Automatic registration of microarray images. II. Hexagonal grid

Vitaly L. Galinsky

Illumina, Inc., 9885 Towne Centre Dr., San Diego, CA 92121, USA and ECE Department, University of California, San Diego, 9500 Gilman Dr., MC 407, La Jolla, CA 92093-0407, USA

Received on January 8, 2003, revised on April 24, 2003; accepted on April 25, 2003

ABSTRACT
Motivation: In the first part of this paper the author presented an efficient, robust and completely automated algorithm for spot and block indexing in microarray images with rectangular grids. Although the rectangular grid is currently the most common type of grouping the probes on microarray slides, there is another microarray technology based on bundles of optical fibers where the probes are packed in hexagonal grids. The hexagonal grid provides both advantages and drawbacks over the standard rectangular packing and of course requires adaptation and/or modification of the algorithm of spot indexing presented in the first part of the paper.

Results: In the second part of the paper the author presents a version of the spot indexing algorithm adapted for microarray images with spots packed in hexagonal structures. The algorithm is completely automated, works with hexagonal grids of different types and with different parameters of grid spacing and rotation as well as spot sizes. It can successfully trace the local and global distortions of the grid, including non-orthogonal transformations. Similar to the algorithm from part I, it scales linearly with the grid size, the time complexity is \( O(M) \), where \( M \) is total number of grid points in hexagonal grid. The algorithm has been tested both on CCD and scanned images with spot expression rates as low as 2%. The processing time of an image with about 50,000 hex grid points was less than a second. For images with high expression rates (\(~90\%) the registration time is even smaller, around a quarter of a second.

Contact: vit@ucsd.edu
Supplementary information: http://fleece.ucsd.edu/~vit/Registration_Supplement.pdf

INTRODUCTION
In spite of all of the success and proliferation of different types of high-density genomics, gene expression and DNA arrays (Bowtell, 1999; Lockhart and Winzeler, 2000) no single dominating array technology has emerged yet. Most of the existing arrays are based either on a robotic dispensing of genetic material or on a photolithographic imprinting on glass or silicon substrate (Lipshutz et al., 1999). Photolithography provides more dense packing of probes and allows the creation of arrays (Zarrinkar et al., 2001).

Although these microarray methods create rectangular grids of DNA probes on a DNA biochip or slide, there is an alternative high-density high-sensitivity array technology, that provides different packing of probes. This technology is based on microspheres (or beads) encoded with DNA probes and placed on optical fiber substrate (Ferguson et al., 2000; Epstein et al., 2002; Oliphant et al., 2002). The fiber substrate is created by a two stage process, [Walt, 2002; Illumina, Inc. (2002), BeadArray™, http://www.illumina.com/tech_plat.htm]. Initially several single core fibers are bundled together forming a fiber minibundle. Then several of these minibundles are melted together forming a rigid multi-fiber microarray. This assembly process provides a high-density packing of individual fibers and creates a grid having hexagonal structure.

This hexagonal grid packing has some advantages over rectangular grids. It is more rigid and allows more features to be packed in the same area. At the same time the processing of fiber-optic microarray images requires more sophisticated algorithms for spot indexing than in the case of rectangular grids.

The objective of the second part of the paper is to present a modified version of the spot indexing algorithm from the first part suitable for indexing spots organized in hexagonal grids from the imaging of the fiber-optic microarray bundles. The algorithm is completely automatic and requires only minimal information to proceed. This information includes the geometrical structure of the hexagonal grid, such as number of single core fiber layers in each minibundle and number of layers of minibundles in the microarray bundle. The algorithm does not have any restrictions on the bundle rotation and is stable to image distortions, including non-orthogonal transformations and non-linear elastic deformations. It can reliably process images with low spot expression rates (as low as \(~2\%)). The tests showed less than a second processing time for a single bundle with about 50,000 grid points on a 1.6 GHz Athlon
It should be emphasized that the spot indexing algorithm for the hexagonal grid is inherently much more complex than the algorithm for the rectangular grid. The most important differences and features unique for the hexagonal grid can be summarized as follows:

- The spot indexing for the hexagonal grid is a two level procedure, a hexagonal subgrid is constructed for each minibundle and all the minibundles are combined together to form the final grid.
- For template or high spot expression rate images the algorithm uses an information about overlapping minibundles in order to correct errors due to distortions or defects.
- For low spot expression rate images the use of template grids is beneficial.
- The iterative procedure of grid centering for the hexagonal grid is completely different and much more complicated than in the case of the rectangular grid. It utilizes only the information from the boundary grid points of the outward layer of minibundles.

To avoid repetitions the second part of the paper is focused on these major differences.

### Spots finding

The first step of the algorithm uses the same adaptive spot detection techniques that were used for rectangular grid. The calculations of average grid parameters (spacing $d_3$ and rotation angle $\theta_3$) are also similar and based on the detect and spread algorithm. But because the hexagonal grid has three axes of symmetry instead of two for rectangular grid, only one spreading pattern is used and the rotation angle is normalized to two different ranges—one is $[-\pi/6 : \pi/6]$ and the second is $[0 : \pi/3]$. After that the rotation angle should be chosen from the range with lowest variance. As a result all restrictions on the value of bundle rotation can be avoided.

Because of the more dense packing of spots the method is even more robust for a hexagonal grid than in the case of a rectangular grid. Hence, for hexagonal grids with comparable total number of points it requires lower spot expression rates to be successful. The probability of failure $p_{sf}$ for a hexagonal grid can be estimated from similar considerations as

$$p_{sf} \simeq ((1 - R_s)^6)^N, \quad \text{or}$$

$$p_{sf} \simeq [1 - R_s[1 - (1 - R_s)^6]]^M,$$

where the boundaries are neglected and six neighbors are counted for each grid point ($R_s$ is the probability that there is a spot at one grid point and $(1 - R_s)^6$—the probability that there are no spots at the positions of all of its close neighbors). The conditional probabilities are also neglected here. For large grids ($N_f = 5, N_m = 13$, hence, $M = 49777$) and for as low expression rates as $R_s = 1\%$ both expressions estimate the probability $p_{sf}$ to be less than $10^{-12}$.

The set of detected spots is used to create incomplete Voronoi diagram or $3 \times 3$ proximity cells for each spot. There is no difference in this step from the rectangular grid.
Grid placement

The spot indexing or grid placement is a more complicated procedure for a hexagonal grid than for a rectangular one. It consists of two levels of indexing, one is indexing of grid points inside each minibundle and the second is indexing of minibundles in the entire fiber bundle.

The indexing inside the first minibundle starts with positioning of the smallest hexagon of the grid (seven grid points) using average grid parameters $d_g$ and $\alpha_g$ and adjusting the position of the hexagon for best fit with spots found. As an approximate center position the center of mass of all spots detected in the image can be used. Then the indexing proceeds as shown in Figure 1 such that each new grid point is interpolated from the locations of two or more of its already indexed neighbors.

After the indexing of all the spots for the particular minibundle has been finished the center of the next minibundle is interpolated from several grid points located at the edges of either two or three of the already indexed neighboring minibundles, as shown in Figure 2. Only the location of the second minibundle center is determined from the single neighbor—the points at the edge of the very first minibundle.

The interpolation of the minibundle center from several neighboring minibundles positioned during different, and not necessarily adjacent, stages of placement allows to avoid accumulation of interpolation errors from early stages to the end. But due to possible presence of local and global grid deformations it does not guarantee that the center of each minibundle will fall in the proximity cell of the correct spot. Hence, it is possible that at the end of indexing some minibundles will be shifted in one of the six possible directions of the grid and will be assigned to the wrong spot. As a result the minibundle will either overlap with some other minibundles or will have one or several edges not associated with any spots in the image (e.g. Supplementary information, Figure 3sup). This error is more likely to happen at the outermost layer of fiber bundle where the distortions of the grid are the strongest.

To correct for this type of minibundle misplacement different strategies can be applied for low and high values of spot expression rates. For grid template images or images with high spot expression rates ($\geq 90\%$) it is easy for each minibundle to record the number of grid points at each of the six edges that either overlap with neighboring minibundle (that is they fell in the proximity cells of spots already assigned to different minibundle) or empty (do not have any spots close enough to them). If two adjacent edges have the number of overlaps or empties larger than the other four edges of the minibundle then the minibundle is forced to jump in the direction uniquely determined by the orientation of overlapping or empty edges. After several iterations through all badly placed minibundles all the defects in placement due to deformations of the grid will be corrected. The time complexity of the entire grid placement for a high spot expression rate image will be determined by $O(M)$ term from the first stage of the placement and the correction part will usually require much less than $O(M)$ operations.

For images with a small number of expressed spots the above correction procedure is not appropriate. Therefore the least square fit approach based on the affine transformation and grid centering (similar to the method described in the first part of the paper for rectangular grid) should be used.

Affine transformation and grid centering

After the initial crude placement of the grid where the interpolation has been done using averaged-over-all-spots grid spacing $d_g$ and rotation angle $\alpha_g$ the parameters of the affine transformation from either an ideal hexagonal grid or from the template grid for this bundle can be calculated. The affine transformation should be used in a similar way as in the first part of the paper

$$x'_i = a_{11}(x_i + \bar{x}) + a_{12}(y_i + \bar{y}) - \bar{x}'$$
$$y'_i = a_{21}(x_i + \bar{x}) + a_{22}(y_i + \bar{y}) - \bar{y}'$$

where $x'_i$ and $y'_i$ ($i = 1, \ldots, M$) are the coordinates of constructed grid in the center of mass $(\bar{x}', \bar{y}')$ coordinate system.
$x_i$ and $y_i$ are the coordinates of ideal hexagonal or template grid also in the center of mass $(\bar{x}, \bar{y})$ coordinate system. The parameters of transformation $a_{11}, a_{12}, a_{21}$ and $a_{22}$ can be obtained using the same expressions as in the first part.

There is no guarantee that the center of the bundle will be found correctly after all these placement steps. A procedure for finding the correct center of the hexagonal grid differs more significantly from the centering of the rectangular grid and is not an easy task especially for low spot expression rate images. It can be done iteratively, but instead of performing an unguided search through all the neighbors of the incorrect center, the direction of grid movement on each iteration can be estimated from the numbers of spots assigned (or unassigned) to each point of the grid in the outer edges of the minibundles located in the outermost layer of the bundle.

The number of empty grid points at the edges of the bundle can be assigned to the vector $S_e$ as follows:

$$S_e^i = \sum_j n_{ij},$$

where $i$ is the index of one of the six edges of the bundle (Fig. 3), $j$ is the grid point index in this edge and

$$n_{ij} = \begin{cases} 0 & \text{if there was spot assigned,} \\ 1 & \text{if there was no spot assigned.} \end{cases}$$

Finding from $\{S_e^i\}$ the largest value of combined empty grid points at any of two adjacent edges allows the unique identification of the required direction of grid shift for the next iteration. This results in significant speed up of the search.

The probability of failure $p_{gf}$ of this guided search can be estimated in a similar way as for the rectangular grid. If all expressed spots are uniformly distributed on the grid this probability can be approximated as

$$p_{gf} = 1 - \left[ 1 - (1 - R_s)^K \right]^6,$$

where $R_s = N/M$ is the spot expression rate and $K$ is the number of grid points at one of the six edges of the bundle. This estimate confirms that hexagonal grids are more error prone than rectangular grids with comparable total number of points. For example for fiber bundle composed of 13 layers of minibundles each having five layers of fiber cores or 49 777 total grid points and for $R_s = 1\%$ the probability of failure is about 34\% (the number of grid points at the edge is 270 in this case). At the same time the similar estimate for the rectangular grid with 50 625 total grid point from the first part was 36\%.

The objective functions for the search are the same as in the first part—the number of spots assigned and the largest aggregate intensity of spots from the same sequences—and they are also used to control the quality of grid placement.

**RESULTS**

The algorithm has been tested on several thousands microarray fiber bundles with different resolutions obtained by different image acquisition hardware including scanners and CCD cameras. The processing time depends on image resolution and on the number of spots but a typical image with 1000–50 000 spots takes from a quarter of a second for template images (large number of spots expressed) to a little more than a second for low spot expression rate images on AMD Athlon 1.6 GHz.

Several real data examples of registration of fiber bundle images are included in Supplementary information.

The first example of grid placement for the template image is shown in Figure 3sup. After the initial stage of placement the center of one of the minibundles was found incorrectly (as shown in enlargement), but this error was easily corrected by the procedure described above.

The next example (Supplementary information, Figure 4sup) is the low spot expression rate fiber bundle image. The total number of spots detected was 756, that corresponds to 1.5\% spot expression rate. The image was correctly registered.

Last example (Supplementary information, Figure 5sup) shows spots indexing in the image with strong local distortions. The algorithm has been able to trace the distortion and create the correct grid for this fiber bundle.

**CONCLUSIONS**

In this paper the author presented an effective and a completely automated algorithm for registration or indexing of
hexagonal fiber-optic bundle arrays. The time complexity of the algorithm $O(M)$, where $M$ is the number of grid points. It uses minimal amount of input data, namely the number of layers of minibundles and the number of layers of fiber cores in each minibundle. The algorithm successfully overcomes the orthogonal and non-orthogonal transformations and even local and non-linear distortions.

ACKNOWLEDGEMENTS
The author thanks Dr Francisco Garcia and Dr Jurgen Haas for sharing insights for the algorithm and all the people at Illumina, Inc. The comments and suggestions of the editor and anonymous reviewers are also very much appreciated.

REFERENCES


