APPLICATION OF COMPUTERIZED IMAGE PROCESSING IN FUNCTIONAL GENOMICS: PRELIMINARY RESULTS

Negar Memarian1, Javad Alirezaie1,2, Ashkan Golshani3

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

Recent strives in the developing field of functional genomics calls for computerized systems that are capable of providing accurate quantitative data for researchers in biology and genetics. In a current research, 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. Here we report the development of an automated image analysis system, which is able 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. For those experiments that were attempted multiple number of times, a precision test has been designed to monitor the level of harmony between the results of trials. Image processing techniques such as spatial adjustments, segmentation and region growing are utilized in the development of system. Preliminary results show that this automated system offers a significant improvement over the manual scoring of yeast plates.

KEY WORDS
Image analysis, image processing, functional genomics, statistical measurements.

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. Consequently the 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 microarray addressing or gridding [3]. Against the computationally expensive nature of most of those works [4], [5], [6], the present image analysis method produces accurate results in large scale with simplicity and with no need for complex and expensive equipment. In contrast with microarray analysis that has been attempted by numerous groups, our work, which can be defined as growth array analysis 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 sections 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

As the first step of setup, the plate of yeast colonies is placed inside a holder tray at time of image capture (figure 1-a). Although it is tried to keep a fixed zoom and angle among various plate images, still some deviations occur mainly due to the variety of time and location of image captures and also because of human hand error in placing the plates. This results in images that are all similar in the sense that they show an arrangement of yeast colonies but they may be different in terms of size, alignment and zoom. If not considered and compensated, each one of these discrepancies can lead into big errors in the results of analysis. The system performs the following procedures to make all input digital images equal in terms of zoom, angle and size:

- **Detection of marginal corners:**
  Color recognition is used to detect the margins of plate. Three identical red color indicators are positioned on the corners of holder tray (figure 1-a). The system seeks red color components in the image and therefore recognizes the approximate region occupied by plate.

- **Standardization of plate image alignment and zoom:**
  As it is apparent in the example of figure 1-a, the plate image may be tilted. The program aligns the plate image by taking pairs of control points and uses them to infer a spatial transformation. One set of control points, namely the input points, is the centroid coordinates of two of the red indicators explained in previous section. The other set is a pair of points on a precisely horizontal line which are a fixed distant apart. The latter set is called the alignment points. The algorithm calculates both the scale the image should be resized and the angle it should be rotated accordingly. First it resizes the image and then rotates it in a linear conformal manner so that the input points match up with the desired alignment points. The result of these steps is shown in figure 1-b.

- **Cropping background:**
  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. must be removed from the image. It is important to note that pinning of colonies on the plate may vary slightly in different images. For example in one plate the first column of colonies may start d1 distant from the left border of plate and in another plate it starts d2 distant from left border, where d1 and d2 are different in quantity. The same situation may happen for the other three (top, right and bottom) borders of plate. Figures 2-a and 2-b depict an extreme case. Bearing the above constraint in mind, it is imperative to be cautious in the cropping stage. Therefore the plate image is cropped with a generous margin from the boundaries of colony plate. This margin prevents cutoff of useful data (white colonies) in asymmetric cases, such as those shown in figure 2-a and 2-b.

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 white 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 \textit{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 the thresholded and filtered image is depicted in figure 3.

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
Figure 2: Pinning of colony substances may deviate from one plate to another. Distance between the plate border and the first column of colonies from left is very different among plate in (a) and (b).

Figure 3: (a) One original plate image, (b) Final result after thresholding, size filtering and eccentricity (circularity) filtering.

Figure 4: An example of failure of choosing a mask with fixed dimensions. The mask (red square) is selected to contain the largest object in image, which is marked with a cross. When the mask is sledded horizontally and vertically, however, it may fail in fully enclosing the other objects. The arrows show a couple of such instances.

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 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. This 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:

\[
\begin{align*}
  y_{\text{start}} &= \min_y \text{centroid} - \text{round}(\max_y \text{width}/2) \\
  x_{\text{start}} &= \min_x \text{centroid} - \text{round}(\max_x \text{width}/2) \\
  y_{\text{end}} &= \max_y \text{centroid} + \text{round}(\max_y \text{width}/2) \\
  x_{\text{end}} &= \max_x \text{centroid} + \text{round}(\max_x \text{width}/2)
\end{align*}
\]

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 the width of the object with maximum 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 5-a where the down bottom row of colonies does not exist.

adaptive object specific mask to handle this issue. For every object (colony) present in the image, a mask is automatically calculated which includes 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 white pixels that fall inside the adaptive 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. In other words, it is like encapsulating the area of each object in one point. This method is advantageous in two ways: Firstly by applying the adaptive 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. Figure 4 depicts the problems that may arise if instead of the adaptive object specific mask, a fixed size mask is utilized. Such method is not successful because all colonies in the image are not identical in size and shape and therefore, while the fixed mask fully surrounds some of the colonies, it overlaps with or misses some others. Even in an ideally lucky case where all the colonies in one image fit in the fixed mask, positioning and size of colonies of another image (plate) may not necessarily agree with this mask. Consequently user will have to calculate the size of mask for every plate image, which is both against the automatic notion of this system and is also very error prone.
This leads to wrong calculation of \( y_{\text{start}} \) and \( x_{\text{start}} \) 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 5-b, that value is shown as a sequence of stars imposed on original image. As it is obvious, now the star coordinates fall outside the threshold rectangle, meaning calculation of \( x_{\text{start}} \) and \( y_{\text{start}} \) 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 \( x_{\text{start}} \) and \( y_{\text{start}} \) and heading for \( x_{\text{end}} \) and \( y_{\text{end}} \), regions of \( m \times n \) 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).

\[
m = (x_{\text{end}} - x_{\text{start}}) / 24 \tag{5}
\]

\[
n = (y_{\text{end}} - y_{\text{start}}) / 16 \tag{6}
\]

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 (figure 6). 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 region. 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 white 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 final result of processing stage. For every plate the average value of white pixels is calculated from equation 7.

\[
S_{\text{ave}} = 1 / N \sum_{i=1}^{N} S_i \tag{7}
\]

where \( N \) is the total number of white 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 \( S_{\text{ave}} \) from the plate’s 1D ordered area array (equation 8).

\[
\Delta S_i = S_i - S_{\text{ave}} ; i = 1, \ldots, 384 \ (24 \times 16 = 384) \tag{8}
\]
biologists are interested to know that out of three trials how many times a difference is observed and how big this difference is. In other words it is desired to trace the harmony between multiple trials of the same experiment and find out how likely it is that the observed difference may not be due to pure chance. A ranking scheme has been designed to evaluate agreement 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 +1 is assigned to every experiment in which area difference of a pair of corresponding colonies is a positive number, a -1 where it is a negative number and 0 where colony area remains unchanged. The mean score of three trials cannot fully represent the harmony between results of three runs. For example compare combinations 6 and 10 in table 1. The mean score of both of these combinations is 1/3, however case 10 is practically a lot worse than case 6. This is because in case 6 the +1 in one of the trials implies that there is a slight chance that either gene deletion or drug treatment (whichever was applied) may affect the colony size. However in case 10, a -1 indicates that the effect of gene deletion/drug treatment on the colony in one of the trials was absolutely reverse of the other two trials (one size reduction against two size enlargements). Thus the score standard deviation for each combination is also examined. If standard deviation is big, it means that the results are more due to chance that being due to a true pattern. If standard deviation is zero (the ideal case), it means that there is no dispersion between the results.

3. Results

We evaluated the efficiency of our algorithm to detect growth differences in yeast colonies, by comparing colony growth for yeasts treated with Echinacea extract to the untreated ones (used as control). Echinacea has been used as an antifungal compound by the people of North American First Nations, but to date its mode of action has remained unknown [7]. We spotted 384 yeast strains, each containing a specific gene deletion, on a standard omni plate containing appropriate yeast media with or without 9 ug/ml of purified Echinacea extract from Echinacea purpurea plant. This is 1/16 of the plates required for a complete experiment. As expected, treatment of colonies resulted in an overall reduction in yeast growth. This makes it very difficult to manually (by eye) identify those reductions which are specific from the non specific ones. It took about one hour of manual activity to compare the two plates (an average time of 10 seconds per yeast colony). This resulted in the identification of 7 yeast strains which appear to have higher sensitivity to the drug treatment. The 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 result is compared to 30 seconds of calculations and detection of 15 potential target strains using the automated system. Therefore, the automated system showed 120 fold increase in performance time and more importantly a 210% increase in detection of target strains. Consequently for a complete experiment (16 plates), an average of 128 target sites are identified with the automated system that would have been missed otherwise. This is a very significant improvement for biologists.

The sensitivity of the detected yeast strains was then confirmed by growing individual yeast strains in a liquid media, in the presence or absence of the extract, followed by measuring the growth, spectrophotometrically. Our results thus far indicate that of the 10 yeast strains tested in this manner, 7 showed an increase in sensitivity to the extract, indicating that about 70% of the automated hits may be true positives. This suggests that the increase in sensitivity in the automated system also results in a reduced specificity. However the amount of new information gathered by the automated system significantly outweighs this reduced specificity.

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.

<table>
<thead>
<tr>
<th>Combination</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>η</th>
<th>σ</th>
<th>A.R</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>+1</td>
<td>+1</td>
<td>0</td>
<td>2/3</td>
<td>0.816</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-2/3</td>
<td>0.816</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>+1</td>
<td>1/3</td>
<td>0.816</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-1/3</td>
<td>0.816</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>+1</td>
<td>-1</td>
<td>0</td>
<td>1.414</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>-1/3</td>
<td>1.633</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
<td>1/3</td>
<td>1.633</td>
<td>1</td>
</tr>
</tbody>
</table>

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.
The automated system developed however, can change this.

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.

It should be noted that although the developed system analyzes yeast arrays, its modality is different from microarrays. As mentioned earlier, the purpose of our proposed system is to automatically and efficiently analyze the data produced by generating deletions (in large scale) in two target genes, the function of one is known and the other is not. Hence, our developed system measures yeast colony growth. On the other hand microarrays are generally for RNA or protein arrays and they look for RNA levels. Hence, microarrays quantify hybridization signals. Consequently, we cannot compare our results to those obtained by microarrays.

Altogether preliminary results indicate that the automated system offers a significant improvement over the manual scoring of the 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 and Future Work

An automated system based on image processing techniques and statistical analysis has been developed. The system receives pairs of images, which show colony arrays of yeast. The developed automated system measures growth changes in those colonies and performs statistical analyses to present meaningful comparison of results to the user.

It should be emphasized that without such automatic analysis system, accurate measurement of colony areas and monitoring drug effect on different genes is extremely laborious and in case of a huge number of colonies, an impossible task.

Upon requirement of the genomics research, the system can be further expanded to monitor other features of colonies in addition to area. A variety of other statistical features such as reproducibility, reliability, validity, likelihood and etc. can be investigated for assessing the results of multiple trials of the same experiment. In order to do so, a deeper study of statistical parameters and their measurement techniques must be accomplished.

In terms of demonstration, the system can be further developed to visually show the growth or shrink rate of colonies with a range of colors. That way the area change can be mapped to a linear tone color map, where an original colony is matched with a certain color (e.g. orange) and it changes in a fuzzy scheme to stronger shades (e.g. red) or weaker shades (e.g. yellow) according to the degree of size increase or decrease respectively. This offers a very quick and in the same time meaningful visual demonstration of experiment results to the user. Consequently the user will be able to find a qualitative idea of experiment’s overall results even before studying big tables of numbers and only by looking at the colors assigned to colonies in plate.

References