Estimating Probe Cell Locations

Harry Zuzan
Institute of Statistics and Decision Sciences
Duke University
Why analyse pixels?

Nature Genetics Supplement on Microarrays (1999)

- Miniaturisation provides a massively parallel assay that removes temporal sources of variation
- The small quantities of RNA required remove sources of variation due to treatment
- Hybridisation takes place on a rigid non-permeable substrate reducing variable conditions that affect hybridisation and data acquisition

All these are low-level gains that facilitate the collection of highly multivariate and highly reproducible observations, each containing stores of biological information. Further improvements in any of these low-level aspects of obtaining gene expression data benefit analyses that take place downstream. Analyses of hybridisation images at the pixel level can make improvements in quantifying image features, which in turn improves data quality.
For each probe cell, Affymetrix reports the 75th percentile of pixels, the standard deviation of pixels and the number of pixels attributed to the probe cell. In the older Hu6800 and Mu11K chips, 36 pixels were usually attributed to each probe cell. In the newer U95 chips, usually 20 or 25 pixels are attributed to each probe cell. The raw hybridisation image is 4733×4733 pixels and these small “plugs” of pixels should be located interior to probe cells in order to accurately represent the scanned fluorescent intensity.

The plot in slide 3 was generated from the cell file of a Mu11KA hybridisation with a tuned up photo multiplier tube (PMT). There are 534×534 probe cells on the Mu11KA chip. In the fan-shaped plot, the effect of saturation can be seen on the right side and on the left side a scattering of probe cells with a low 75th percentile but large standard deviation indicate unreliable probe cell intensities. The points are densely packed at the base of the funnel.
The coefficient of variation for each probe cell can be obtained from the same information used to generate the plot in slide 3. In slide 4 the CVs from slide 3 are plotted as a grayscale image. Pixel intensities increase with the CVs of their corresponding probe cells. The pattern of bands indicates that there is a relationship between a probe cell’s CV and its location in the grid of probe cells.

Misalignment of the probe cell grid that Affymetrix superimposes over the hybridisation image was suspected to be the source of this pattern and examination of the grid revealed that pixel intensities were not always correctly attributed to probe cells. Removal of this spatial contribution to variation should produce improved estimates of probe cell intensities, motivating the development of an alignment algorithm.
The alignment algorithm

- Estimates the locations of probe cell centres
- Scanned probe cell locations are modelled as a continuous deformation of a lattice
- The algorithm uses penalties to balance:
  1. Minimising the variance of pixel intensity near probe cell centres
  2. Maintaining local lattice structure of probe cell locations

Each pixel in the scanned image of an hybridisation represents a small region on the physical hybridisation surface that could be interior to a probe cell, straddle as many as four probe cells or be partly or entirely in the border area surrounding the array of probe cells. Evident in hybridisation images is the effect of a blurring process. Due to the discrete approximation of the hybridisation surface provided by pixels and the effect of the blurring process, the image of an hybridisation is not segmented by probe cells, even though the physical hybridisation surface is segmented. The lack of image segmentation has the greatest effect on intensities of pixels representing regions on or near the perimeter of probe cells, in the sense that these pixel intensities do not represent signal accumulated from a single probe cell. As a consequence, pixels near the edges of probe cells may need to be discarded when computing hybridisation summaries. Accurate estimates of where probe cells are located are required for the remaining pixel intensities to be representative of probe cell response to hybridisation.

Probe cells are laid out in a rectangular array but past experience has provided evidence that probe cells are not equally spaced. The deviation from equal spacing is gradual and can be modelled as a continuous deformation of a perfect lattice.
Probe cells boundaries are evident by abrupt changes in pixel-to-pixel intensities near neighbouring probe cells. A small region of pixels located on a probe cell will tend to have smallest variance of pixel intensity if it is aligned to the central region of the probe cell. Misalignment will in most cases cause an increase in variance. The alignment algorithm takes into account the tendency of misalignment to increase this variance. The alignment algorithm also takes into consideration the local lattice structure of the gradually deformed hybridisation image.
Let $j$ be a variable that indexes the array of probe cells and let $c_j$ be the current estimate of the coordinates of the centre of probe cell $j$. Although pixels occupy discrete locations, consider the elements of $c_j$ to be continuous. Let $\mathcal{N}_j$ be the set of indices of the eight neighbours of probe cell $j$ and let $\bar{c}_j$ be the centroid of $\{c_k\}_{k \in \mathcal{N}_j}$. Based on the criterion of retaining local lattice structure, the optimal revised estimate of $c_j$ is $\bar{c}_j$. 

Maintaining local lattice structure

- Let $j$ index probe cells
- Let $c_j$ be the current estimate of the location of the centre of probe cell $j$
- Let $\mathcal{N}_j$ be the set of indices of the eight neighbours of probe cell $j$
- Let $\bar{c}_j$ be the centroid of $\{c_k\}_{k \in \mathcal{N}_j}$
Use $\bar{c}_j$ as the centre of a $3 \times 3$ rectangular grid of evenly spaced locations to sample in order to propose revised estimates of $c_j$. Call these nine locations $a_{uv}$, with $u, v \in \{-1, 0, 1\}$, $a_{00} = \bar{c}_j$ and let $\delta \in (0, 1]$ be the distance separating adjacent locations.
Penalty scores

Revise $c_j$ to minimise penalties

- The penalty for deviation from $\bar{c}_j$ is $\delta$
- The penalty for fit is $s^2$, the computed variance of pixel intensity
- Overall penalty is a weighted average $a\delta + bs^2$

At each $a_{uv}$, the penalty for locating the probe cell out of alignment with its neighbours is $t_{uv} = \sqrt{u^2 + v^2}$ and the variance, $s_{uv}^2$, of pixel intensities is computed within a small square region of fixed size centred at each $a_{uv}$. The boundary of this region can partially include pixels it intersects and the contribution of pixel volumes to variance is weighted according to their partial areas within the boundary.

The decision of which $a_{uv}$ to choose as the revised estimate of $c_j$ is based on minimising a weighted average of the penalties $t_{uv}$ and $s_{uv}^2$. After evaluating alignments over trial runs, we found it best to first log transform the pixel data and an effective weighted penalty to be $\delta t_{uv} + 5s_{uv}^2/S^2$, where $S^2$ is the mean of $s_{00}^2$ over all probe cells.
1. Obtain initial estimates for all $c_j$
2. Initialise $\delta$ ($\delta = 0.5$; works well)
3. While $\delta > 0$
   (a) Sequentially revise all $c_j$
   (b) Decrement $\delta$ ($\delta = \delta - 0.05$; gives 10 iterations)

Prior to the first iteration of the alignment algorithm, initial probe cell locations were estimated by interpolation between the positions of the probe cells at the four corners of the probe cell array, $S^2$ was computed and $\delta$ was set to 0.5. Call this initialisation the completion of iteration 0. For each subsequent iteration, the probe cell locations, $c_j$, $j = 1, \ldots, 534^2$ were updated sequentially. In each case, the revised $c_j$ immediately replaced the estimate from the previous iteration as did the contribution of $s_{00}^2$ to $S^2$. Thus, each computation of location $\bar{c}_j$ was based on the most recently revised members in $\{c_k\}_{k \in N_j}$. After each iteration, $\delta$ was decremented by 0.05 and the iterations ceased when $\delta$ was no longer greater than 0.
The same scanned hybridisation used to generate the plots in slides 3 and 4 was reanalysed at the pixel level. First the probe cells locations were initialised by linear interpolation between the 4 corners of the array of probe cells in the raw image scan. The mean and variance was computed from 6×6 regions centred on the interpolated locations. A plot of standard deviation versus mean intensity of these regions is shown in slide 11. This plot can be compared to the plot in slide 3. One major difference is the way that saturation is approached along the right hand side. This is partly due to fact that the mean was used instead of the 75th percentile.
The initial interpolated locations were revised using the alignment algorithm. Slide 12 shows the plot of standard deviations versus mean intensities within probe cells after alignment. Compared to the plot in slide 11 the standard deviation has substantially decreased. But this is what the alignment algorithm set out to do. It will almost certainly find probe cell locations with decreased overall variance and more evidence is needed to corroborate that decreased variance resulted from improved alignment.
A plot of probe cell means computed from initial interpolated locations versus probe cell means computed from aligned locations is shown in slide 13. It appears that prior to alignment incorrect attribution of pixel intensities to probe cells had caused low intensity probe cells to pick up signal from their high intensity neighbours. After alignment it appears that high intensity probe cells had recovered their signal.

The comparisons thus far compare means and variances at aligned locations to means and variances at interpolated locations. This is not yet a direct comparison to Affymetrix’s probe cell summaries. Slides 11, 12 and 13 indicate that linear interpolation is inadequate and the alignment algorithm seems to be operating as intended. Direct comparisons with Affymetrix’s summaries follow.
The data in slide 12 was used to produce the image in slide 14. It can be seen that after the alignment algorithm was employed the banding pattern disappeared. The narrow band of larger than typical coefficients of variation at the bottom of the image correspond to a region where little or no hybridisation takes place on the Mu11KA chip causing small denominators in the CVs. The reduction in variance of intensities within probe cells can be attributed to the removal of a spatial source of variation caused by misalignment.
In slide 15 is a plot Affymetrix’s 75th percentiles versus post-alignment means. Although it is tempting to conclude that a linear relationship exists, the fact that almost all of these probe cells contain different probe sequences must be considered. Each probe cell binds target RNA with an affinity that is unique to its probe sequence and the apparent strength of the linear relationship may be due to the broad range of probe-target affinities present on an HSDM. The disparity encountered by choosing the mean in place of the 75th percentile combined with improvements in alignment may prove to be large in relation to a given probe cell and its probe sequence. When deciding how individual probe cell intensities should be represented, accuracy is at stake. This is in contrast to comparisons of entire hybridisations, where reproducibility is at stake. The need for accuracy suggests that a measure of central tendency such as the mean or median should be used to represent probe cell intensity.
The banding pattern observed in the previous data does not appear with nearly the strength in more recent data. Possible sources for this correction are changes in Affymetrix’s software, use of different chips and scanner to scanner variation. Maybe others.

The alignment algorithm does not depend on any special probe cells and can be applied to all the current chip designs. It was applied to a U95A hybridisation image that was scanned using a tuned down PMT. In slide 16 is a plot of the 75th percentiles reported by Affymetrix versus probe cell means using the alignment algorithm. Because of the smaller size of probe cells on U95 chips a 4×4 pixel area was used to compute probe cell means. The overall intensity of the hybridisation was low but not unusual given the tuned down PMT. The maximum aligned probe cell mean was 26332. Since most of the data lies near the base of the scatterplot it is useful to trim the range of points plotted.
Slide 17 shows the plot in slide 16 cut off at the 99th percentile for mean aligned probe cell intensity. The 99th percentile is 2029.
Slide 18 shows the plot in slide 16 cut off at the 97th percentile for mean aligned probe cell intensity. The 97th percentile is 1031. Looking at narrower ranges of the data brings out the difference between the use Affymetrix’s 75th percentile and the mean probe cell intensity after alignment. The following table lists percentiles of the aligned probe cell means from the hybridisation.

<table>
<thead>
<tr>
<th>percentiles</th>
<th>intensities</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>53.93</td>
</tr>
<tr>
<td>20</td>
<td>59.81</td>
</tr>
<tr>
<td>30</td>
<td>67.05</td>
</tr>
<tr>
<td>40</td>
<td>77.35</td>
</tr>
<tr>
<td>50</td>
<td>93.12</td>
</tr>
<tr>
<td>60</td>
<td>119.37</td>
</tr>
<tr>
<td>70</td>
<td>165.56</td>
</tr>
<tr>
<td>80</td>
<td>245.61</td>
</tr>
<tr>
<td>90</td>
<td>443.77</td>
</tr>
<tr>
<td>95</td>
<td>732.88</td>
</tr>
<tr>
<td>96</td>
<td>852.09</td>
</tr>
<tr>
<td>97</td>
<td>1031.74</td>
</tr>
<tr>
<td>98</td>
<td>1335.90</td>
</tr>
<tr>
<td>99</td>
<td>2029.49</td>
</tr>
<tr>
<td>100</td>
<td>26332.28</td>
</tr>
</tbody>
</table>
Improvements

- A source of variation was identified and removed
- All the details of the alignment algorithm can be specified.
- Provides within probe cell outlier detection, alternatives to using the 75th percentile.
- Possibly probe cells could be made smaller

In concurrent work, knowing probe cell locations has permitted the implementation of an image model for the entire hybridisation image. The image model subtracts background noise and permits perfect match intensities to be measured without reference to mismatches.
Many thanks to the people at Duke who have helped out and who I have had the chance to work with in the collaboration between ISDS and the Department of Genetics.