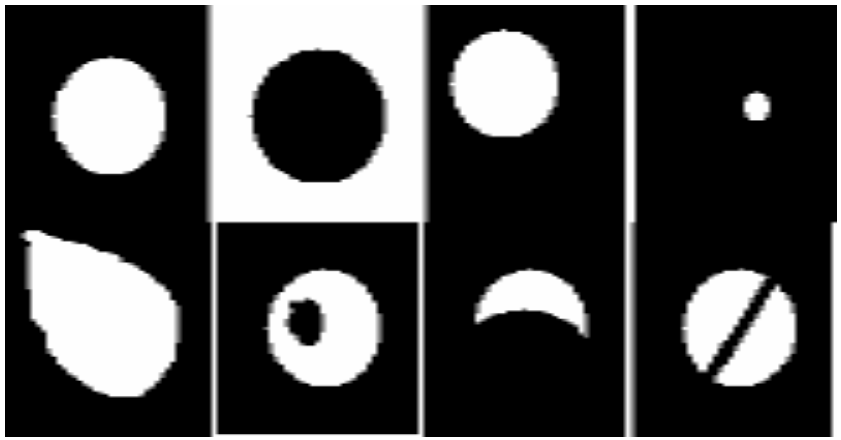
Variation of spot morphology

- Spot morphology other than circular (e.g. rectangles)

- Spatial and morphological variations of spots:
  a. regular spot
  b. inverse spot
  c. spatially deviating spot inside of a grid cell
  d. spot radius deviation
  e. tapering spot or comet shape
  f. spot with a hole
  g. partially missing spot
  h. scratched spot



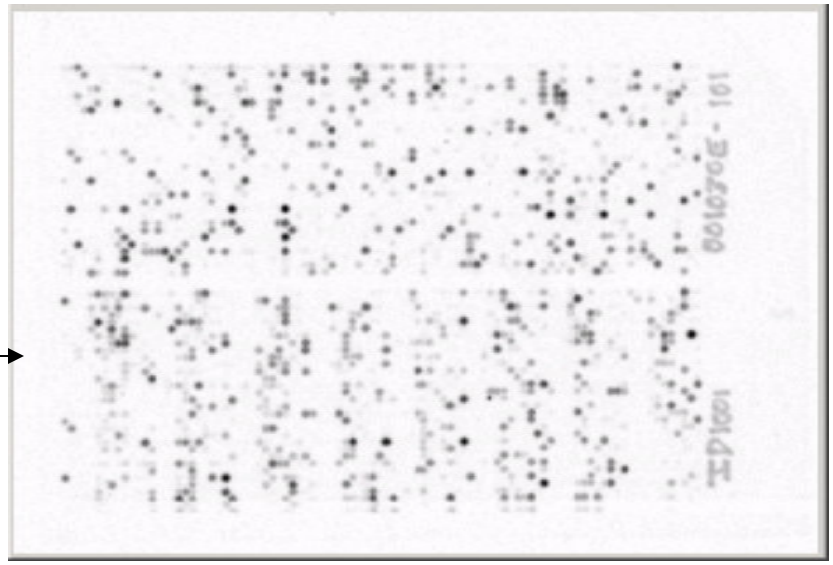
a. b. c. d.



e. f. g. h.

Variations of spot vs. background intensity

- Fluorescent labeling: dark background, bright spots
- Radioactive labeling: bright background, dark spots



Only background/foreground difference is relevant to biological meaning

BUT

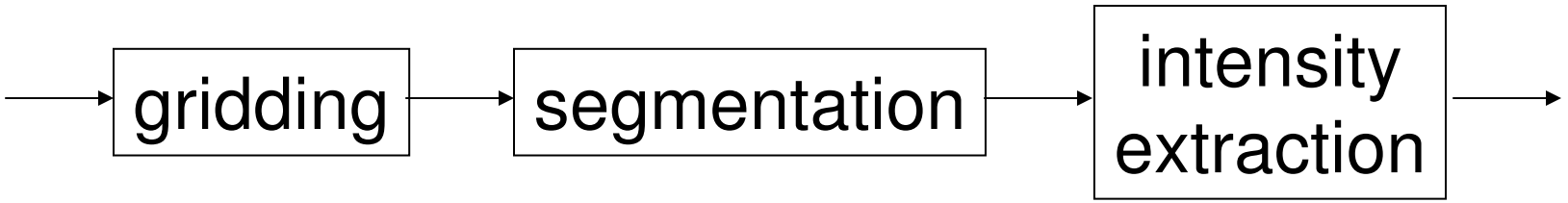
background and foreground variations affect the discrimination of the two classes

Microarray image analysis

- Goal: design automated microarray image processing algorithms that are robust to all variations:
  - technology
  - grid geometry
  - background
  - spot location
  - spot morphology
  - dark-bright scheme

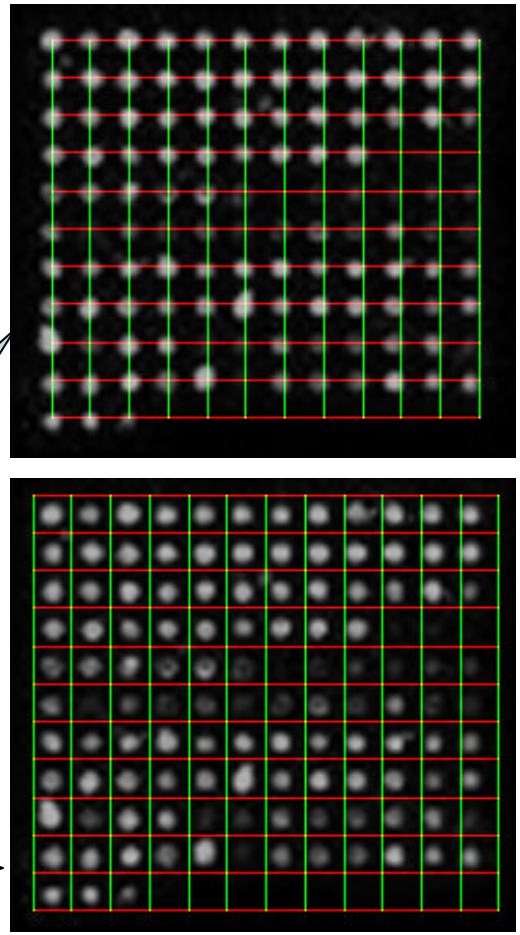
Steps in microarray image analysis

- gridding (*addressing*, *spot finding*, or *grid alignment*)
  - Assign coordinates to all the spots
- segmentation (*foreground separation*)
  - Classification of pixels either as foreground or as background
- intensity extraction (for each spot)
  - Foreground fluorescence intensity pairs (R, G)
  - Background intensities
  - Quality measures



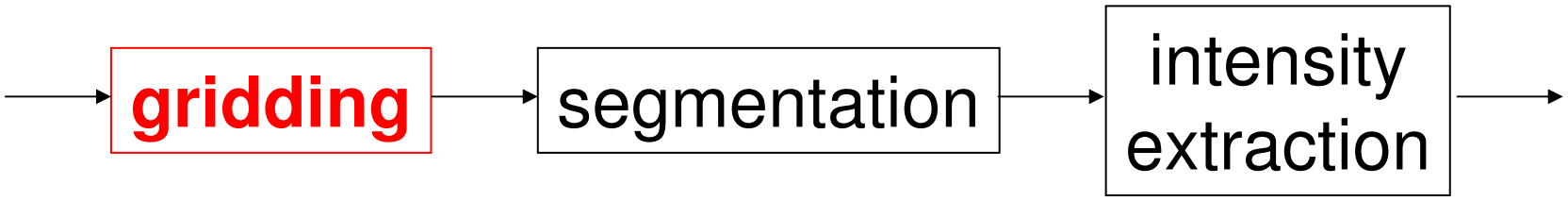
Gridding

- Objective: localize a 2D array of spots in a microarray scan before any information is extracted from the spots
- Localization is usually performed determining an orthogonal grid registered with the microarray image content, so that:
  - pairs of perpendicular lines intersect at the center location of each spot, or (equivalently)
  - each spot is centered in a net mesh



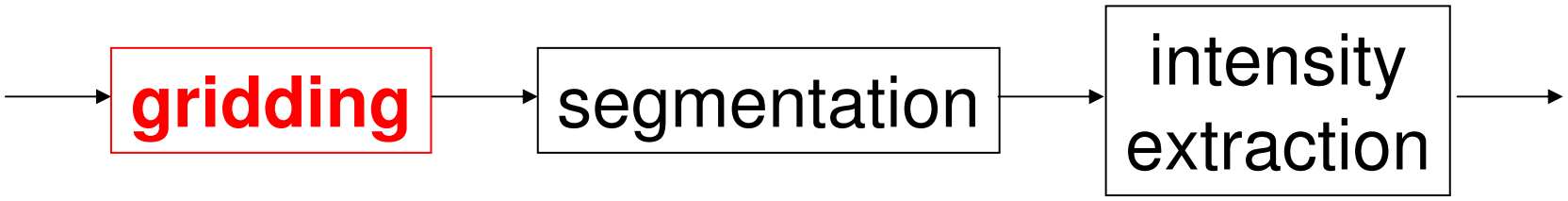
1st classification of gridding methods

- **Automation of methods:**
  - Manual: a grid template of spots is manually adjusted
    - time consuming, tedious, results not reproducible
  - Semi-automated: manual grid initialization followed by automated refinement
    - less time consuming, increase of results reproducibility
  - Fully automated: data driven without any human intervention, based on one-time human setup (for incorporating prior knowledge on image microarray layout)
    - reduced time consuming, high results reproducibility, but highly dependent on data content



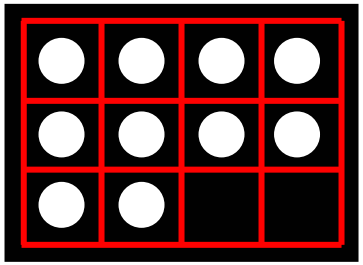
2nd classification of gridding methods

- **Image analysis approach:**
  - Template-based: define a template by specifying information about the microarray image and then adjust template location and parameters to match the spots (*grid refinement*)
  - Data-driven:
    - based on statistical analysis of 1D image projections, or
    - used as part of image segmentation algorithms

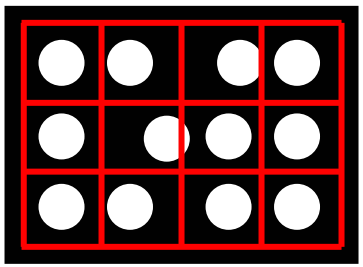


Template-based approaches

- Adopted in most literature and existing software packages (e.g., *GenePix Pro* by Axon Instruments, *ScanAlyze* or *GridOnArray* by Scanalytics)
- Also adopted for initial grid, to be later refined (e.g.,[Yang et al., '00], [Antoniol et al., '05])
- Pros:
  - incorporates knowledge about *ideal* grid
  - appropriate if measured grid geometry does not deviate too much from grid model defined by the template
- Cons: if measured spots are unpredictably regular, leads to inaccurate results or unacceptable costs for custom-tuned templates

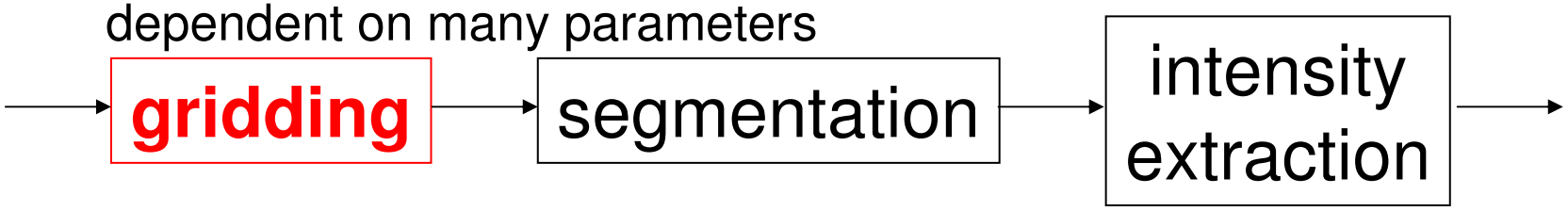


uniformly spaced templates



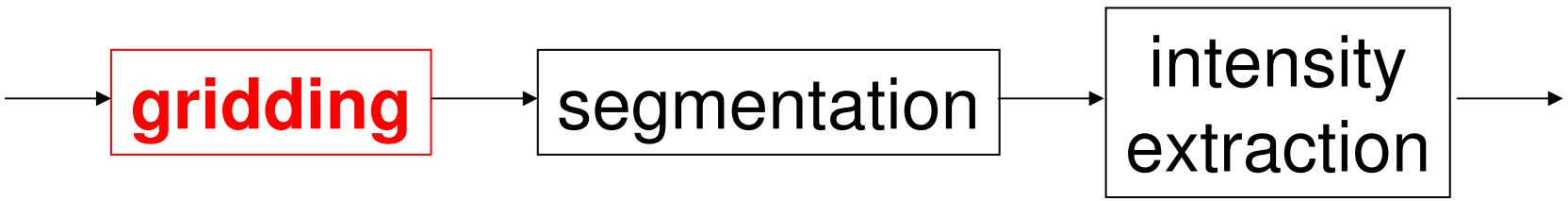
Data-driven approaches

- based on statistical analysis of 1D image projections:
  1. projection along rows and columns
  2. detection of local maxima in projections
  3. individuation of lines determined by local maxima, incorporating input parameters (e.g. number of lines), and computation of local maxima spacing
  4. intersection of orthogonal lines gives estimates of spot centers
- used as part of image segmentation algorithms:
  - Mathematical morphology [Angulo et al., '03; Hirata et al., '01], Markov Random Field (MRF) models [Katzer et al., '03; Antoniol et al, '04], graph models [Jin et al., '05]
- Pros: automatic alignment, also for non-uniform grids
- Cons: prone to misalignment due to spurious or missing spots, dependent on many parameters



Further problems related to gridding

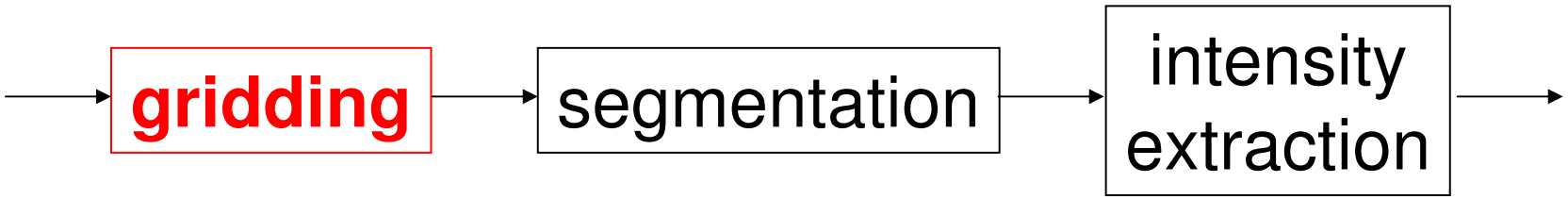
1. Processing of multi-channel images
2. Taking into account grid rotations
3. Accuracy vs. speed
4. Processing multiple grids



Channel fusion

How to process multi-channel images?

1. Gridding separate channels and then fusing grids
   - Pros: intensity variations are not propagated
   - Cons: computationally demanding, problems in merging multiple grids
2. Fusing channels and then gridding fused image
   - Pros: computationally less demanding, no need for merging multiple grids
   - Cons: intensity variations are propagated



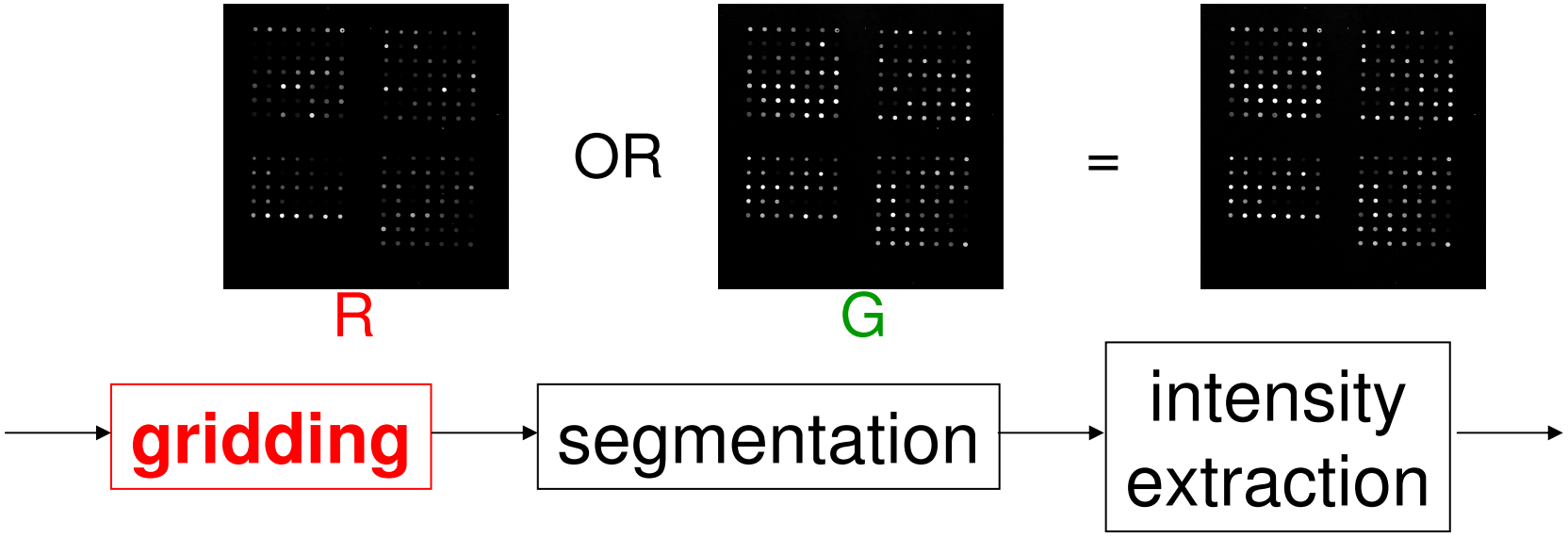
Channel fusion (cont.)

2. Fusing channels and then gridding fused image.

   Channels fusion by:

   a) Linear combination weighted by median values (e.g *Spot* [Yang et al., '02])

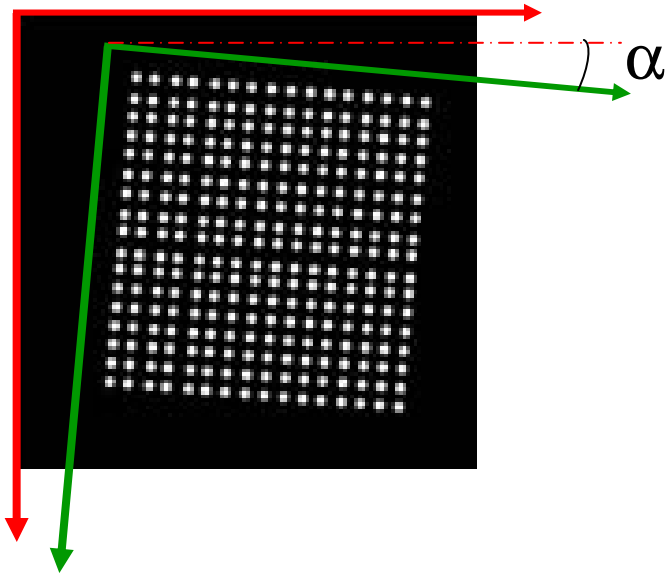
   b) Boolean OR function (e.g. *Gridline* [Bajcsy, '04])



Grid rotation

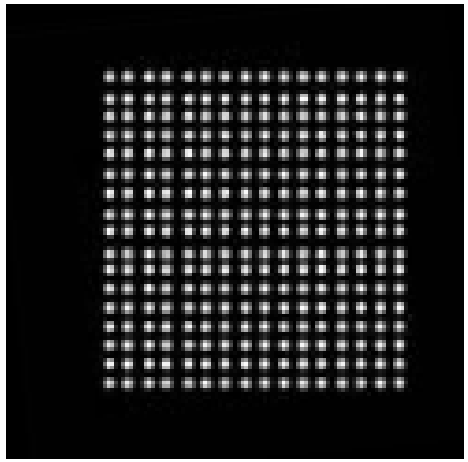
input images can be rotated

(coordinate system of robot printing the array may be slightly rotated with respect to the microarray image coordinate system)



compute α and:

- rotate input image by -α
- OR take into account α in subsequent steps



gridding → segmentation → intensity extraction →

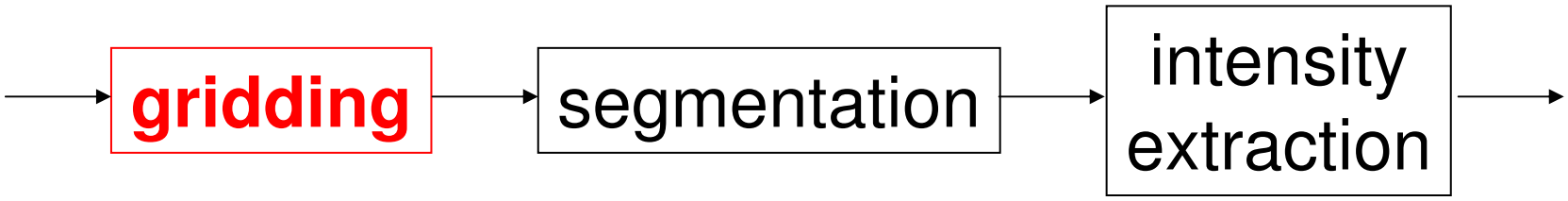
Grid rotation (cont.)

Exhaustive search of all expected rotation angles

(usually in $[-\pi/4, \pi/4]$ or in a user defined interval $[\alpha_{min}, \alpha_{max}]$)

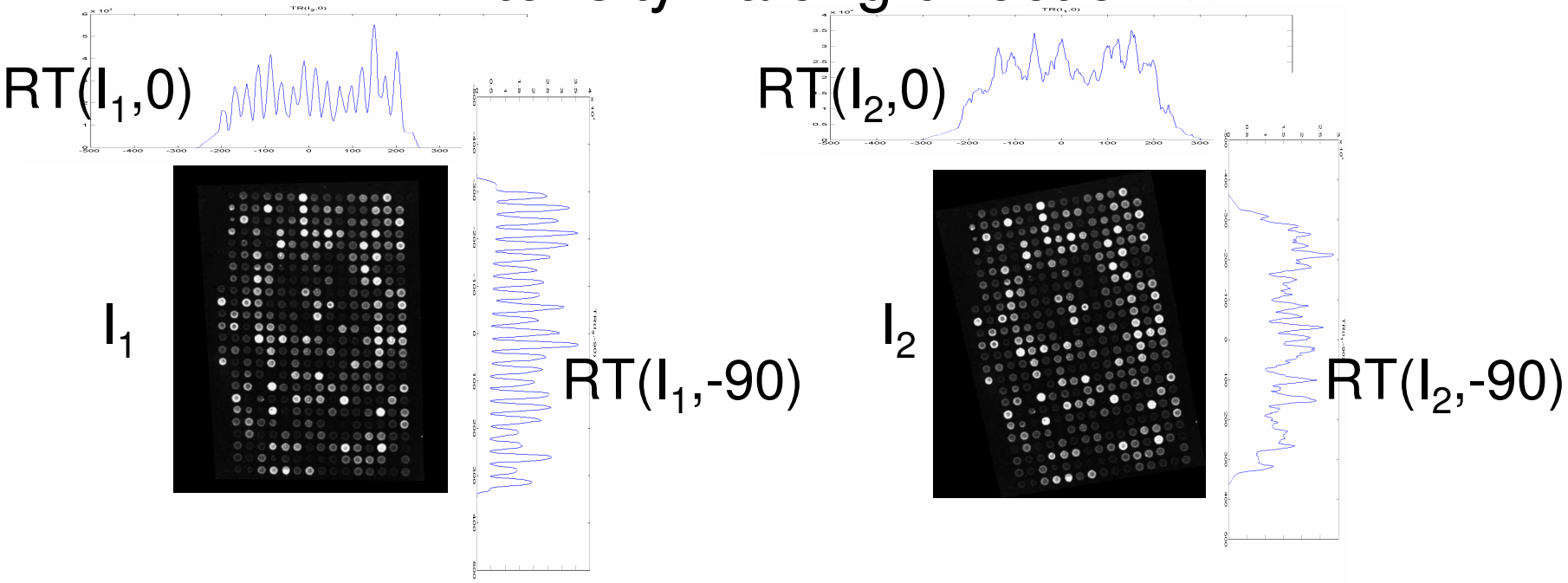
to find angle α that:

- minimizes a grid score function [Bajcsy, '04]
- maximizes median values of input image Radon Transform (RT) projection along direction α [Brandle et al., '03]
- maximizes values of OMT filtered input image RT projection along direction α [Antoniol et al., '05]
- maximizes combination of values of RT projection along the rows and columns of α-rotated binarized input image [Battiato et al., '07]



Grid rotation (cont.)

Radon Transform RT(I, α): projection of image intensity I along direction α

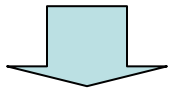


$I_2 = \text{rotate}(I_1, -\alpha)$ ⇒ $RT(I_2,\alpha) = RT(I_1,0)$, $RT(I_2,\alpha-90) = RT(I_1,-90)$

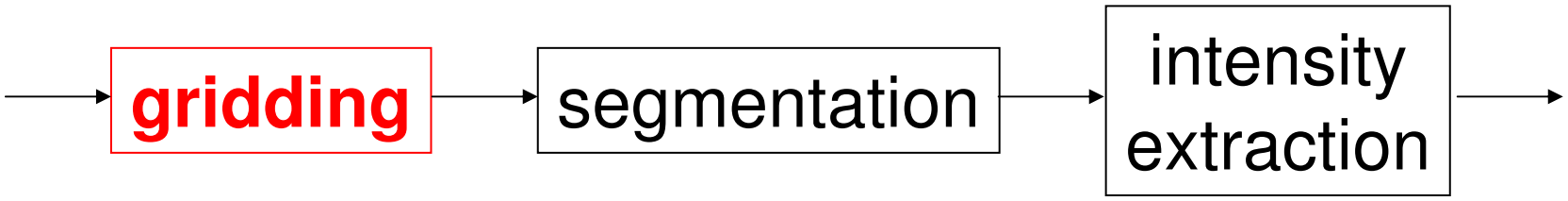
gridding → segmentation → intensity extraction →

Reducing the image size

Accessing and processing whole image (millions of pixels) is time consuming

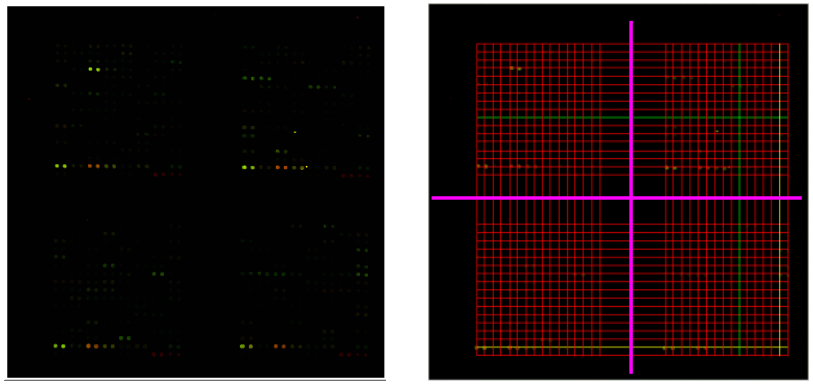


- Reduce input images to a smaller size, to reduce the quantity of information to be processed [Bajcsy, 2004]
  - *sub-sampling*: select a single pixel from a group and use it to represent the entire group
  - *down-sampling*: use a statistical sample (e.g. the mean) to create a new representation of an entire group of pixels
- Accuracy vs. speed

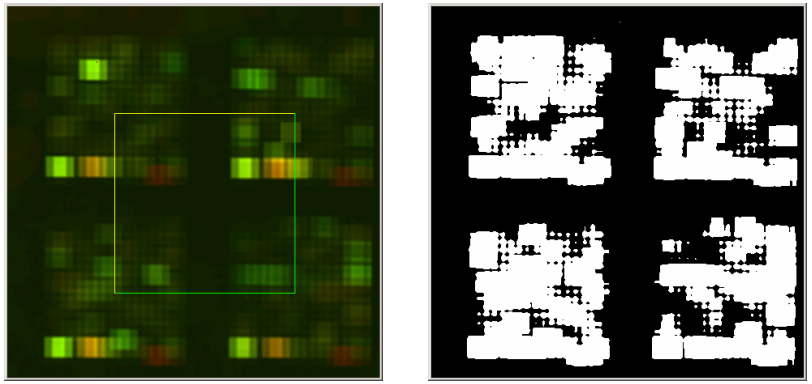


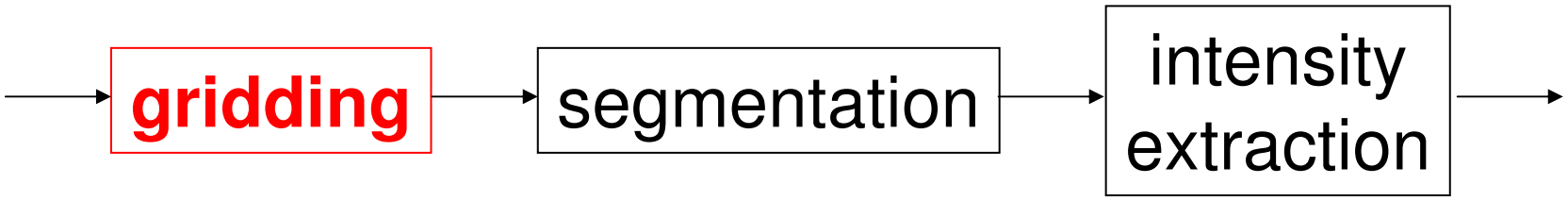
Processing Multiple Grids

- Line discontinuity approach [Bajcsy, '04]



- Filtering approach [Angulo et al., '04]





Segmentation

- Classification of pixels as foreground or background
  -> fluorescence intensities are calculated for each spot as measure of transcript abundance
- Production of a spot mask: set of foreground pixels for each spot

Classification of segmentation methods

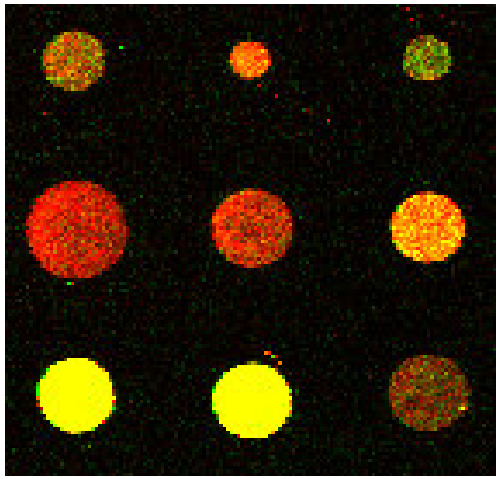
According to geometry of produced spots:

- fixed circle [e.g. ScanAlyze, GenePix, QuantArray]
- adaptive circle [e.g. GenePix, Dapple]
- adaptive shape [e.g. Spot]
- histogram-based [e.g. ImaGene, QuantArray, DeArray]

Fixed circle segmentation

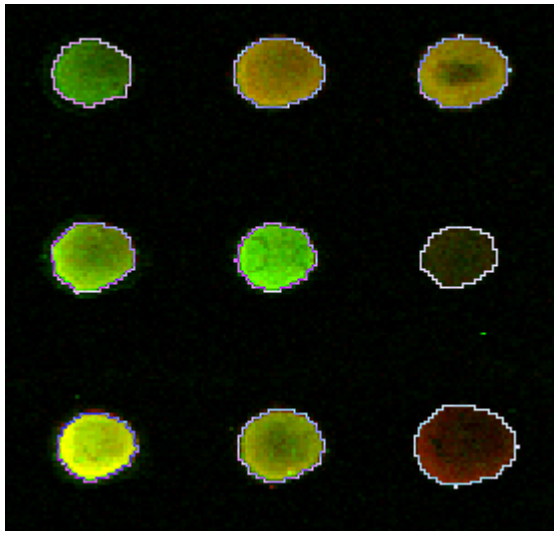
- Fits a circle with a constant diameter to all spots in the image [e.g. ScanAlyze]

- Pros: easy to implement
- Cons: the spots need to be of the same shape and size



Adaptive circle segmentation

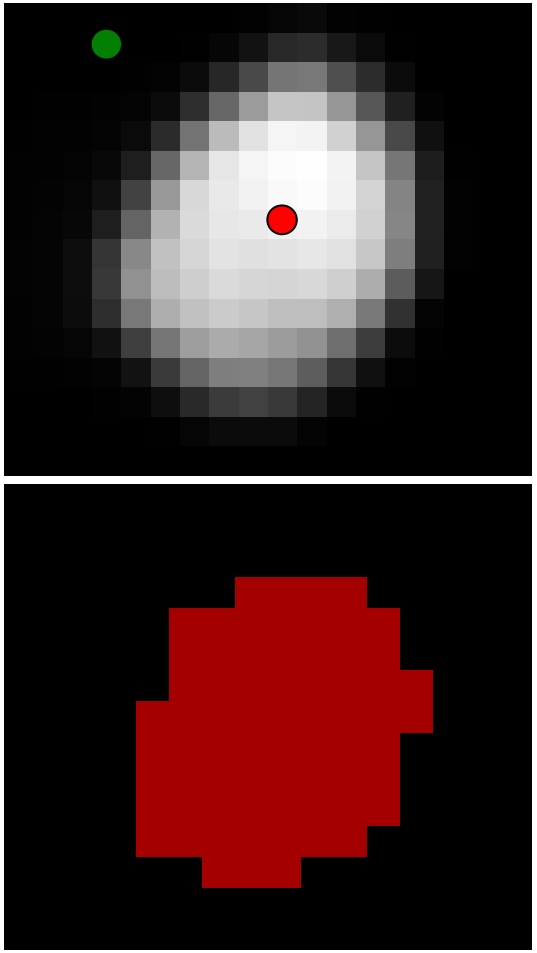
- Fits a circle to all spots in the image, estimating the circle diameter separately for each spot [e.g. GenePix – older versions]
- Problematic if spot exhibits oval shapes



[GenePix Pro 6.0]

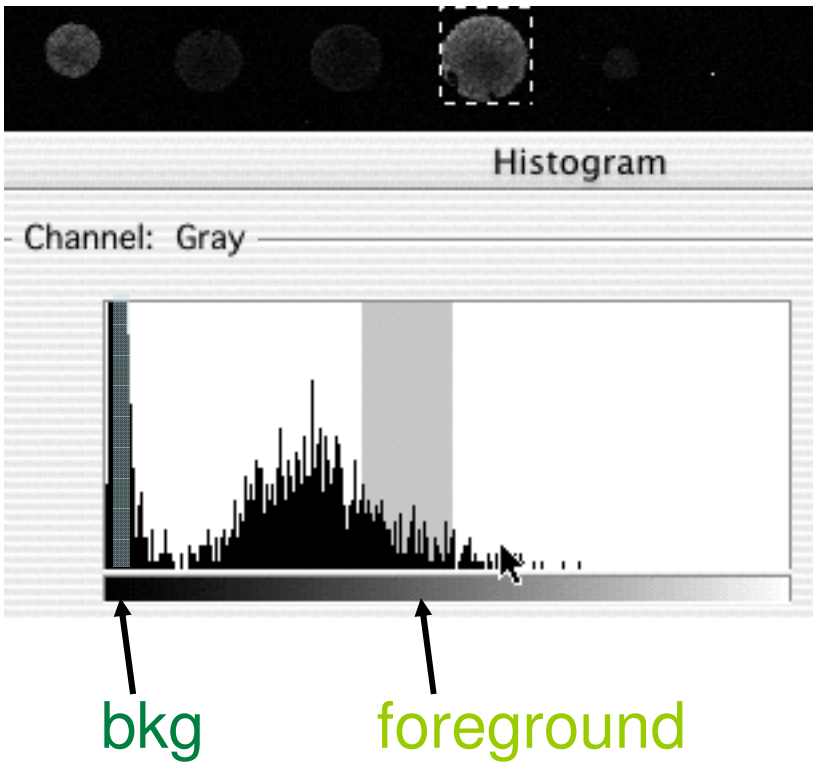
Adaptive shape segmentation

- Seeded Region Growing
  [Adams et al, 94]
  1. specification of seeds (number and position known)
  2. bkg and fg regions grow from the seeds outwards simultaneously preferentially according to the difference between a pixel's value and the running mean of values in an adjoining region



Histogram segmentation

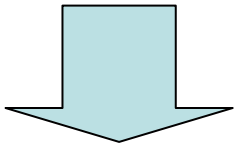
- Uses a target mask chosen to be larger than any other spot
- fg and bkg intensities determined from the histogram of pixel values for pixels within the masked area
- Pros: resulting spot masks are not necessarily connected
- Cons: Unstable when a large target mask is set to compensate for variation in spot size



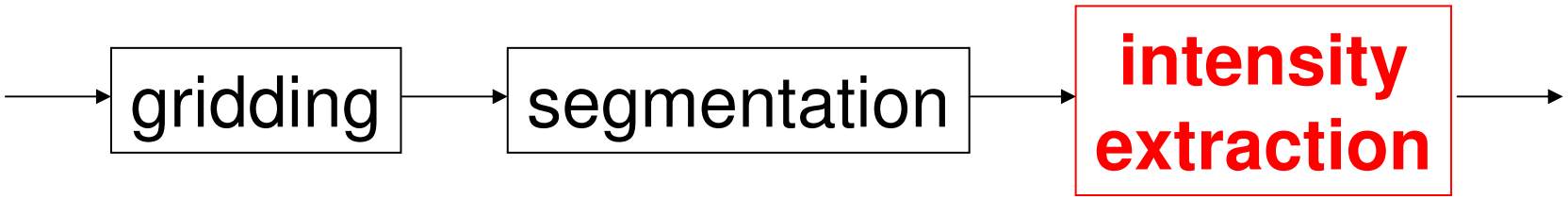
gridding → **segmentation** → intensity extraction →

Intensity extraction: spot intensity

Total amount of hybridization for a spot proportional to the *total fluorescence* at the spot



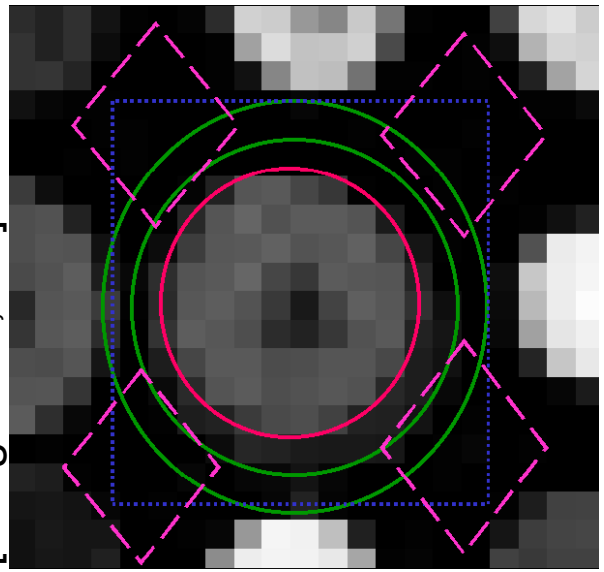
Spot intensity = mean (or median) of pixel intensities within the segmented spot mask



Intensity extraction: bkg intensity

spot measured intensity includes a contribution of non-specific hybridization and other chemicals on the substrate

background intensity = median of pixel intensities within selected regions surrounding the spot mask



[Yang et al., '00]

Blue square: Scanalyze

Green circles: QuantArray

Pink diamonds (valleys): Spot

gridding → segmentation → **intensity extraction** →

Our approach to gridding

Objective: construct an uniform orthogonal grid registered with the microarray image content

Input data:

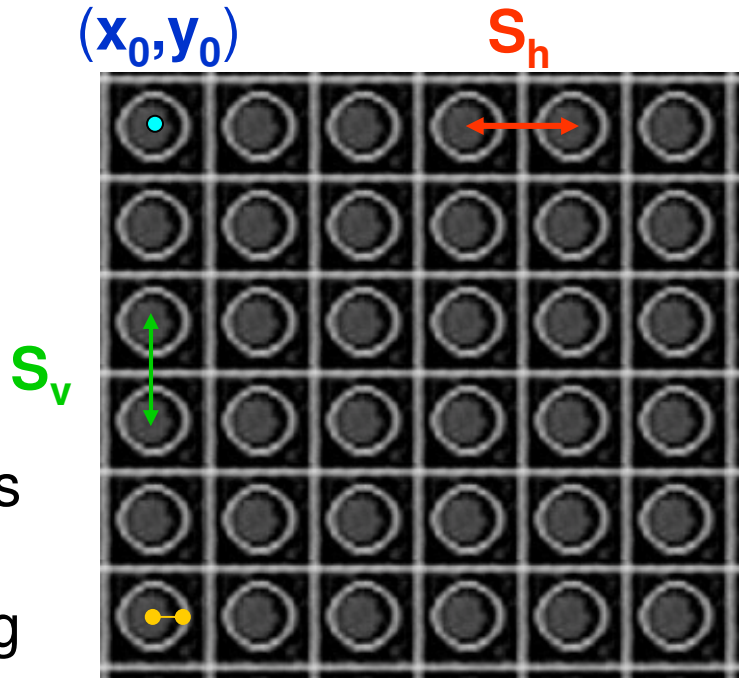
- **PR** = # of spots per row
- **PC** = # of spots per column
- **R** = spot radius

Output data:

- (x_0, y_0) = row and column coordinates of first spot center
- S_h, S_v = row and column grid spacing

Uniform grid uniquely determined:

$$(x_i, y_j) = (x_0 + i \cdot S_h, y_0 + j \cdot S_v) \quad \forall i, j$$



Our approach to gridding (cont.)

[L. Maddalena, A. Petrosino, '07]

- **Computation of (x_0,y_0)**: search for array **A** of centers of circles with approximate radius R and computation of minimum x and y coordinates:
  - Circular Hough Transform (CHT)
  - Orientation Matching Transform (OMT)
- **Computation of S_h, S_v**:
  - Average distances of centers in **A**
  - Most frequent distances of centers in **A**
  - Discrete Fourier Transform (DFT)

Computation of (x_0,y_0) based on CHT

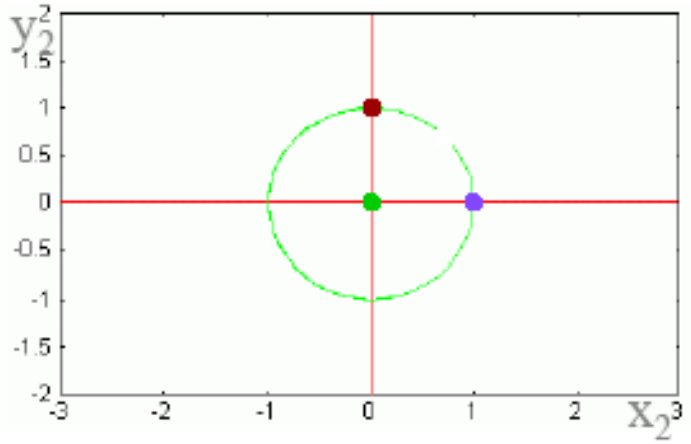
Circular Hough Transform

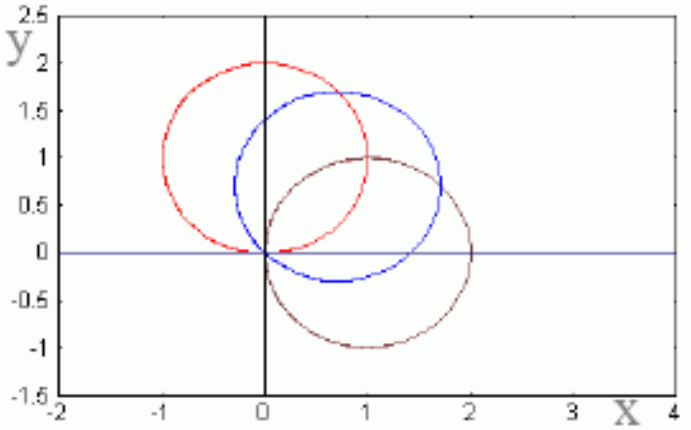
[Duda et al.,'72]

$$(x - x_c)^2 + (y - y_c)^2 = R^2$$

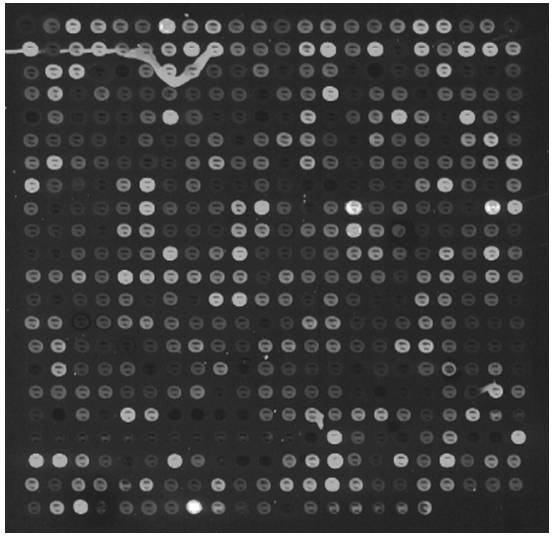
Computation of (x_0,y_0):

1. preprocessing of input image (edge detection and image binarization)
2. computation of the Hough space
3. search of maxima in the Hough space
4. (x_0,y_0) given by minimum x and y coordinates of maxima

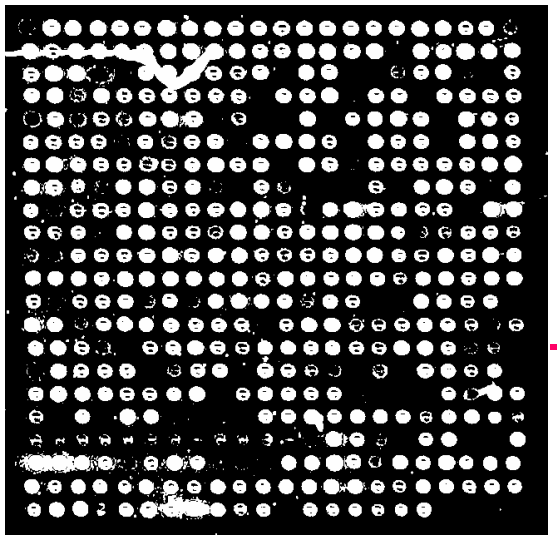




Computation of (x_0,y_0) based on CHT (cont.)



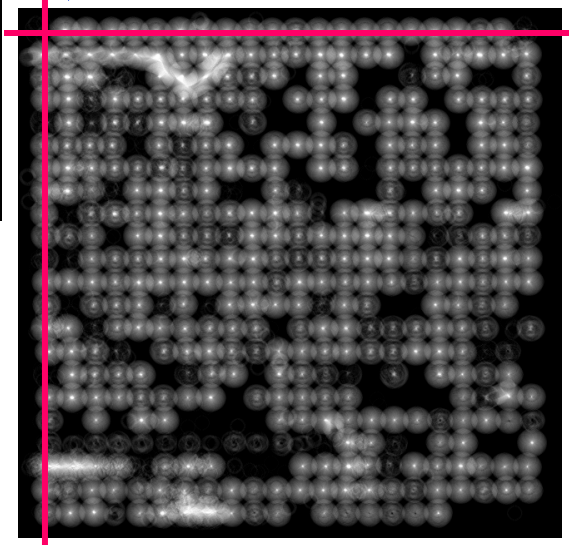
image



pre-processed image

(x_0,y_0)

Hough space



N.B. first spot barely evident in original image

Computation of (x_0,y_0) based on OMT

OMT (Orientation Matching Transform) [Ceccarelli, Petrosino, '99]

$$OM(u,v) = \frac{1}{2\pi(R-r)} \iint_{A_r^R(u,v)} \frac{\cos(\phi^*(x-u,y-v) - \phi(x,y))}{\sqrt{(x-u)^2 + (y-u)^2}} dxdy$$

$A_r^R(u,v)$ = annulus of radii r and R centered in (u,v)
$\Phi^*(x,y)$ = orientation of gradient of ideal circle centered in (0,0) of radius $(x^2+y^2)^{1/2}$
$\Phi(x,y)$ = orientation of image gradient at (x,y)

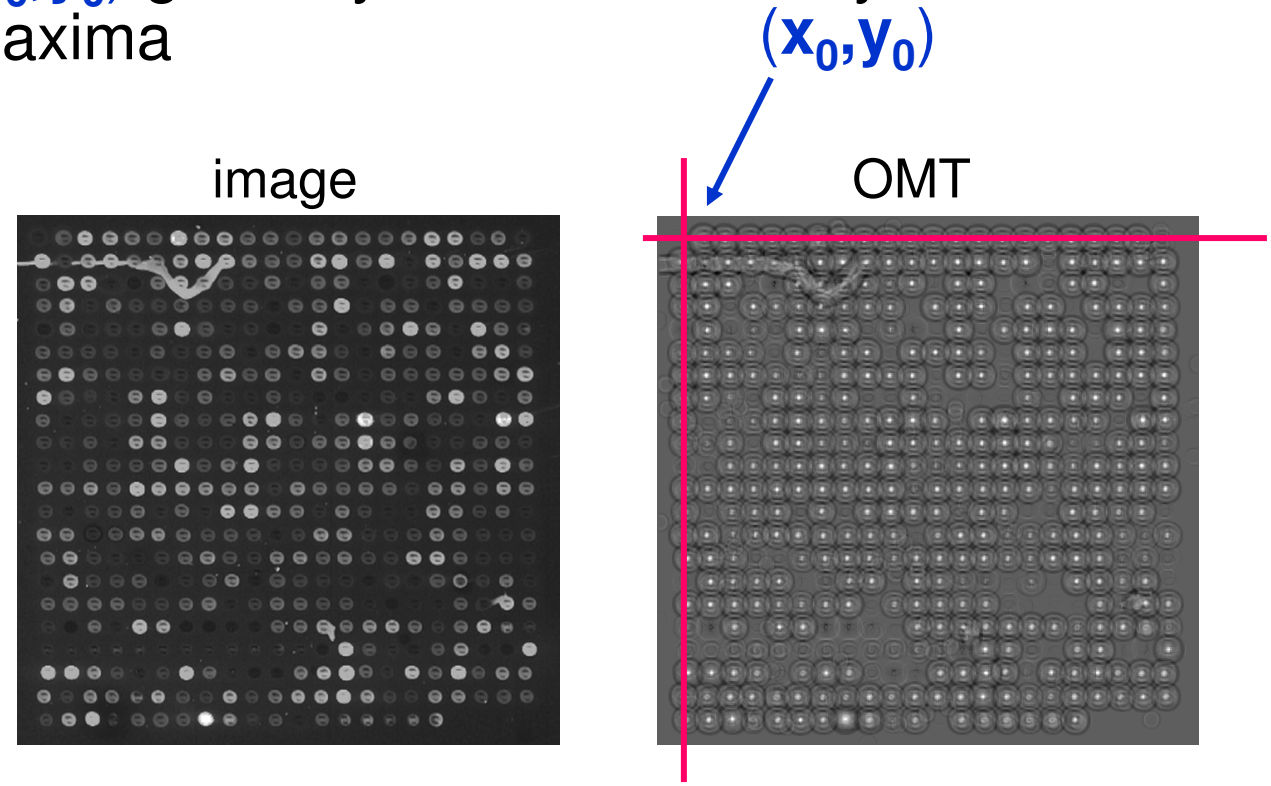
OM(u,v) represents the evidence that pixel (u,v) is the center of a circular object

- extension of CHT
- correlation-based transform
- invariant to contrast changes
- allows to deal with in the same manner with different radii
- can be tailored to recognize clear spots on dark bkg and viceversa

Computation of (x_0,y_0) based on OMT (cont.)

Computation of (x_0,y_0):

1. computation of the OMT
2. search of maxima in the OMT
3. (x_0,y_0) given by minimum x and y coordinates of maxima



Computation of S_h and S_v

1. Given the array A of maxima (either from CHT or OMT), S_h and S_v computed as:

   a) average (horizontal and vertical) distances of centers in A

   b) most frequent (horizontal and vertical) distances of centers in A

2. Discrete Fourier Transform

Computation of S_h and S_v based on DFT

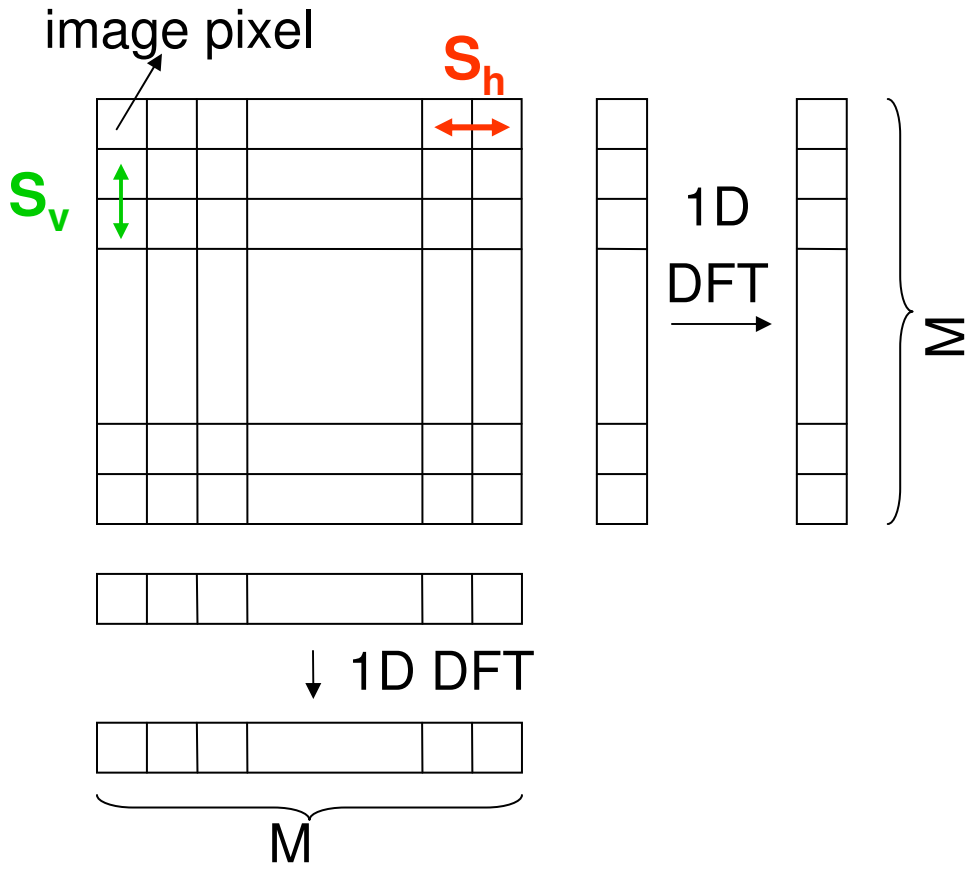
Compute S_v and S_h using 1D DFTs (Discrete Fourier Transforms) of image projections along rows and columns

$$F(u) = \frac{1}{M}\sum_{x=0}^{M-1} f(x) e^{-j2\pi ux/M}$$

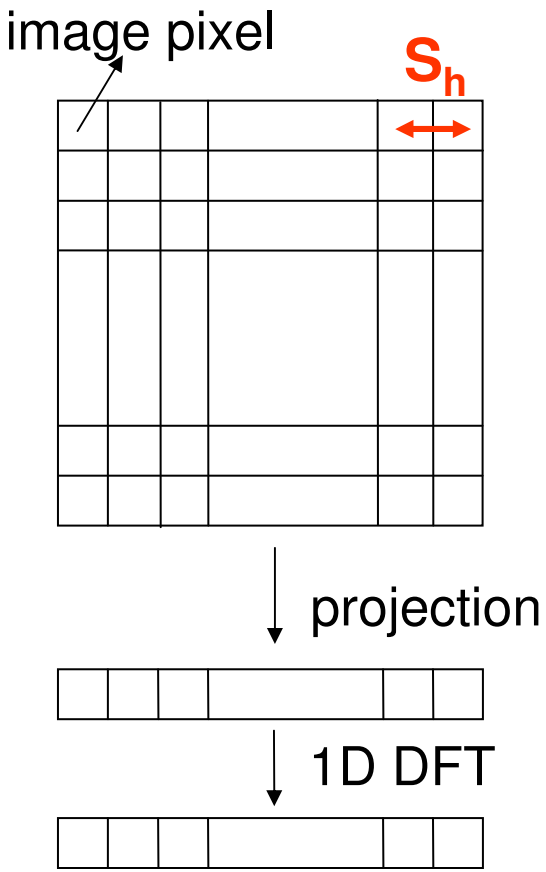
Regular spacing of f(x) results in a local maximum of F(u) at frequency $freq_{max}$

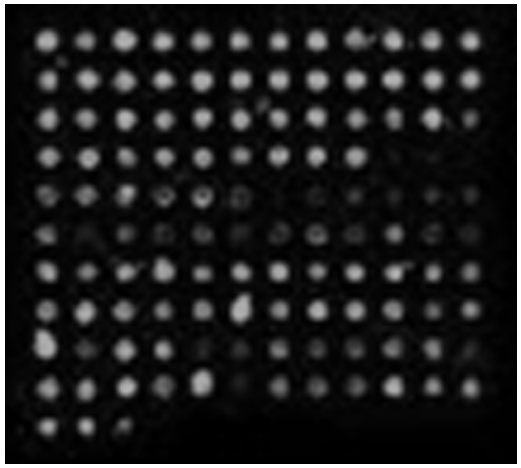
(Row or column) spacing is estimated by:

$$S = \frac{M}{freq_{max}}$$

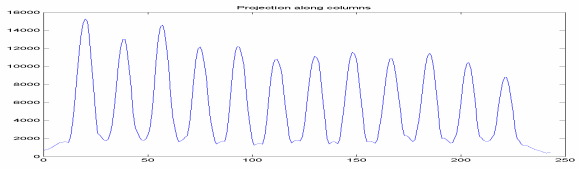


Computation of S_h and S_v based on DFT (cont.)

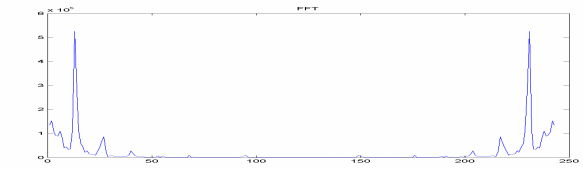




image



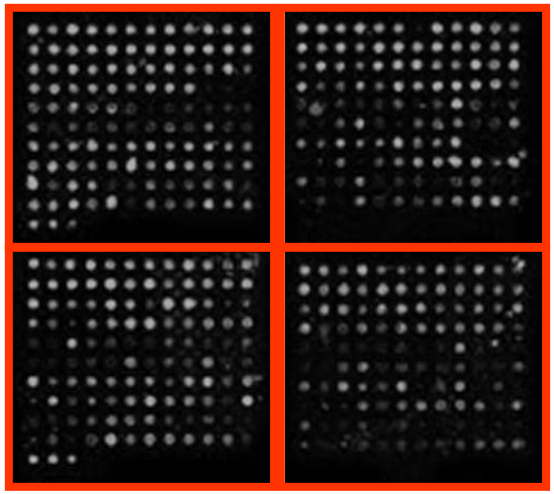
projection

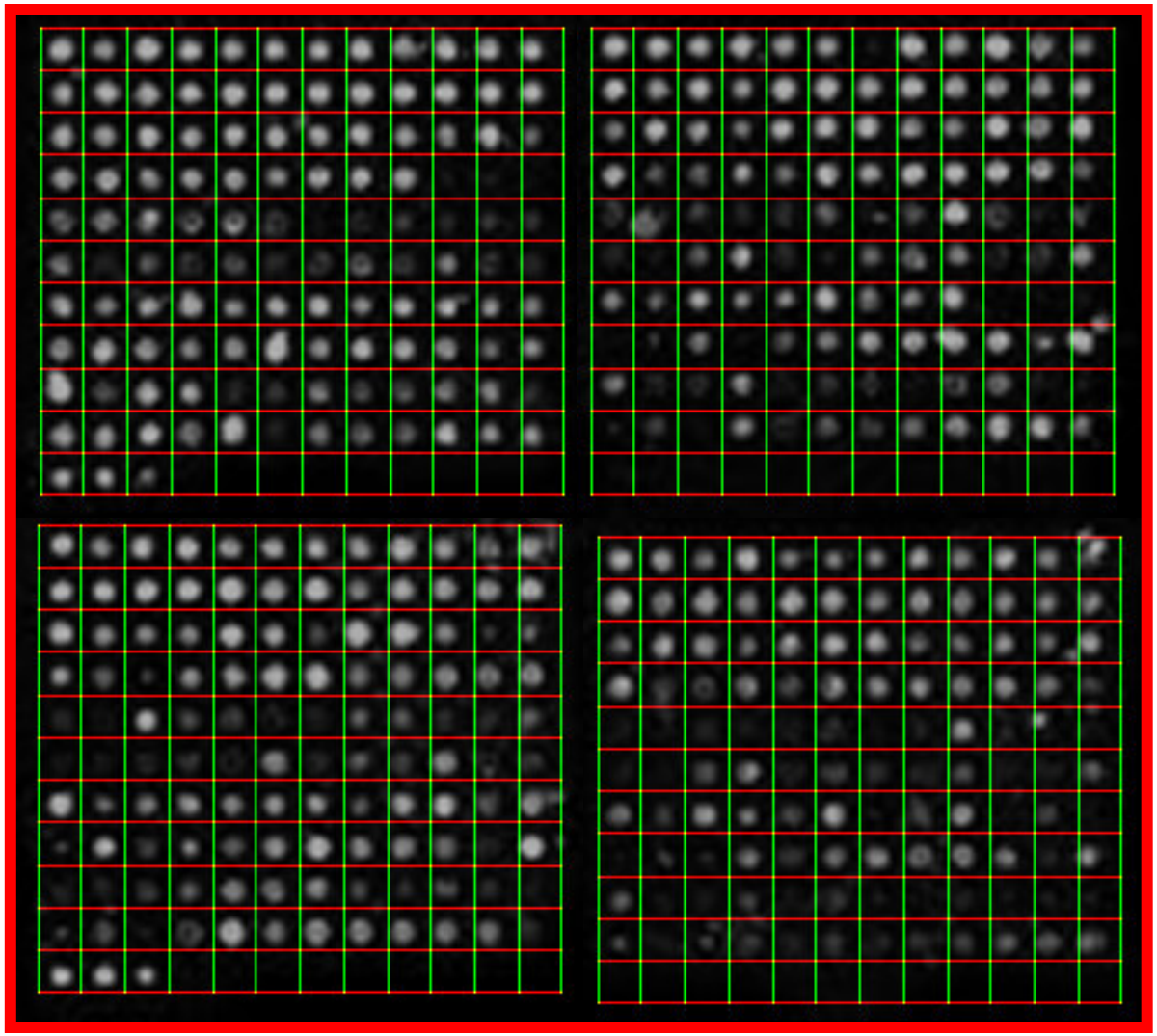


1D DFT

Some experiments

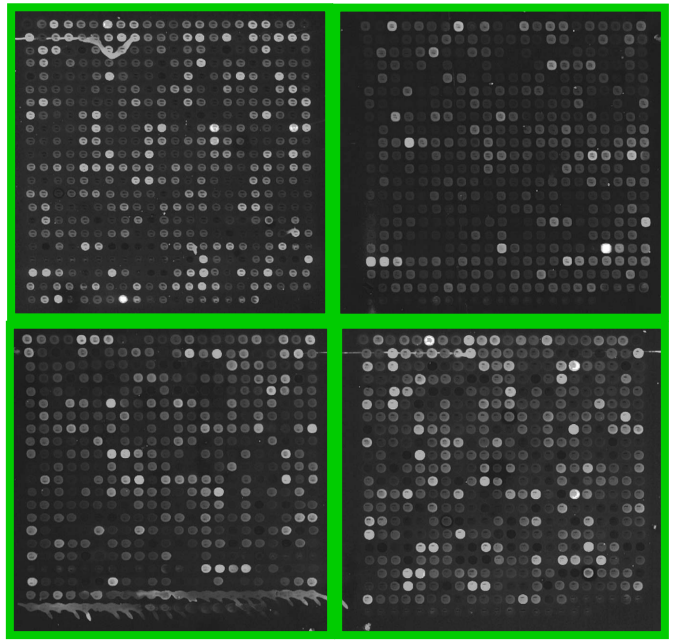
Microarray A
(PR=11, PC=12, R=6)

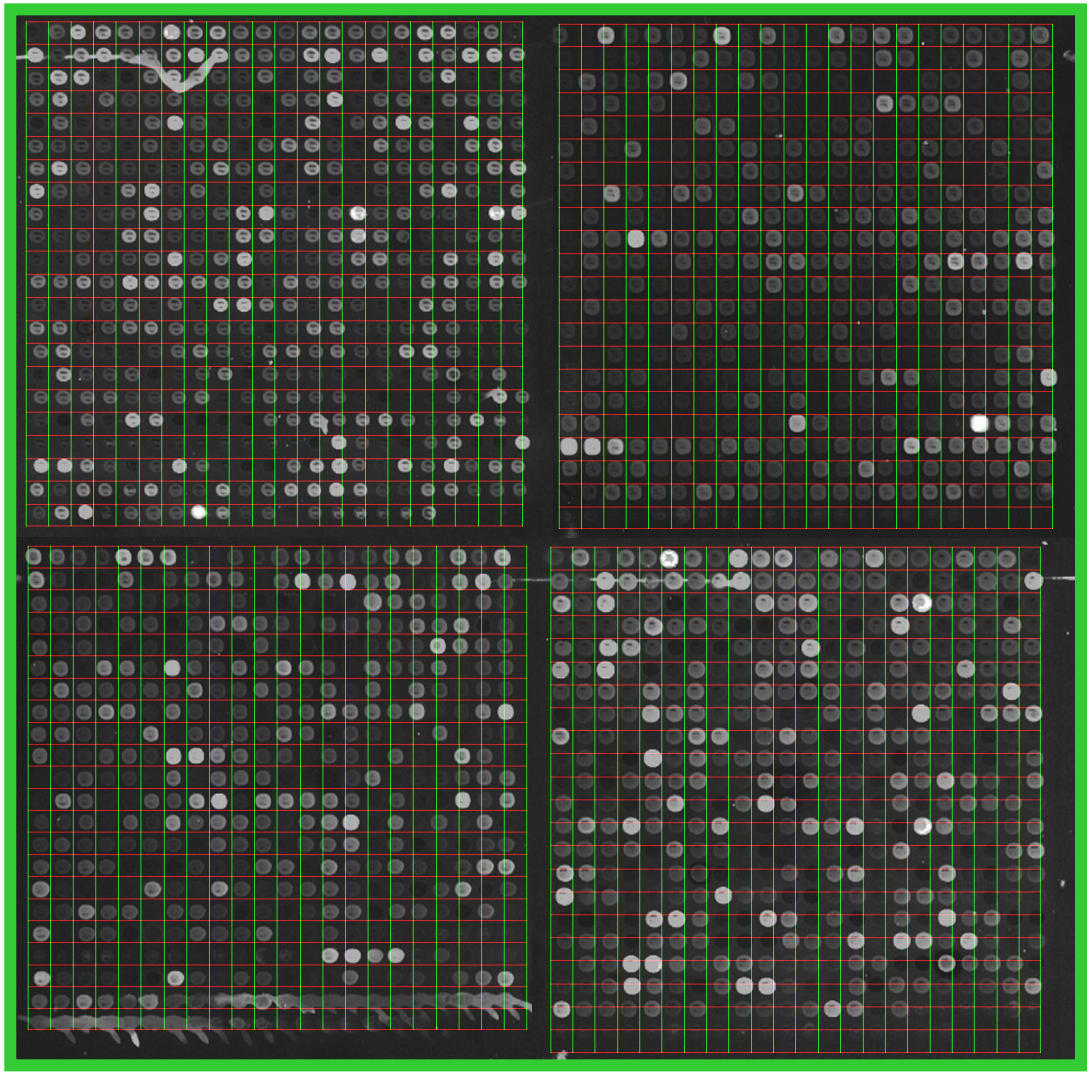




Some experiments (cont.)

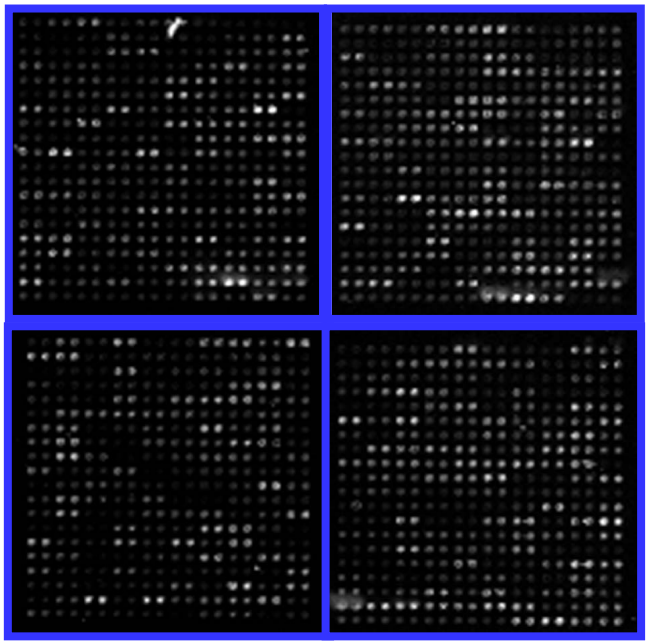
Microarray B
(PR=22, PC=22, R=9)

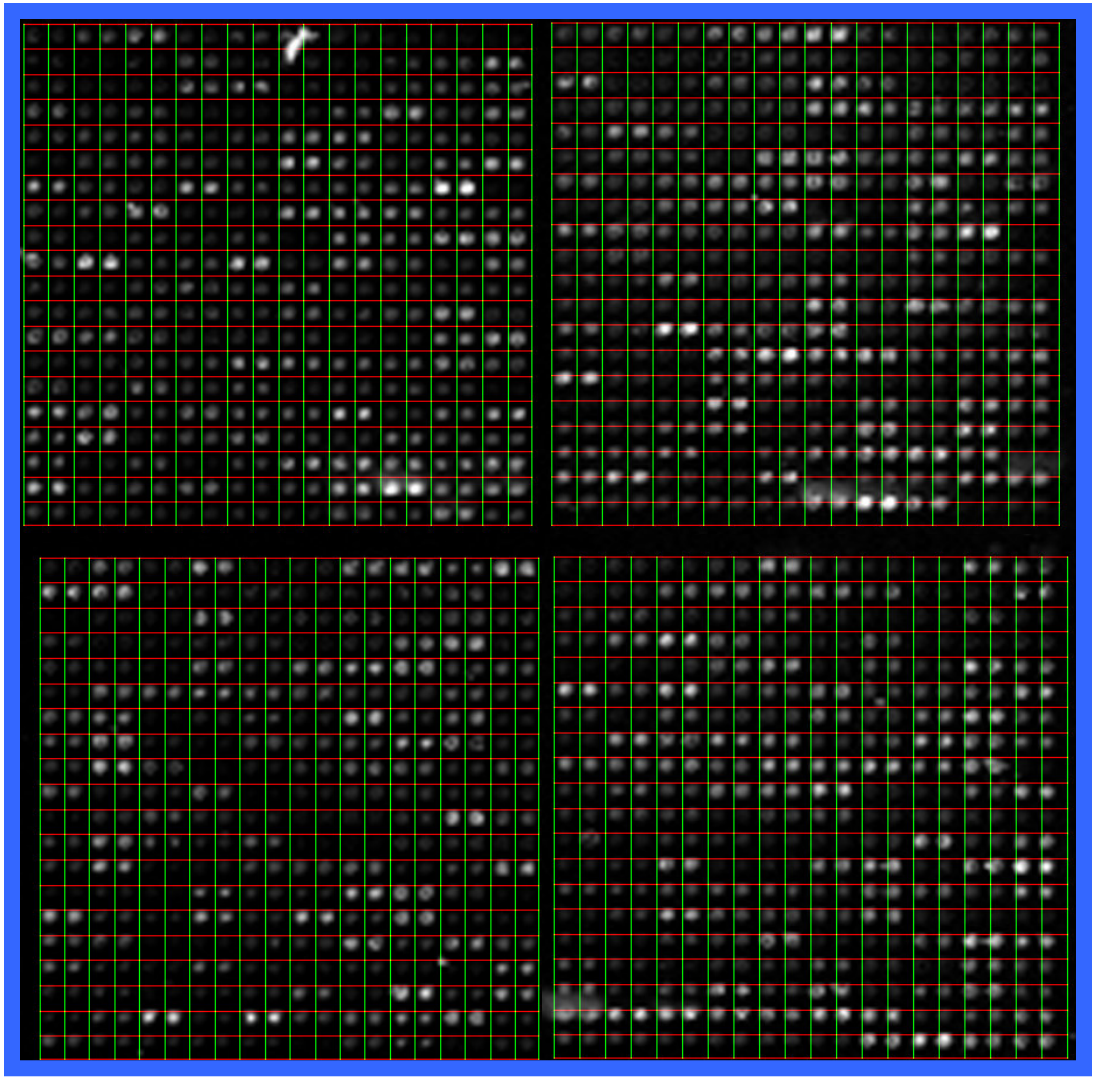




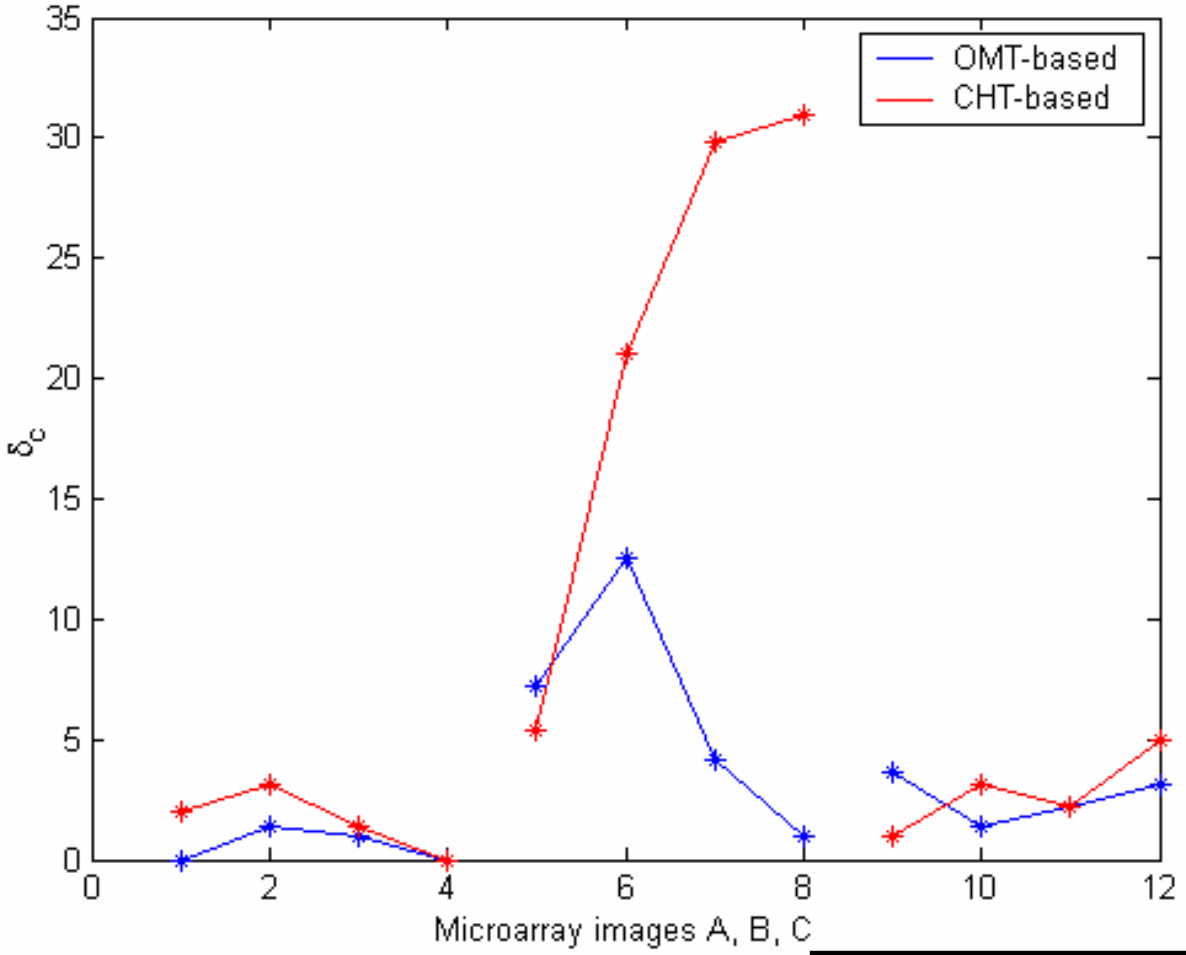
Some experiments (cont.)

Microarray C
(PR=20, PC=20, R=7)





Errors in computation of (x_0,y_0)

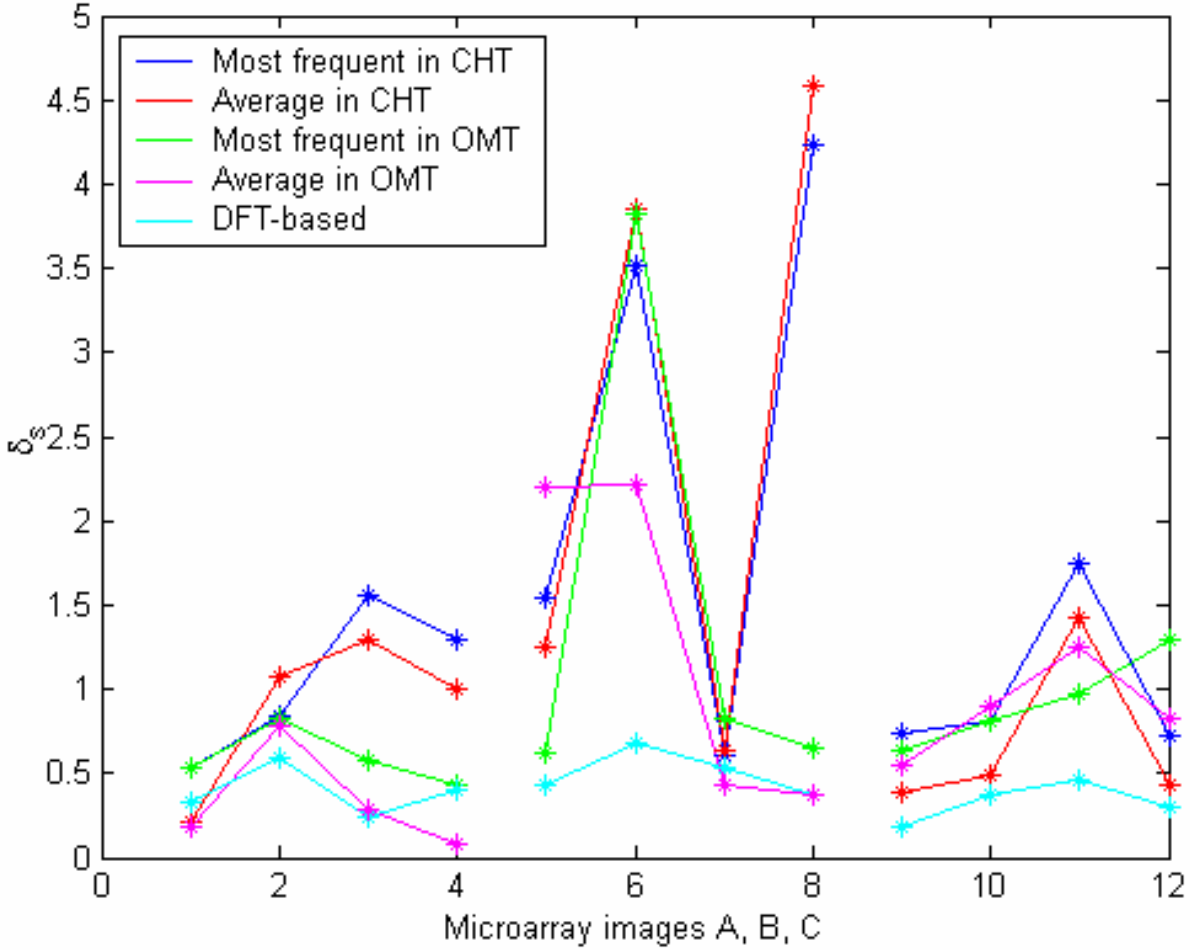


(x_0, y_0) = coordinates computed with ScanAlyze

$(\overline{x_0}, \overline{y_0})$ = coordinates computed with our methods

$$\delta_C = \sqrt{\left(\overline{x_0} - x_0\right)^2 + \left(\overline{y_0} - y_0\right)^2}$$

Errors in computation of S_h and S_v



(S_h, S_v) = average periods computed with ScanAlyze

$(\overline{S_h}, \overline{S_v})$ = periods computed with our methods

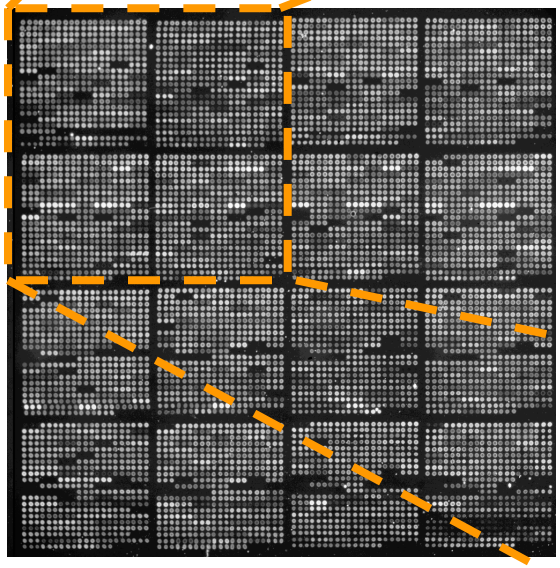
$$\delta_S = \sqrt{\left(\overline{S_h} - S_h\right)^2 + \left(\overline{S_v} - S_v\right)^2}$$

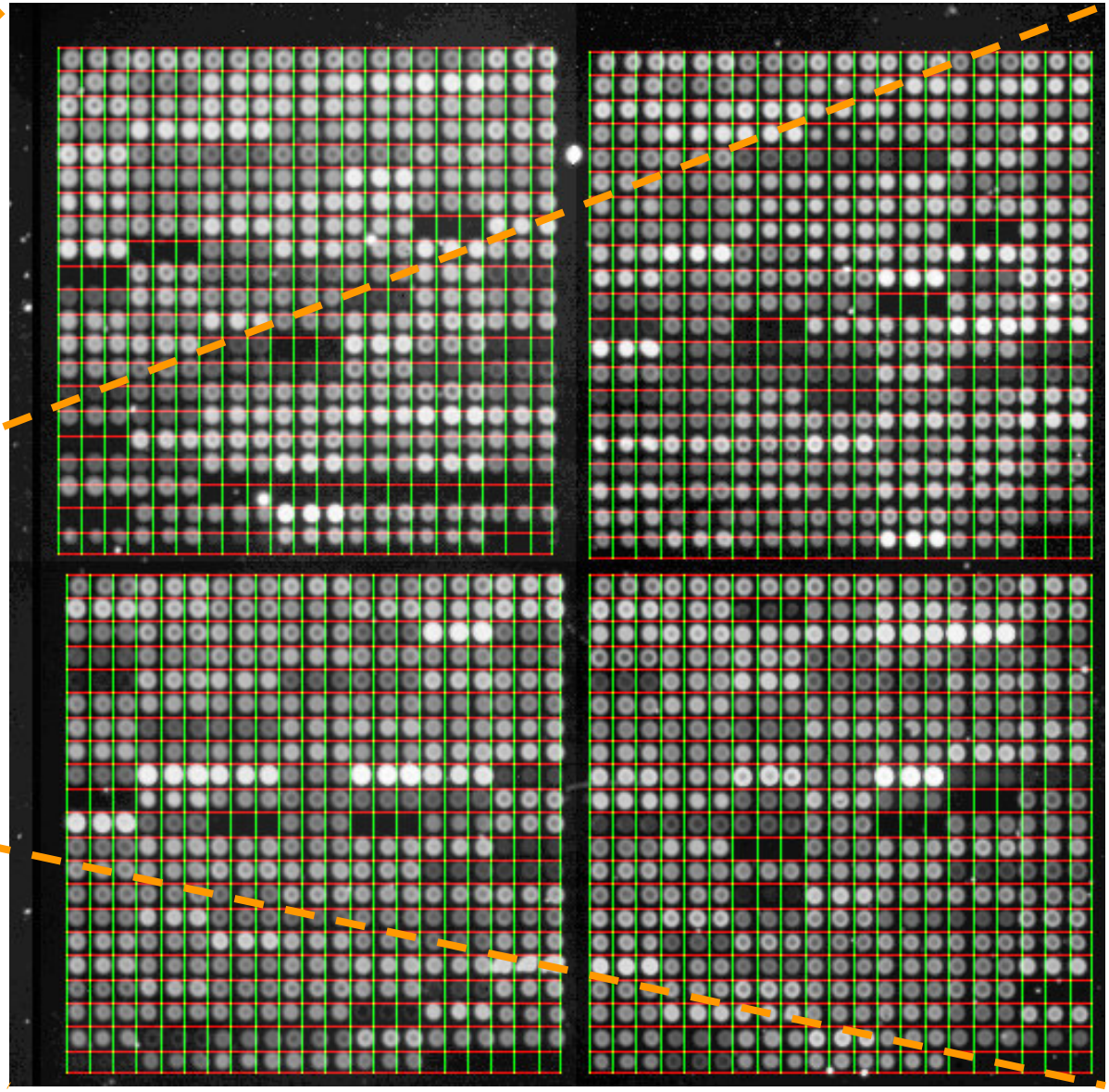
Further experiments

Example from *Spot* [1]

4x4 grids each with PR=21, PC=21, R=4

$$\delta_C, \delta_S \in [1, 2.25]$$



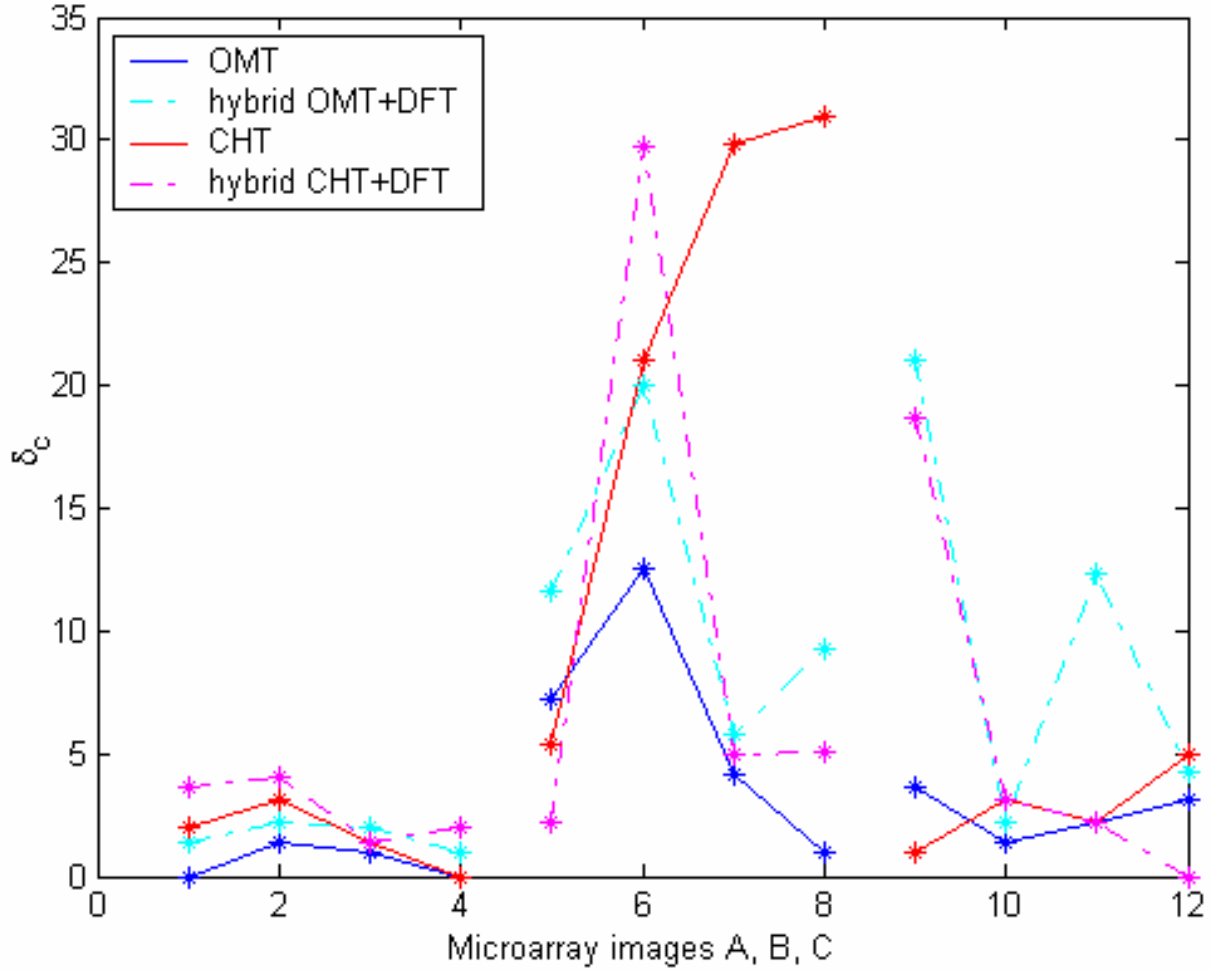


[1] http://experimental.act.cmis.csiro.au/spot/demodownload/demodownload/SpotExamples.zip

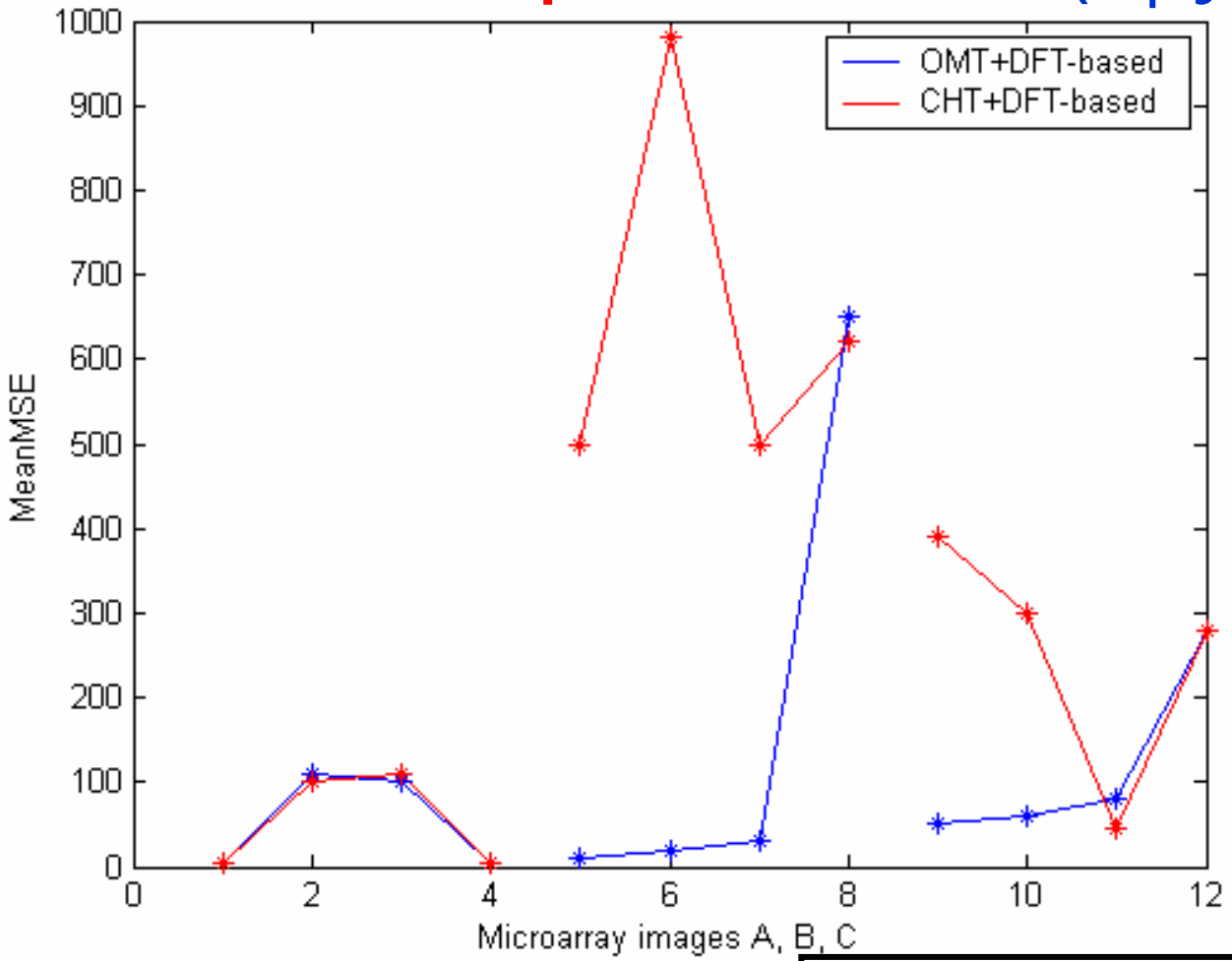
Hybrid methods

Combination of methods for period and first center computation

1. Compute S_v and S_h
2. Use S_v and S_h to identify the ROI where to look for circles
3. Compute (x_0,y_0) searching only into the ROI



Errors in computation of (x_i, y_i) $\forall i$



(x_i, y_i) = coordinates computed with ScanAlyze

$(\overline{x_i}, \overline{y_i})$ = coordinates computed with our methods

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\overline{x_i} - x_i\right)^2 + \left(\overline{y_i} - y_i\right)^2$$

Conclusions

- Brief overview of image processing issues for microarray images:
  - The data
  - Sources of variations in the data
  - Some methods usually adopted for the three steps
    1. Gridding
    2. Segmentation
    3. Intensity extraction

Conclusions (cont.)

- An approach to microarray image gridding:
  computation of (x_0,y_0), S_h, and S_v and different methods to accomplish it

  Pros:
  - incorporates knowledge about *ideal* grid (PR, PC, R), without requiring parameters that are data dependent ((x_0,y_0), S_h, S_v)
  - does not produce misalignment due to spurious or missing spots
  - Appropriate if:
    - measured grid geometry does not deviate too much from grid model defined by the template, OR
    - adopted for initial grid, to be later refined

  Cons:
  - if measured spots are unpredictably regular, leads to inaccurate results