Original Paper

High-throughput Automated Image Processing System for Cell Array Observations

Daisuke Tominaga,[†] Fukumi Iguchi,[††] Katsuhisa Horimoto[†] and Yutaka Akiyama[†,†††]

In many cases of biological observations such as cell array, DNA micro-array or tissue microscopy, primary data are obtained as photographs. Specialized processing methods are needed for each kind of photographs because they have very wide variety, and often needed automated systems for modern high-throughput observations. We developed a fully-automated image processing system for cell array, high-throughput time series observation system for living cells, to evaluate gene expression levels and phenotype changes in time of each cell.

1. Introduction

In many cases of biological observations such as 2-dimensional electrophoresis of proteins, DNA micro-array, and cell array, the primary data are obtained as photographic images. Experimental methods developed in recent years are characterized by high-throughput, then large volumes of photographic data are now obtained with the aim of comprehensive observation. In cell array [1)~5)] (**Fig. 1**), which is currently the focus of our research, there are small pits (called "spots") on the surface of a slide glass (TFA chip). The cells are cultured in these holes while gene expression levels and phenotype changes are simultaneously observed using fluorescent proteins and other methods.

For example, in experiments using a TFA chip with 1,500 spots, if a photograph is taken of each spot every 15 minutes for 2 days, a total of 288,000 images ($1500 \times (24 \times 60/15) \times 2 = 288{,}000$) will be produced. For this approximate 300,000 images, background luminance intensity subtraction, noise on a spot elimination, recognition of cell shape, and the integration of luminance intensity are performed, and as time series processing, depending on the circumstances, each cell needs to be traced and any changes in gene expression levels and cell shapes in time are digitized.

In general, since this type of image processing requires experience at discriminating between noise and artifacts which are not noise, often manipulations by the researcher who did the experiments are needed. For example, based on knowledge concerning general cell size, shape, the approximate intensity of background luminance, and so on, the selection of optimum parameters for the experimental conditions is performed by trial and error iterations using software that comes with the observation equipment, etc. and semiautomatic processing is performed while monitoring for the appearance of miss-recognition and outliers by operators. In such cases, researchers who conducted the experiments are occupied for long periods of time because the most detailed and precise knowledge are needed to select optimum parameter values which reduce noise and extract the highest quality informations from photographed images.

Furthermore, manual processing including search optimum parameter values for the images on a video display monitor generally takes long time for the operation and may also damage the health of the operator. In our past cases, an operator could process approximately 20,000 images of cell array per month. Therefore, it will take about 1 year to process images obtained by 2 days observation.

Thus, support from bioinformatics is indispensable in order to achieve the maximum level of high-throughput, comprehensive observation ability of cell array. To solve problems above, many software systems have been developed. One of the most popular image processing/recognition software is ImageJ [6),7)] (formerly known as NIH Image). It is made as general-purpose and many algorithms are im-

† Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology
†† Information and Mathematical Science Laboratory, Inc.
††† Graduate School of Information Science and Engineering, Tokyo Institute of Technology

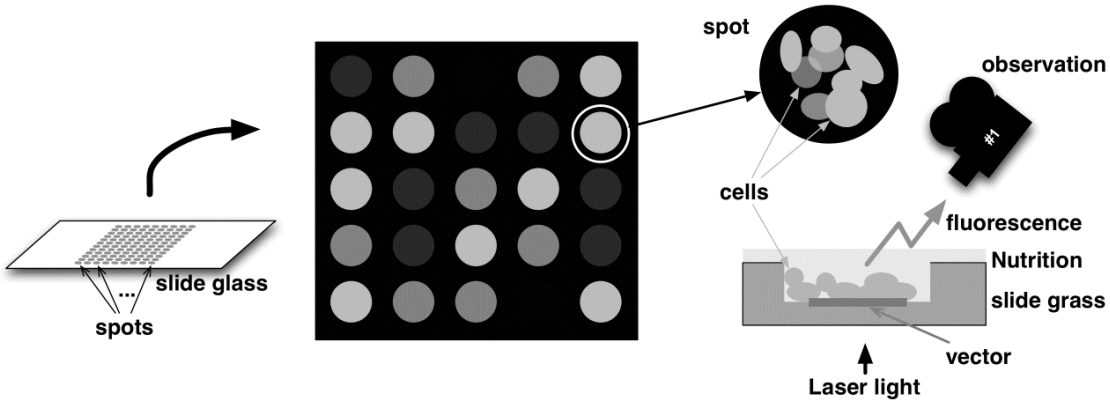

Fig. 1 Time series observation of living cells by the cell array. Several micrometers diameter holes (called 'spot') are digged on a slide glass. Other surface than spots is covered by polymer to prevent adhesion of cells. Virus vectors can be transfected through cell membrane from spot's surface into cells, then into nuclei. Transfected vector will make cells luminary if the vector contains a fluorescent protein gene. If a siRNA is transfected, target gene of the siRNA will be knocked out.

plemented based on it. Software especially for cell array are available in this few years [8),9)]. Although these software can digitize intensity of cell luminance, recognize each cell and its shape, count a number of cells, etc. [10)], nothing available to track each recognized cell in time. Since behavior of cells varies widely even in a same condition, cell tracking is significant to analyze how and why their behavior varies. Especially for large-scale high-throughput data, automated cell tracking must be introduced.

We therefore have developed a system using parallel computing and a minimum of manual operations to process the data. The software used was based on a published image processing system, ImageJ. The software runs on Linux and Windows XP.

2. Input-Output

2.1 Primary Data

Input data to our system are series of monochrome images in time obtained by a cell array equipment with one TFA chip. The input units are multiple images (a number of images equals the number of spots on the chip multiplied by the number of times each spot is photographed) obtained from a single experiment. Each image is a TIFF or JPEG microscopic image of a spot with a diameter of several μm (an example is shown as **Fig. 2**), and the file name included the spot number and number of a time the spot had been photographed. Information such as the magnification and exposure time that is not directly related to the image processing results is stored in separate files.

2.1.1 Optional Parameters

The numerical values of the upper and lower limits of the cell diameters are inputted into the software because of cell size, amount of movement, and so on are depending on cell species and experimental conditions being observed. The directory of the input file and the directory of the output on the computer in which the software runs are also specified by the user.

2.2 Output Data

From one image, the total of the luminance values in the image (total for a spot), the mean luminance intensity (mean of a spot is obtained by dividing the luminance by the total cell area), the total and mean of the luminance of each cell, and centroid and circularity of each cell are calculated, and then it is determined whether the cell in the image at a certain time point corresponding to a cell in an image taken the next image (**Table 1**). All outputs are stored as text files, and the values of spot total and spot mean are stored in files for each spot, while values for each cell are stored in files for each cell in a directory for each spot. By plotting these numerical values, it is possible to survey and analyze the changes in intensity and shape of cells over time. In addition, for verification, the image files with the outlines of the cells recognized are generated in lower resolution (**Fig. 3**).

Table 1 Values calculated by our software system.

For each image	shape of cells (as an image file) total and mean intensity for each spot
As time series	total and mean intensity for each cell circularity and centroid of each cell correspondence of cells between images time series of calculated values for each cell

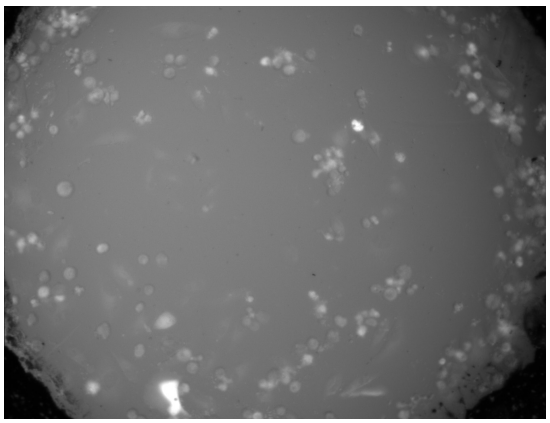

Fig. 2 Example of a photographed image of a spot. An edge of a circle shaped spot appears at four corners of the image.

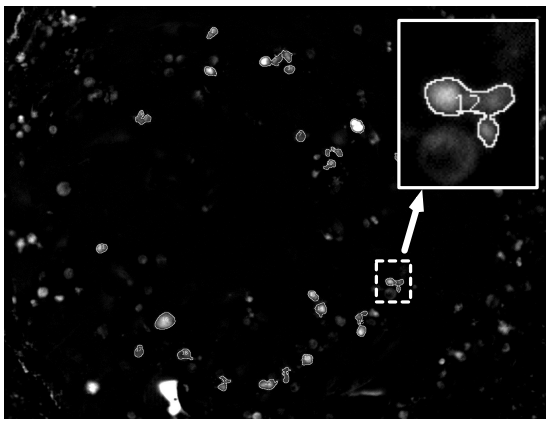

Fig. 3 Processed image from photograph of Fig. 2. Background luminance has been subtracted. Each recognized cell is bordered by solid line and numbered inside their border (numbered as '12' in this example shown at upper-right corner).

2.3 Processing Method

Each image is downloaded into a computer and then 1) background correction, 2) threshold value processing, and 3) shape recognition and luminance intensity integration are performed sequentially. After processing of all the images has been completed, cell tracking is done concerning whether the cells in multiple images are the same cell or not (**Fig. 4**).

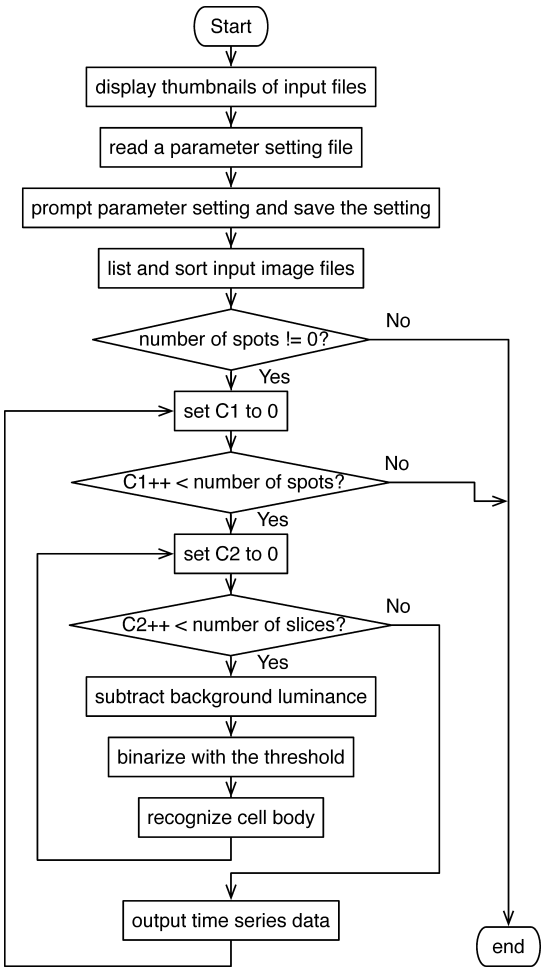

Fig. 4 A flowchart of the image processing algorithm of our system. Outer (C1's) loop illustrates process for each spot. Inner (C2's) loop is for each image file consisting of background subtract, cell body recognition, integration of intensity and calculation of circularity and centroid. Each process is applied in this order.

2.3.1 Background Correction

Taking into consideration that they are images of round spots, picture elements (pixels) of parts of the images other than the cells are removed by the shading correction [11)~13)]. First, the scale of the image is reduced to 1/10 in order to shorten the processing time. An image is smoothened using morphology filters (expansion/contraction filters) with the radius of 2 pixels and the block noise generated by expanding to the original size is smoothened using the

Gaussian filter [13]. The resulted image by these processes is used as the background image. The luminance intensity of the background image is subtracted from the original image, and then the corrected image is obtained by adding the mean intensity of luminance of the background image to each pixel of this image that underwent the subtraction.

2.3.2 Threshold Value Processing

The threshold value is subtracted from the entire image in order to separate the cells and the background. First, the initial value of the threshold value is calculated using the maximum entropy algorithm [14]. Since cells that appear dark are eliminated if this value is subtracted from the entire image, the value from which the standard deviation of the luminance intensity over the entire image has been subtracted is taken as the threshold value.

2.3.3 Cell Shape Recognition

For the entire image, binarization is performed by assigning pixels that have higher luminance intensity values than the threshold value described above as 1, and those with lower pixels as 0. For pixels that are 1, the cell shape is obtained by tracing the contour line to make closed curves [13]. A number is assigned to each closed curve. On the background corrected image, the values of intracellular luminance intensity are added up, and the total of the luminance intensity values of all the pixels in the cell areas in one image is used as the luminance intensity value of that image. The mean luminance intensity value is derived by dividing the luminance intensity value of the image by the total area of cells.

The values of luminance intensity of each cell in the image are calculated, and simultaneously the centroid coordinates and circularity ($4\pi\times$ area/circumference) are also calculated.

2.3.4 Cell Tracking

Based on the location of the centroid of a cell in each image, time series data consisting of the luminance intensity, mean intensity of luminance, centroid, and circularity for each cell are constructed for the cell are assumed to be the same cell in more than one image.

Two consecutive images are compared and it is concluded that a cell in the chronologically preceding image and a cell within the subsequent image whose centroid is the closest to the previous cell and is within the judgment distance, are the same cell. The judgment distance is taken to be a given threshold depends on a diameter of circle whose area is the same as the average area of cells in a preceding image. For this reason, when observing cells with a large amount of movement, it is necessary to shorten the photographing interval. If there is no corresponding cell in the later image, it is assumed that the cell has disappeared. However, if a cell that meets the conditions appears in a later image, it is assumed to be the same cell. On the other hand, if a cell appears where there was no cell, it is assumed to be a new cell.

3. Implementation

The shading correction, Gaussian filter, morphology filter, a part of threshold value processing, and shape recognition implemented for ImageJ v. 1.37 are used. Background correction, another part of threshold value processing, calculation of circularity, and time series processing have been programmed newly and prepared as plug-ins for ImageJ. The program was constructed so that it could be run as the parallel execution by the native thread function of Java 1.4.2 or higher runtime environment.

4. Case Study

Twelve types of virus vectors were introduced into HeLa cells, and we processed images of 9 types of gene reporter vectors that were observed by one experiment using one TFA chip. Figures 2 and 3 had been obtained by this experiment. The experimental conditions were $12 \times 9 = 108$ sets, and 4 spots were assigned for each condition. By statistically processing these 4 spots, we attempted to improve the reliability of calculated values. The total number of spots was $108 \times 4 = 432$, photographed over 42 hours every other hour (photographed a total of 43 times). The total number of images was 18,576 (432×43). A size of obtained image files is 2800 (width) $\times$ 2133 (height) pixels. The format is 12-bit depth gray-scale TIFF. Approximately it is 40 GB in total.

Typical shape of a cell is a circle whose diameter is approximately 100 pixels. Especially dead cells show apparent circle shapes by its surface tension. In this experiment, most of both living and dead cells do not move because they are stick to the glass surface of their spot. So we have set judgment distance (Section 2.3.4) to 2.5 $\times$ larger than the average circle. If we take more larger value for the distance, several separated cells are recognized as identical cells even if they completely stick. If the value is

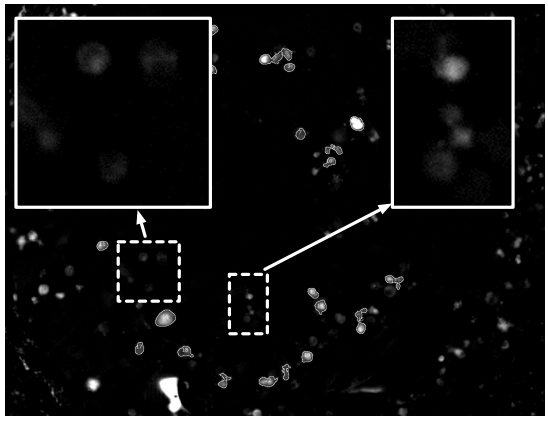

Fig. 5 Effect of background subtraction and cell size threshold. Original image is shown as Fig. 2. Left solid line square indicates cells subtracted simultaneously with background luminance. Right solid line square indicates cells having smaller area than a circle which has a lower limit of specified diameter (Section 2.1.1).

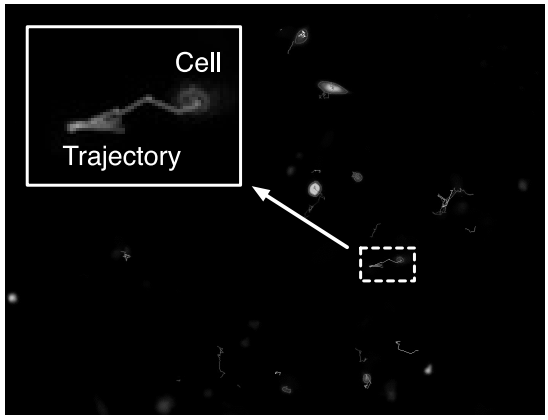

Fig. 6 Recognized cells and its trajectories. A trajectory is a time series of centroid coordinate of a cell. It is saved in a text file automatically.

too small, identical cell are not recognized after they divide. The distance should be chosen for each experimental condition.

Figure 5 shows automatic elimination of ambiguous cell shapes by background subtraction and specifying the lower limit of cell size. Each cell which has dark shape (in the left solid square) and too small shape (in the right square) is not recognized as a cell and ignored when performing digitizing of luminance of cells and spots and tracking of cells.

A result of cell tracking is shown in **Fig. 6**. Trajectories are tracked for all recognized cells. If a cell disappears and appears again in the distance threshold, the later cell is recognized

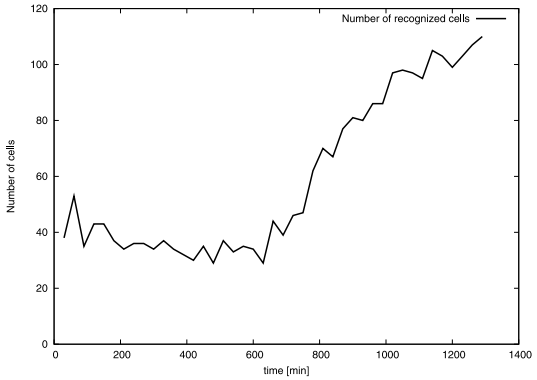

Fig. 7 Change of the number of recognized cells on a spot in time. Because of observed cells are cancer cells (HeLa cells), the number is approximately increasing. Hoever, it shows small decrease at many points.

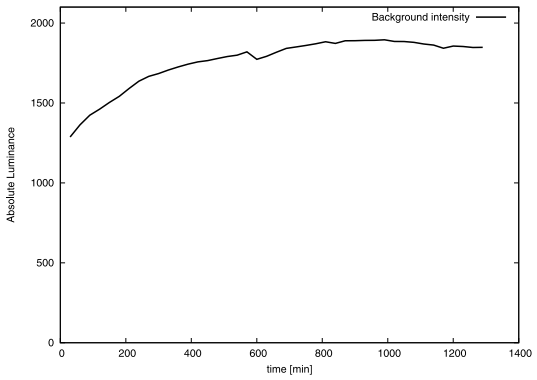

Fig. 8 Change of background luminance of a spot in time. Its increase may caused by progression of transfection of reporter vectors.

as same cell as former one, and the trajectory is represented by broken lines that have a same color.

For each spot of this experiment, averagely 51.1 cells were recognized (SD 19.4). A plot of a number of recognized cells changing in time of the spot No.43 is shown in **Fig. 7** and changes of estimated background luminance of the same spot is shown in **Fig. 8**. An image of this spot at the last time point is shown in Fig. 2. Due to a kind of observed cells (Hela, a well established cancer cell stock), total number of cells is increasing (Fig. 7). Average number of cells of the spot is 59.4. Simultaneously, by the progression of reporter vector's transfection total luminance of the spot is increasing and it affects to background luminance (Fig. 8).

A number of cells are varying depending on spot's condition. Time series of mean luminance of each cell in the spot No.87 are shown

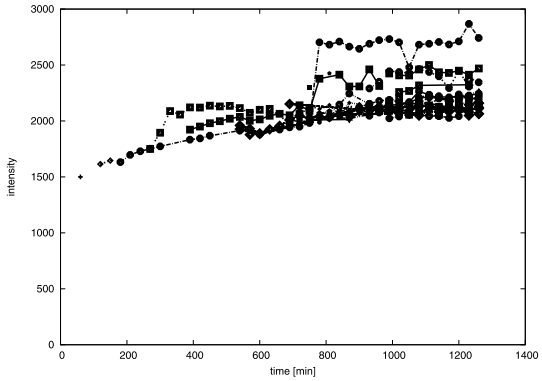

Fig. 9 Average intensity of all cells in a spot (total intensity of each cell divided by area of the cell). Each line corresponds to each cell. Total number of cells is 56.

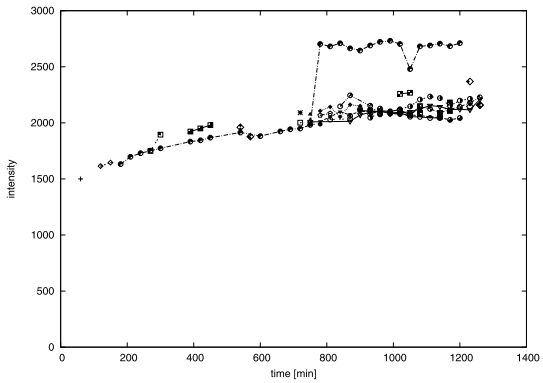

Fig. 10 Plots of cells except whose circularity are 0.8 or upper. Total number of plotted cells is 29 (27 cells have been assumed dead, and eliminated from Fig. 9). Compared with a cell which has the highest intensity in Fig. 9, a plot of correponding cell in this figure ends earlier than in Fig. 9. It means the cell shape changed to circular and judged the cell died at the end of the plotted line.

in **Fig. 9**. Average number of recognized cells in No.87 is 56. In this experiment, taking into consideration that dead cells would nearly peel off the glass surface and become circular due to surface tension, we assumed those cells with a high circularity index (0.8 or more) had died. By eliminating these cells we are able to follow the changes in intensities and shapes of cells over time for only viable cells. The plot is shown in **Fig. 10**. Compared a cell whose intensity is the highest in both two figures, a plot of the cell is terminated earlier in Fig. 10 than Fig. 9. The terminating point is determined by the circularity index. It means a plot is ended at a time of death of the cell.

The program was run in parallel on a SMP calculator which had 4 dual-core CPUs. All of the images were processed and it took approximately 8 hours to produce the whole time series data (Dual-core Opteron 886/2.4 GHz/1 MB cache × 4 CPU (8-core SMP architecture), memory 32 GB, J2RE 1.4.2, Linux kernel 2.6.5).

5. Discussion

We are able to follow and simultaneously measure the activities of 9 types of genes as well as cellular shape in each viable cell within the cell array spots using the software we developed. Since luminance intensity is not lost even if a cell dies (exponential decrease with half-life of several hours), when fluorescence only is observed, there is concern that gene activity will be calculated from both dead and living cells. However, this problem can be avoided by judging the viability of a cell based on its shape, making it possible to achieve an analysis with a higher degree of reliability.

In molecular biology or biology in general, if there is a large volume of images, the processing of the images usually progresses as far as being semiautomated because there are great differences in the characteristics of the images depending on objects being observed and the method used for observation. In other words, if we consider the case of cell array, we see visually with our eyes how the cells should be recognized, and if there are cells that should be recognized but are not, this should be corrected, and in the case in which noise is miss-recognized for a cell, all miss-recognitions should be pointed out in general. With the software we developed as well there are cells that cannot be recognized, however, it can be confirmed with the image that miss-recognition of noise is slight by specifying the size of the cell and performing background correction. It is believed high quality numerical values can be obtained in this manner.

Our method can process varying background luminance in time, however, dark cells in a totally bright spot cannot be recognized, even if the shape is clear. Cells that can be clearly identified when seen by the naked eye are eliminated simultaneously with the background luminance in such images. In order to solve this problem, other filters such as differentiation filters, fast Fourier transformation (FFT) filters, wavelets, etc. should be introduced, or background calculation should be improved as such

that initial threshold value is set lower and background luminance is calculated for each cell, etc. Template matching for cell size and shape recognition will also improve measurements and judgments. Although these methods require more computational resources, possibly improve cell recognition performance.

If it is assumed that approximately 30 seconds are required to perform the same processing using the semiautomatic software that comes with the experimental equipment used in the photographing, the present data would have taken about 155 hours to process, which equals 19 days at 8 hours per day. Taking into consideration that the observations were completed in slightly less than 2 days, it can be seen that image processing acts as a bottleneck. In order that the high through-put of cell array is not lost, in the very least the processing should be faster than the observation period. The processing system that we developed and described here achieved this by parallel computing. When executed on a computer with a single CPU core, it is envisioned that it would take 60 hours for the same processing, which would not be able to keep up with the pace of the experimental observations. Furthermore, because a large increase in the calculation time is expected if intelligent algorithm are introduced for a noise filter or shape recognition, increasing the speed by parallel computing is inevitable, and it is believed not only SMP computing but also grid computing should be used.

Acknowledgments This work has been supported by the project for development of gene function analysis technology based on cell array from the New Energy and Industrial Technology Development Organization, Japan.

References

1) Ziauddin, J. and Sabatini, D.M.: Microarrays of cells expressing defined cDNAs, *Nature*, Vol.411, pp.107–110 (2001).
2) Bailey, S.N., Wu, R.Z. and Sabatini, D.M.: Applications of transfected cell microarrays in high-throughput drug discovery, *Drug Discovery Today*, Vol.7, No.18 (Suppl.), S113–S118 (2002).
3) Yoshikawa, T., et al: Transfection microarray of human mesenchymal stem cells and on-chip siRNA gene knockdown, *Journal of controlled release*, Vol.96, 227–232 (2004).
4) Kato, K., Umezawa, K., Miyake, M., Miyake, J. and Nagamune, T.: Transfection microarrays of nonadherent cells on an oleyl poly (ethylene glycol) ether-modified glass slide, *Biotechniques*, Vol.37, No.3, pp.444–452 (2004).
5) Neumann, B., et al.: High-throughput RNAi screening by time-lapse imaging of live human cells, *Nature methods*, Vol.3, No.5, 385–390 (2006).
6) Abramóff, M.D., Magalhães, P.J. and Ram, S.J.: Image processing with ImageJ, *Biophotonics International*, Vol.11, No.7, pp.36–44 (2004).
7) http://rsb.info.nih.gov/ij/
8) Shah, N.A., Laws, R.J., Wardman, B., Zhao, L.P. and Hartman IV, J.L.: Accurate, precise modeling of cell proliferation kinetics from time-lapse imaging and automated image analysis of agar yeast culture arrays, *BMC Systems Biology*, Vol.1, No.3, pp.1–14 (2007).
9) Lamprecht, M.R., Sabatini, D.M. and Carpenter, A.E.: CellProfiler: Free, versatile software for automated biological image analysis, *BioTechniques*, No.42, No.1, pp.71–75 (2007).
10) Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., Golland, P. and Sabatini, D.M.: CellProfiler: Image analysis software for identifying and quantifying cell phenotypes, *Genome Biology*, Vol.7, Issue 10, Article R100 (2006).
11) Serra, J.: *Image Analysis and Mathematical Morphology*, Academic Press, New York (1982).
12) Wählby, C., Sintorn, I.-M., Erlandsson, F., Borgefors, G. and Bengtsson, E.: Combining intensity, edge and shape information for 2D and 3D segmentation of cell nuclei in tissue sections, *Journal of Microscopy*, Vol.215, No.1, pp.67–76 (2004).
13) Woods, G.: *Digital Image Processing*, Prentice Hall, New Jersey (2003).
14) Shannon, C.E.: A Mathematical Theory of Communication, *The Bell System Technical Journal*, Vol.XXVII, No.3, pp.379–423 (1948).

(Received April 12, 2007)
(Accepted August 27, 2007)

(Communicated by *Susumu Goto*)

Daisuke Tominaga received B.E. in 1992, M.E. in 1996 and Dr. of Information Engineering in 2001 from Kyushu Institute of Technology. He is currently a Research Scientist at Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology from 2001. His current research interests include linear and nonlinear modeling for time development of biological network systems, model selection for time series data, heuristic non-linear numerical optimization, statistical noise reduction and symbolic computation for differential equation models of biological networks.

Fukumi Iguchi studied acoustic holography under supervision of Professor Keinosuke Nagai, and received B.E. in 1998 and M.E. in 2000 from University of Tsukuba. She is currently working for development of image processing algorithms and procedures mainly at the Information and Mathematical Science Laboratory, Inc., and has been developed softwares for analyses of photographic images from wide variety of specific area, especially for images by artificial satellites, and microscopic images of development and differentiation of living nerve cells and cancer cells.

Katsuhisa Horimoto received Ph.D. from Tokyo University of Science in 1991. After positions of assistant professor of Tokyo University of Science, associate professor of Saga Medical University and professor of the University of Tokyo, he is currently a leader of the biological network team at Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology. His current research interests include algebraic biology and graphical models.

Yutaka Akiyama received B.E. in 1984, M.E. in 1986 and Ph.D. in 1990 from Keio University. After positions of Research Officer at Electrotechnical Laboratory of the National Industrial Technology Center, Associate Professor at Molecular Biology and Information Research Section, Kyoto University, Director of Parallel Applications TRC Laboratory of the Real World Computing Partnership, Engineering Research Association, Chief Researcher of Electrotechnical Laboratory of the National Industrial Technology Center and Director of Computational Biology Research Center, Advanced Industrial Science and Technology, he is currently a professor of Graduate School of Tokyo Institute of Technology. His current research interest includes parallel computing architecture, large scale simulation, pathway analysis of biological networks and analysis of mass-spectrometry data.