MIA: A UIMA-Based Microarray Image Analysis System

Wei-Bang Chen[1], Chengcui Zhang[1], Wen-Lin Liu[2], and Richa Tiwari[1]
[1]*Department of Computer and Information Sciences,*
[2]*Department of Accounting and Information Systems,*
University of Alabama at Birmingham, Birmingham, Alabama 35294, USA

Abstract

Microarray is an emerging technology that allows biologists to monitor expression levels of thousands of genes in parallel. In this proposal, we report an UIMA based microarray image analysis (MIA) system for solving two existing critical problems in microarray image analysis and data management. MIA system is designed to offer a flexible, scalable, and extensible environment, and to provide accurate microarray analysis results without the need of manually correcting the image processing results.

1. Introduction

Microarray is a chip-based tool which simultaneously monitors expression level of thousands of genes, and thus, is ideal for functional genomics study. In the experiment, test and control specimens, labeled with red and green fluorescent dyes respectively, are hybridized with affixed probes on the slide. Complementary specimens and probes are tightly bound, and unbound specimens are washed off. The intensity values of the two fluorescent dyes are detected, and expression level for each gene is defined as the red-to-green intensity ratio of each spot. However, there exist challenges in microarray image analysis and data management.

There are two major issues in microarray image analysis: (1) spot addressing, and (2) spot segmentation [1]. For spot addressing, existing approaches are not fully automatic and require users to provide layout parameters, also cannot solve the errors/imperfections in block alignment since they process the entire slide as a whole. For spot segmentation, manufacturing defects and improper operations may have adverse impacts on data precision and the subsequent analysis [2].

Data management is another issue. Microarray experiments contain abundant information, such as specifications and annotations of probe sets, and descriptions and conditions of specimens. Slide images and their related information are unstructured since computer cannot understand the contents and extract meanings directly. However, the information contains direct and indirect evidence for knowledge discovery and need to be managed efficiently.

In previous work, we reported a fully automatic analysis algorithm for spot addressing and fluorescent intensity measuring in [3]. In this study, to manage the unstructured data more efficiently, the algorithm is tightly integrated with the component-based Unstructured Information Management Architecture (UIMA) as the microarray image analysis (MIA) system. UIMA requires domain experts to implement *Analysis Engines* (AEs) which enable computers to extract useful information from unstructured contents, such as images, and store extracted contents as annotations in a *Common Analysis Structure* (CAS), an object-based data structure, for representing and sharing the structured information in the framework. MIA system provides the flexibility, scalability, and extensibility of data management [4].

In this proposal, we briefly introduce MIA system in Section 2. Section 3 compares MIA with related systems. Section 4 summarizes this system.

2. MIA system design

Our goal is to design an automatic system for extracting spot intensity values accurately from red and green channels of a microarray slide. The system first addresses each spot, and then extracts fluorescent signals from each spot. The addressing process detects and corrects slide tilt, discovers block boundaries, generates gridlines for each block, and finally recognizes spots on a slide. The signal extracting process applies a segmentation algorithm to retrieve spot intensity values accurately.

MIA system follows a Model-View-Controller (MVC) design pattern, and is implemented based on the UIMA framework. It offers a friendly GUI for specifying imported and exported data locations. Moreover, users can easily plug in various analysis components for performing various analysis tasks. MIA system includes four major modules: (1) Slide Information Module, (2) Slide Blocking Module, (3) Slide Gridding Module, and (4) Slide Segmentation Module. Modules are implemented as either a primitive AE or an Aggregate AE. A primitive AE contains an Annotator and a Component Descriptor. An annotator is codes for analyzing unstructured contents. A component descriptor is a XML file describing the data structure and the input/output requirements of the annotator. An aggregate AE includes one or more primitive AEs, and it contains only a component descriptor with additional

0-7695-3084-2/07 $25.00 © 2007 IEEE
DOI 10.1109/ISM.Workshops.2007.24

information about the data processing flow involved in primitive AEs. Slide Information Module analyzes parses, retrieves slide information in XML documents, and stores information in CAS for further analysis. This module collaborates with an agent-based automatic information update module to retrieve the latest gene annotations from various public databases, thus can provide the up-to-date information to researchers. Slide Blocking Module includes three sub modules: signal/noise detector to identify foreground and background pixels, tilt detector to detect and correct tilted slides, and block boundary detector to separate blocks. Slide Gridding Module is realized as an aggregate AE which has two primitive AEs (bounding box generator and grid line detector). This module generates grids for blocks and stores grids as annotations in CAS. Slide Segmentation Module segments each cell in grids and extracts real signal pixels from each spot. We use Otsu's thresholding algorithm in a progressive manner to get a local threshold for a spot region by minimizing the intra-class and inter-class variance of pixel intensities [5].

3. Comparisons with related systems

We applied MIA system on 8 slides (48 blocks/slide), the slide blocking module correctly returns 384 identified blocks. The recall and precision are both 100%. The result shows the robustness of proposed algorithm. For the performance of automatic gridding on 384 blocks (60 gridlines/block), our method missed 175 gridlines, and returned 1 false gridline. The recall and precision are 99.24% and 99.99% respectively. The proposed method discovers the grid template information automatically (e.g., number of rows/columns), while the existing software require users' input to proceed. Thus, it is improper to compare our performance with those in [6-8].

In spot segmentation, we compare MIA system with commercial software GenePix, which adopts a circle with adaptive radius, as shown in Figure 1. The 1st row shows an irregularly shaped spot with noise pixels. MIA system successfully removes the noise while GenePix fails. The 2nd row shows a donut-shaped spot with signals on both the outer circle and in the center. Again MIA system successfully identifies the signals and removes the background pixels on the intermediate circle, while GenePix misrepresents signals in that spot. The 3rd row shows a regularly shaped spot surrounded by non-signal pixels at the outer rim. MIA system correctly identifies the central signal spot, while GenePix produces lots of false positive pixels.

Comparing both systems applied on 187 donut-shaped spots and 191 noise-contaminated spots, the success rate of MIA system is 89.84% for donut-shaped spots and 90% for noise-contaminated spots, while GenePix cannot handle donut-shaped spots with the adaptive circle method. Besides, the success rate of GenePix for noise-contaminated spots is only 54%.

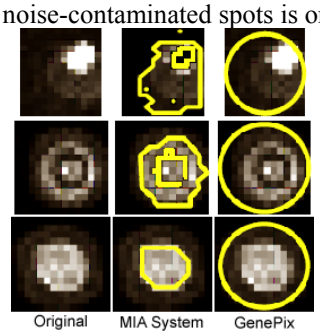

Figure 1. Segmentation results

4. Summary of MIA system

MIA system accurately addresses each spot, and handles uneven background and noises in slide. Noise removal is one of the major contributions in this study. A simple yet effective method for spot segmentation is proposed and applied to each cell in grids. The experimental results show that MIA system is robust and outperforms GenePix. Also, we implement an information update module to retrieve related information from online public databases like MeSH and update corresponding information in database.

5. Acknowledgement

The research of Dr. Chengcui Zhang is supported in part by NSF DBI-0649894.

6. References

[1] O. Demirkaya, M. H. Asyali, and M. M. Shoukri, "Segmentation of cDNA microarray spots using markov random field modeling," *Bioinformatics*, vol. 21, pp. 2994-3000, 2005.

[2] A. A. Ahmed, M. Vias, N. G. Iyer, C. Caldas, and J. D. Brenton, "Microarray segmentation methods significantly influence data precision," *Nucleic Acids Res*, vol. 32, pp. e50, 2004.

[3] W.-B. Chen, C. Zhang, and W.-L. Liu, "An Automatic and Robust Method for Microarray Image Analysis and the Related Information Retrieval for Microarray Databases," presented at ICDE Workshops, 2007.

[4] D. L. Ferrucci, A., "Building an example application with the Unstructured Information Management Architecture," *IBM Systems Journal*, vol. 43, pp. 455-475, 2004.

[5] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, pp. 62-66, 1979.

[6] A. N. Jain, T. A. Tokuyasu, A. M. Snijders, R. Segraves, D. G. Albertson, and D. Pinkel, "Fully automatic quantification of microarray image data," *Genome Res*, vol. 12, pp. 325-32, 2002.

[7] M. Katzer, F. Kummert, and G. Sagerer, "A Markov Random Field model of microarray gridding," in *Proceedings of the 2003 ACM symposium on Applied computing*. Melbourne, Florida: ACM Press, 2003.

[8] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Res*, vol. 30, pp. e15, 2002.