AUTOMATED SYSTEM FOR IMAGE ANALYSIS OF YEAST COLONIES: A NOVEL APPLICATION IN FUNCTIONAL GENOMICS

Negar Memarian[1], Javad Alirezaie[1,2], Ashkan Golshani[3]

1. Department of Electrical Engineering, Ryerson University, Toronto, Canada.
2. Department of Systems Design Engineering, University of Waterloo, Waterloo, Canada.
3. Department of Biology, Carleton University, Ottawa, Canada.

nmemaria@ee.ryerson.ca javad@rousseau.uwaterloo.ca agolshan@connect.carleton.ca

ABSTRACT

An automated image analysis system has been implemented for a novel genomics application. The biologists are interested in exploring the effect of various drugs on functionality of genes. Study of size change in drug treated colonies of yeast, implies information about those gene pathways that are affected by the drug. The role of developed system is to distinguish and extract true yeast colonies from other objects in a digital image, accurately measure their area, and provide a coordinate oriented map of colony areas. The developed system also executes post processing calculations and presents useful statistical parameters associated with corresponding colony pairs. A precision test is designed to monitor the precision of experimentation trials. Image processing techniques such as spatial adjustments, segmentation and region growing are utilized in development of system. Preliminary results show robustness and significant improvement of this system over conventional methods.

1. INTRODUCTION

In recent years functional genomics has attracted considerable attention from a variety of disciplines such as computer science, mathematics, engineering, chemistry and of course life and medical sciences. This field has been described as the development and application of global (genome-wide or system-wide) experimental approaches to study various biological questions such as those in gene function, mode of action of bioactive chemicals (such as drugs) and protein interaction mapping, by making use of the information provided by genome sequencing and mapping projects [1]. Integration of different functional genomic data enables biological hypotheses to be formulated with increasing levels of confidence [2].

In the present work, biologists are interested in 1) identifying novel functions for genes in yeast *Saccharomyces cerevisiae* and 2) elucidating the molecular mechanism of action for various bioactive chemicals. Yeast is the most commonly used model system in biological research. One way of accomplishing this goal, is by generating deletions (in large scale) in two target genes, the function of one is known and the other is not. In this approach the first gene is disabled by a deletion mutation, and the second gene is deactivated by either a second mutation or by the use of a target bioactive compound. If deletion of both genes combined results in slow or altered growth of yeast colonies, then one can hypothesize that these two genes are genetically interacting and thus their functions are related. Similarly if the second gene is deleted by means of a chemical treatment then it can be hypothesized that the chemical targets the same pathway as the first mutation. There are currently no automated tools published which can efficiently analyze such produced data. Previous strives for screening colony size change were manual, i.e. by human eye and therefore quite laborious and error prone. As a result, quantitative analysis of a pool of thousands of colonies proved to be impossible.

Here we report the development of a computerized system that performs accurate colony area measurement and calculates required statistical parameters. From the image processing point of view, comparable work has been done in micro scale, in the dissimilar area of DNA microarray addressing or gridding. Against the computationally expensive nature of most of those works [3], [4], [5], the present image analysis method produces accurate results in large scale with simplicity and with no need for complex and expensive equipment. Our work is a novel application of image processing and analysis in the developing area of functional genomics. The developed system is robust to experimental deviations.

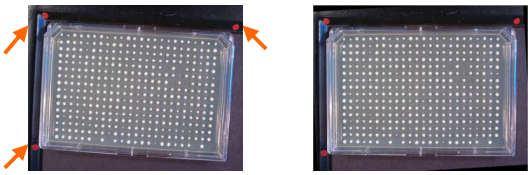
In the following section the material and development methods will be discussed. Then the results will be presented. We will conclude our statements in the conclusion section.

2. MATERIALS AND METHODS

The developed automated system processes digital images in Joint Photographic Experts Group (JPEG) format. It performs three stages of preprocessing, processing and post-processing on the input images.

2.1. Preprocessing

The purpose of this stage is to equalize the zoom, angle and size of input digital images. Inconsistency of any of these parameters can lead into big errors in the results of analysis. The plate of yeast colonies is placed inside a holder tray at time of image capture. First, color recognition is used to detect the margins of plate. Three identical red color indicators are positioned on the holder tray (figure 1-a). The system seeks red color components in the image and therefore recognizes the approximate region occupied by plate. As it is apparent in the example of figure 1-a, the plate image may be tilted. The system aligns the plate image by taking pairs of control points and uses them to infer a spatial transformation. One



(a) (b)

Figure 1: (a) A sample input image to the system. Arrows show the red indicators. (b) Image in (a) is aligned using a linear conformal spatial transformation.

set of control points, namely the input points, is the centroid coordinates of the top two red indicators. The other set is a pair of points on a precisely horizontal line, which are a fixed distance apart. The latter is called the alignment points. The algorithm calculates both the scale image should be resized and the angle it should be rotated to attain the standard alignment. Figure 1-b shows an image aligned using this method.

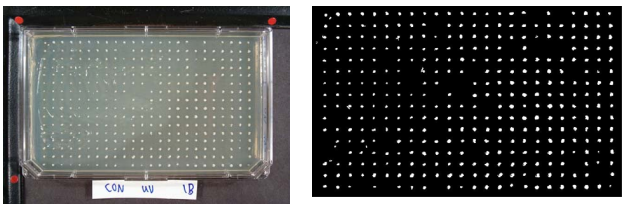
In the last step of preprocessing and before entering the actual processing stage, background objects such as plate frame, metallic tray, red indicators, plate name sticker and etc. are removed from the image.

2.2. Processing

In order to extract the regions of interest –yeast colonies-, we apply image segmentation. First the cropped image is binarized through global thresholding. Then a size and shape analysis is performed to distinguish between true colonies and mistaken blemishes. Only those objects with size greater than 0.0005 of total area of bright pixels in thresholded image are reserved. The value 0.0005 is reasonably small not to allow undesired isolated pixels and has been selected empirically. Also according to our *a priori* knowledge of the shape of regions of interest, we know that colonies normally keep a rather circular shape. Therefore the objects in the image are examined based on their eccentricity and consequently long entities are filtered. An example of segmented image is depicted in figure 2.

Next step of work is providing a coordinate oriented map of colony areas. Conventional labeling and region property detecting methods do not work accurately for this purpose. We combine region growing with an adaptive object specific mask to handle this issue. For every object (colony) present in the image, a mask is automatically formed, which is a rectangular area slightly bigger than the object. The dimensions of this mask are specific to the shape and size of that particular object. Sum of the bright pixels that fall inside the object specific mask is stored as the total area of that object in an array at the coordinates of object's centroid. We have named this array area map, as it in fact maps the total area of each object to its centroid. This method is advantageous in two ways: Firstly by applying the object specific mask we ensure that an accurate measurement of object area is performed (i.e. a part of the object does not get eliminated or missed due to wrong window size). Secondly by placing the area value of each object at location of its centroid, we can keep track of the correlation between size and position of individual objects later on.

The area map needs to be ordered in the order of master gene list. Master gene list is a database containing thorough information about those genes that are being traced by the plate analysis research. According to the gene list, the bottom-right object in each plate denotes object 1 and the top-left object in that plate represents object 384 (16x24). The objective of this step is to come up with



(a) (b)

Figure 2: The automated system can detect colonies from other undesired objects. (a) Unwanted long entities and small blemishes are present in the image. (b) The system correctly discards most of these objects as can be seen in the binary image after segmentation.

an array that has the results in order from coordinate (1, 1) to coordinate (16, 24). This may seem like a quite trivial task, since the colonies appear in an almost organized 16x24 grid. However it is interesting that dealing with this part of analysis proved to be the trickiest step of work. That is because, very often the colonies do not appear in a strictly aligned grid and since the system is aimed to work accurately without human interaction, it is required to take such deviations into consideration automatically. To solve this issue, we obtain the logical version of area map, which is a binary map of object centroids. The result, named centroid map is a matrix of zeros the same size as thresholded plate image, with ones only at location of colony centroids. The peripheries of map are calculated from prototype equations below:

$$y\_start = min\_y\_centroid - \text{round}(max\_y\_width/2) \quad (1)$$

$$x\_start = min\_x\_centroid - \text{round}(max\_x\_width/2) \quad (2)$$

$$y\_end = max\_y\_centroid + \text{round}(max\_y\_width/2) \quad (3)$$

$$x\_end = max\_x\_centroid + \text{round}(max\_x\_width/2) \quad (4)$$

where the first term at right hand side of equality in each relation is the minimum or maximum centroid coordinates along x or y direction. The second term at right side of equality is half of the maximum object width value. Here occurrence of a worst case should be taken into account. There is a chance that from the four peripheral sequences of colonies, one whole row or column is missing. An example can be seen in image 3-a where the down bottom row of colonies does not exist. This leads to wrong calculation of *y_start* and *x_start* in equations 1 and 2. To compensate for such cases, a distance check has been designed to verify the coordinates of starting and ending margins. If any of the four values at left hand side of equations 1 to 4 are not less than a specific threshold, then a predefined value is imposed for that quantity. To illustrate this more clearly, in figure 3-b, that value is shown as a sequence of stars imposed on the original image. As it is obvious, now the star coordinates fall outside the threshold rectangle, meaning calculation of *x_start* and *y_start* based on these points would now be correct. The threshold is supposed to be 100 pixels distant from the image border line for each side. The predefined values are chosen empirically based on the average location of colony centroids in many images.

As the last step of processing phase, the centroid map and area map are utilized to order the colonies. Beginning from *x_start* and *y_start* and heading for *x_end* and *y_end,* regions of m x n size from centroid map are examined. For every nonzero value found in that region, its corresponding area value from area map is added to an accumulator for that region. Parameters m and n are attained from (5) and (6).

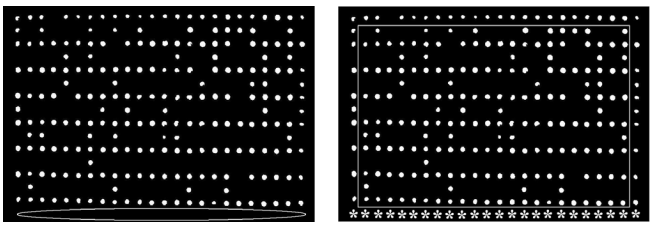


Figure 3: (a) Bottom row of colonies missing. This is marked on the image by a white ellipse, (b) White rectangle showing position of distance checking threshold. Each side of this rectangle is 100 pixels distant from the plate side parallel to it. Imposing the predefined values (stars) compensates for the missing row.

$$m = (x\_end - x\_start) / 24 \quad (5)$$

$$n = (y\_end - y\_start) / 16 \quad (6)$$

Role of accumulator is noteworthy because of heeding those objects which have small disconnected parts. Although the colonies are usually compact objects with one definite centroid, there are cases where a colony is comprised of a big chunk with a few smaller islands around it. Without the accumulator, the ordering algorithm would only consider the area of the first object detected in the examination region as the colony area in that coordinate. With the accumulator however, the system continues the process of centroid detection and adding to the accumulator until there is no more nonzero centroids left in the examination region. A check is executed when the ordering step is finished. This check compares the total value of bright pixels before and after ordering. If these values are not identical, it is revealed that some information has been missed in the midst of ordering.

2.3. Post processing

Some statistical analysis is performed on the final result of processing stage. For every plate the average value of bright pixels is calculated from equation 7.

$$S_{ave} = 1/N \sum_{i=1}^{N} S_i \quad (7)$$

where N is the total number of bright objects present in plate and S_i is the area of object *i*.

The deviation of area of each object from plate's average is calculated by subtracting the scalar S_{ave} from the plate's 1D ordered area array (equation 8).

$$\Delta S_i = S_i - S_{ave} \; ; \; i = 1, \ldots, 384 \; (16 \times 24 = 384) \quad (8)$$

The ultimate objective of this analysis is to monitor the change in colony area in two corresponding experiments, in one of which the yeast colonies are treated with a specific kind of chemical/drug and in the other one, they are left plain. The area difference and percentage of reduction between the two plates plus the normalized value of these quantities are therefore calculated to provide a thorough understanding of area change for each colony within its plate and also compared with its corresponding colony in another plate.

For the biologists to say that a drug is effective on gene deletion X, the entire array has to be run more than once. For this study so far a few of the experiments have been selected and run for at least three times to find out how likely it is that the observed area difference may not be due to pure chance. A ranking scheme has been designed to evaluate agreement (harmony) between the results of sequential trials. Table 1 presents a general case. When three runs of the same experiment are performed, total of ten different combinations of growth, shrink or unchanged size may happen. This is because order does not matter here (i.e. growth, shrink, shrink is the same as shrink, growth, shrink and shrink, shrink, growth. The same condition applies for other combinations). As columns two to four of Table 1 suggest, a score of +1 is assigned to every experiment in which area difference of a pair of corresponding colonies is a positive number, a score of -1 where it is a negative number and 0 where colony area remains unchanged. The mean and standard deviation of scores of three trials are calculated. If standard deviation is big, it means that the results are more due to chance than being due to a true pattern.

Table 1. Classification based on the trend of size change in a sequence of three experiment trials. A.R. stands for agreement of results in three trials: 3=absolutely agree, 2=partially agree, 1=disagree.

Combination	T 1	T 2	T 3	η	σ	A.R
1	+1	+1	+1	1	0	3
2	-1	-1	-1	-1	0	3
3	0	0	0	0	0	3
4	+1	+1	0	2/3	0.816	2
5	-1	-1	0	-2/3	0.816	2
6	0	0	+1	1/3	0.816	2
7	0	0	-1	-1/3	0.816	2
8	0	+1	-1	0	1.414	1
9	+1	-1	-1	-1/3	1.633	1
10	-1	+1	+1	1/3	1.633	1

3. RESULTS

We evaluated the efficiency of our algorithm to detect growth differences in yeast colonies, by comparing colony growth for yeasts treated with various compounds to the untreated ones (used as control). In our experiment with cobalt, we plated equal number of mutant yeast cells on a media containing 0 ug/ml (used as control) or 400 ug/ml of cobalt. The mutant strain shows a significant growth reduction in presence of 400 ug/ml of cobalt indicating sensitivity to the drug. This is shown in figure 4. The spot on the very right of each experiment contains 10^4 cells. The following spots approximately contain 10^3, 10^2 and 10^1 cells, respectively.

We used a spot test analysis to measure the true sensitivity of all the 384 strains on the plate. To do this, wild/type (normal) yeast cells as well as the mutants were first grown in liquid media to saturation (stationary phase). The cells were then diluted 50 times and allowed to grow to mid-saturation (mid-log). The cells were then diluted using serial-dilution and decreasing numbers of cells were plated on the solid media with or without (used as a control) the compound. Relative reduced growth of the cells is scored as sensitivity to the compound. This is a very time consuming procedure and only a maximum of 16 strains can be analyzed in a given day. The analysis of the 384 strains by this technique took 28 full days (this includes the re-runs). Consequently, while the results obtained this way are more reliable, this technique cannot replace the colony size reduction approach for large-scale analysis. Based on the spot test analysis 22 strains were picked as true sensitive and an additional 10 were detected as possible sensitive

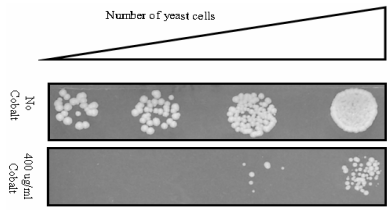


Figure 4: Spot test analysis. The mutant strain shows a significant growth reduction in the presence of 400 ug/ml of cobalt indicating sensitivity to the drug.

(they need further testing).

Manual (by eye) inspection and identification of those growth reductions that are specific from the non specific ones resulted in the identification of 15 strains picked as positives. Based on the spot test one of them was shown not to be a true positive. The manual comparison was done by looking at two plates by eye and estimating the relative growth of an individual colony over the average growth on the plate and comparing it to that on the second plate. This is also a laborious task. It took about one hour of manual activity to compare a pair of plates (an average time of 10 seconds per yeast colony).

The latter method is compared to 30 seconds of calculation and detection of 30 potential target strains using the proposed automated image analysis system. Therefore, the automated system showed 120 fold increase in performance time and more importantly a 200% increase in detection of target strains. Consequently for a complete experiment (16 plates), an average of 240 target sites are identified with the automated system that would have been missed otherwise. This is a very significant improvement for biologists.

A summary of true positive/plate and specificity for the three discussed methods (spot test, by eye and by the proposed automated system) is presented in table 2. This is the result of analyzing growth reduction in total of 384 strains with or without the presence of cobalt. Here, the result acquired by the spot test is considered as ground truth. Although the specificity of the automated image analysis system is slightly lower than the manual method, however the number of true positives and amount of new information gathered by the automated system significantly outweighs this reduced specificity.

We are mainly interested in elucidating the molecular mechanism of action for bioactive chemicals. So far, we have attempted comparing colony growth for yeasts treated with bioactive compound *Echinacea* extract. *Echinacea* has been used as an antifungal compound by the people of North American First Nations [6], but to date its mode of action has remained unknown. Our preliminary results for *Echinacea* show that the automated system has 210% increase in detection of target strains over the manual method, with the possibility that about 70% of the automated hits may be true positives. This is promising and encourages us to further investigate the effect of this bioactive chemical.

Another advantage of the automated system is that by quantification of the growth difference we can prioritize the target candidates for follow up experiments. Multiple positive hits often present the challenge for deciding the priority of each hit. Once the hits are quantified however, those with the best scores get automatic priority.

In addition the automated system has the advantage of reducing the amount of starting material which is not always available. Due to the limitations associated with manual scoring, each experiment is often repeated in multiples. This is not always possible as the amount of the experimental bioactive chemicals is often a limiting factor. Consequently large scale experiments with certain valuable samples for which only small quantities were readily available, were previously thought to lack merit. The automated system developed however, can change this.

Table 2. True positive and specificity results acquired by different methods of monitoring growth reduction of yeast strains, under effect of cobalt.

Method	True positive/plate	Specificity
Spot test	22	100%
By eye	14	99.7%
Automated system	22	97.8%

The precision test was operated on three sets of different experiments, each set including three trials of that particular experiment. This test showed an average of 90.7% absolute agreement of results over multiple runs of a similar experiment.

Altogether preliminary results indicate that the automated system offers significant improvement over the manual scoring of plates. Sensitive additional information can now be extracted from the same experiment, which otherwise would have been missed. This additional information can help biologists to better interpret their results.

4. CONCLUSION

An automated system based on image processing techniques and statistical analyses has been developed for a novel application named functional genomics. The system receives pairs of images, which show arrays of yeast colonies, measures the size of colonies and performs statistical analyses to present meaningful comparison of results to the user. It should be emphasized that without such automatic analysis system, quantitative measurement of colony areas and monitoring drug effect on deletion of various gene pathways is impossible. The conventional method of qualitative analysis by human eye, which is significantly less accurate than the computerized analysis, is a very laborious task even for a small number of colonies.

5. REFERENCES

[1] P. Hietor, M. Boguski, “Functional Genomics: It’s All How You Read It”, Science, vol. 278, pp. 601-602, 1997.

[2] B. J. Andrews et al, “Playing Tag with the Yeast Proteome”, Nature Biotechnology, vol. 21, no. 11, pp. 1297-1299, 2003.

[3] N. Braendle, H. Bischof, H. Lapp, “Robust DNA Microarray Image Analysis”, Machine Vision and Applications, vol. 15, pp. 11-28, 2003.

[4] L. Muresan, B. Heise, E.P. Klement, J. Kybic, “Quantitative Analysis of Microarray Images”, Proceedings of IEEE International Conference on Image Processing (ICIP), pp. 1274-1277, 2005.

[5] M. Katzer, F. Kummert, G. Sagerer, “A Markov Random Field Model of Microarray Gridding”. Proceedings of SAC 2003.

[6] M.L. Smith, P. Gregory, N.F. Bafi-Yeboa, J.T. Arnason, “Inhibition of DNA Polymerization and Antifungal Specificity of Furanocoumarins Present in Traditional Medicines”, Photochem Photobiol, vol. 79, pp. 506-509, 2004.