Image analysis of cDNA microarray data
Marit Holden
Norsk Regnesentral

Marit.Holden@nr.no

Protein synthesis:
transcription and translation

Figure taken from The Human Genome Project at http://www.ornl.gov/hgmis/project/info.html
Gene expression

- "The process by which a gene's coded information is converted into the structures present and operating in the cell. Expressed genes include those that are transcribed into mRNA and then translated into protein and those that are transcribed into RNA but not translated into protein (e.g., transfer and ribosomal RNAs)."
- In a specific cell, at a certain time point, only a few genes in a cell are expressed
DNA microarray applications

• Human disease diagnostics and treatment
  – determination of predisposition and risk factors wrt. certain diseases
  – prediction of risk factors involved using certain treatment schemes
  – monitor disease stage and treatment progress

• Agricultural diagnostics and development
  – identify plant pathogens to allow suitable plant protection to be improved
  – efficiency and economy in plant biotechnology

• Analysis of food and genetically modified organisms (GMO)
  – determine the integrity of food
  – detect alterations and contaminations
  – quantify GMOs

• Drug discovery and drug development
cDNA microarray experiment

- cDNA clones (probes)
- PCR product amplification
- Purification
- Printing

mRNA target

- Hybridise target to microarray
- 0.1 nl/spot

Analysis

- Overlay images and normalise

Copied from talk by Terry Speed at http://www.ipam.ucla.edu/programs/fg2000/fgt_tspeed7.ppt

Norsk Regnesentral
Norwegian Computing Center

November, 2001
Marit Holden
The TIFF image

• Raw data: two 16-bit TIFF images
• Pixel: represents the total fluorescence in the region corresponding to the pixel = area density of dye molecules
• Settings on scanner (adjusted by the user)
  – Scan rate, laser power, PMT voltage
Noise

• “Perfect image”
  – Only reflects the measures of the fluorescence intensities for the dye of interest

• Noise affecting final signal
  – Photon noise and electronic noise
  – The amplification and digitization process
  – Laser light reflection
  – Dust at the slide
  – Treatment of the glass slide
  – Background fluorescence
Comparison of methods/systems

- Image analysis packages
  - Spot – commercial (image analysis package in paper)
  - ScanAlyze – publicly available
  - GenePix – commercial
  - QuantArray – commercial
- Compare methods
- Paper: Effect of image analysis decisions on $\log_2 R/G$
Image processing

• Gridding/addressing
  – Finding the location of each of the spots

• Segmentation
  – Classification of pixels as foreground (fg) or background (bg)

• Intensity extraction
  – Calculating for each spot
    • Rfg, Gfg – foreground intensities
    • Rbg, Gbg – background intensities
    • Quality measures
Combined image

- Used by the Spot image analysis system in the gridding and segmentation process
- Obtain an image 8-bit image from the 16-bit R and G images:
  - \( R' = f(R) \) and \( G' = f(G) \), where \( f \) is a square root transformation
  - \( m_R \) and \( m_G \) are median values
  - New pixel = \( \min(G' + (m_G / m_R) \times R', 255) \)
- A spot mask will be obtained from the 8-bit image
Gridding

Minimize user intervention - automatic

• Quantarray
  – A lot of manual work
• GenePix
  – Less manual work
• ScanAlyze
  – Paper gives no information about that
• Spot
  – Very little manual work

• A method based on Bayesian statistics
Spot’s gridding method

• Based on
  – Constant within a batch (the basic structure):
    • Number of rows and columns with grids
    • Number of rows and columns with spots
    • Print-tip configuration
  – Varies between slides:
    • Overall translation
    • Rotation and skewing (Here: assumed to be no problem)
• User chooses one image as a template
  – Specifies top-left spot in each grid, and bottom-right spot in bottom-right grid
  – Remaining part of gridding: automatic
Gridding - Bayesian image analysis

- Time-consuming for computer, but not for user
- Rotation and skewing allowed
- Applied on phosphor images
  - A robot places a rectangular lattice of spots on a hybridization filter
Gridding - Bayesian image analysis cont.

- A Markov Random Field is used as a structural model of a deformable rectangular lattice.
- Reconstructed lattice: A maximum a posteriori (MAP) estimate is found by simulated annealing.
- Without data (prior): Maximum probability to the perfect regular lattice.
- Simulate realizations from the posterior (with data).
- Some preprocessing for simplifying the model.
Segmentation

• Partition the image into foreground (spot) and background
  – Fixed circle segmentation
    • ScanAlyse, GenePix, Quantarray
  – Adaptive circle segmentation
    • GenePix, Quantarray
  – Adaptive shape segmentation
    • Spot
  – Histogram segmentation
    • Quantarray
Circle segmentation

- **Fixed circle segmentation**
  - Circle with constant diameter for all spots
  - Many packages have this as an option
  - Problematic when not all spots are circular and of the same size

- **Adaptive circle segmentation**
  - Diameter estimated for each spot
  - Many packages: manual adaptive circle segmentation
  - Problematic when not all spots are circular
Adaptive shape segmentation

• Allows spots of different size and shape
• Seeded region growing (proposed in the paper)
  – Require specification of seed points
  – Seeds for foreground and background regions - easy for microarrays
  – Algorithm:
    As long as there are pixels not belonging to one of the regions:
    • Check all pixels which are neighbors to at least one of the regions and find the one with value closest to the mean value of its neighbor region.
    • Extend the neighbor region with this pixel.
Histogram segmentation

• Uses a circular or square target mask larger than any spot
• Fg and bg intensity estimates obtained from the histogram
  – Threshold determines fg or bg or
  – Fg = mean of values between 5\textsuperscript{th} and 20\textsuperscript{th} percentile
  – Bg = mean of values between 80\textsuperscript{th} and 95\textsuperscript{th} percentile
• Advantage: simplicity
• Disadvantages: pixels from neighboring spots + spot masks not connected
Intensity calculation

• Foreground intensities: $R_{fg}$, $G_{fg}$
  – Mean or median of spot mask

• Background intensities: $R_{bg}$, $G_{bg}$
  – Median of selected regions surrounding the spot mask
  – For later background corrections
    • Removing contribution which does not come from the hybridization of the target to the spot
    • Usually performed by subtracting $bg$ from $fg$

• Quality measures
  – Spot size or shape
  – Measures of $bg$ intensities relative to $fg$ intensities
Background correction

- No background correction
- Background corrections
  - Local background correction
    - Background varies gradually and smoothly over the slide
    - Computes the median pixel value in a region near each spot
  - Morphological opening correction
  - Constant background correction
    - Common background for all spots
Computing background intensities

Which pixels belong to the background?

- **ScanAlyze**
  - All pixels within the square centered at the spot center, not within the spot

- **Quantarray**
  - Considers the area between two concentric circles

- **Spot and GenePix**
  - Valleys between spots of the array (average of four median values)
Spot - Computing background intensities

• Morphological opening (non-linear filter) on original data:
  – Square structuring element
  – Removes all spots and obtain an estimate of the background
  – Avoid bright pixels
  – Estimate local backgrounds from a large local window

• Constant background
  – Computes background from set of control spots (no hybridization)
  – Paper: No control spots
    • 3rd percentile of the foreground pixels
Experiments

- 16 slides:
  - 8 with normal mice and 8 with not normal mice
  - Common reference for all 16 slides
- Computed $\log_2(R/G)$
- Used a within-slide spatial and intensity dependent normalization method
- Find out which genes are differently expressed for the normal compared to the not normal mice
- Compared the different methods for this experiment (+ a follow-up dye-swap experiment with the most interesting genes printed several times on two slides)
Conclusion

• Background ajustment method: important
• Segmentation approach: smaller impact on result
• Study indicates: morphological background estimation better than other methods
  – Does not include foreground pixels
  – Uses many pixels in the computation: smaller variability
  – Lower background estimate more correct?
• Results from the experiments when separating eight genes known to be differentially expressed
  – Spot best, closely followed by ScanAlyze, then GenePix
  – QuantArray generally poor