EUSIPCO 2013 1569742669

AUTOMATED IMAGE ANALYSIS AND INFERENCE OF GENE FUNCTION FROM HIGH - CONTENT SCREENS

Priyanka .J. Raja [2],Justin Jacob[1],Byung-Jun Yoon [3], Geofferey Bartholomeusz [1], Arvind Rao[2]

[1]Department of Experimental Therapeutics, [2]Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston TX.
[3] Department of Electrical and Computer Engineering, Texas A&M University, College Station TX.
Email: aruppore@mdanderson.org

ABSTRACT

The study of tumor biology and heterogeneity is of high importance for identifying viable cancer therapeutics. A high content RNAi screen is carried out to identify genes that induce varied tumor morphologies. We present a novel automated pipeline to identify and interpret gene function by extracting morphological features of tumor cell aggregates , in large scale 3D RNAi screens. We use a "bag of words" based clustering approach to distinguish multiple phenotypes. Functional analysis of genes underlying the phenotypic clusters reveals the role of growth and invasion modulators in shaping tumor cell morphology and heterogeneity.

Index Terms— image analysis, image processing, machine learning, high content screening, textural features, affinity propagation clustering, earth movers distance

1. INTRODUCTION

High content throughput screening is an experimental approach applied to identify the role of RNAi in altering cellular phenotype and to study the effects of therapeutics on diseased cell morphology and cellular product expression.

High content screening technology has been used to quantify spatial and temporal variation in cellular events. Automated image-based screening facilitates the detection of perturbagens that alter the morphology or abundance of biological molecules.

A high throughput kinome siRNA screen is carried out in collaboration with the High Throughput Screening Core at the UT MD Anderson Cancer Center, where a set of 880 kinase genes (Dharmacon) were knocked down to study their effects on tumor architecture and hypoxic response induced in tumor spheroids. The screen was carried out in a 96 well plate format and each well corresponds to a certain kinase that has been silenced. Ten such plates were used during the course of the experiment and each plate had 88 wells in use for cell culture. Thus, the screen was carried out within 880 wells with each well associated with 10 imaging fields. The final dataset consists of 8800 images (10 images/well * 880 wells). Each image comprises groups of cell aggregates (referred to as "blobs" in the remainder of this paper).

Our analysis involves detection of these tumor cell blobs to extract textural and morphological features to quantify and discriminate the varied tumor morphologies. These different morphologies are representative of proliferation or growth-inhibition of the tumor cells due to hypoxia induced by gene knockdown.

Clustering of the cellular morphologies represented by feature vectors will provide a key insight into the role of these underlying genes in proliferation or growth inhibition of these tumor cells. Therefore, we have created a pipeline that allows for the inference of gene function and their potential effects on tumor cells via analysis of image-derived phenotypes. The overall workflow in the proposed pipeline is presented in Figure1.

2. METHODS AND RESULTS

2.1. Image Preprocessing

In order to detect and retain data pertaining to the tumor cells in the acquired image dataset, field-specific masks are generated. An image processing workflow is developed using the Pipeline Pilot software from Accelrys Inc. The tool allows us to

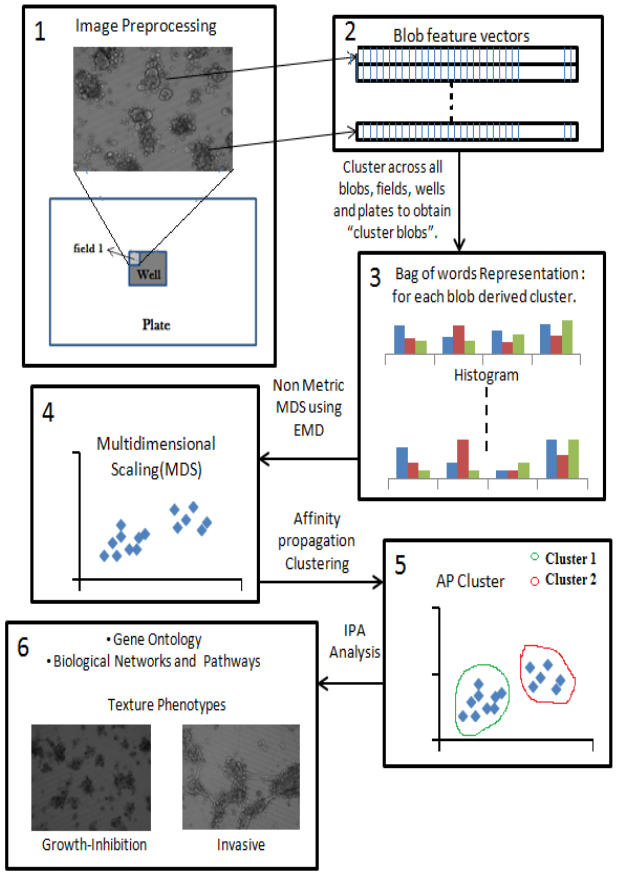

Figure 1: *The workflow adopted for our analysis.*

automate the common steps involved in generating the masks for the 8800 images involved in the analysis.

For cell segmentation, we use a morphological gradient, suitable erosion and dilation operations, followed by binarization. Individual blobs are detected using neighbourhood connectivity followed by a filtering operation to remove spurious blobs.

Every blob is accessed by its assigned label, cropped from the image using a bounding box and processed individually. Morphological and textural features are now computed on these blobs to define a feature vector of texture patterns and structural geometry.

2.2. Feature Extraction

Second order statistics or co-occurrence matrix features have been used extensively for texture extraction and classification [1][2]. Tumor architecture and its varied morphologies can be assessed and analyzed by computing textural and morphological features associated with the invasive and the non-invasive (growth inhibitory) phenotype. Gray level co-occurrence matrices are employed to compute textural features on individual blobs with symmetric offsets at angles of 0, 45, 90 and 135 degrees. Eighteen textural descriptors (including contrast, energy, entropy and correlation) are computed and tabulated for each individual blob within a field.

In addition several morphological features (such as area, diameter, perimeter, solidity and form factor) are computed for these blobs (tumor spheroids). Consequently, each blob can now be represented with a 29(18+11) dimensional feature vector. This procedure is iterated over all the blobs identified across the acquired image data set to obtain a feature matrix (rows representing blobs across plates and columns representing features).The features are normalized to zero mean and unit variance before further processing (as shown in block 2 in Figure 1).

2.3. Adjusting for contrast

Since texture features are sensitive to grayscale variation, we employ a contrast-adjustment step to facilitate comparison across images. This step is also intended to aid in data interpretability and visualization. The images of the blobs are subjected to contrast-adjustment and the co-occurrence matrices are computed on these contrast-adjusted gray level images using symmetric offsets. Such adjustment in contrast may lead to detecting subtle changes in textures and therefore assist in categorization of tumor-associated phenotype.

2.4. Normalized histogram representation of wells

The primary goal of the analysis is to discover functional phenotypes from each well in such high content data. Following a bag-of-words [3][4] paradigm, we represent each well as a histogram over clusters, where a cluster label is a word and the blobs within a well is a bag of words representing cluster memberships (shown in block 3 of Figure 1). These clusters are derived via clustering image features of blobs within fields. The bag of words approach is illustrated in Figure - (2).

Using the 29 features extracted for each blob, we use a Gaussian Mixture Model algorithm for clustering all the blobs (across all the images). The Bayesian Information criterion (BIC) [5]based on model

clustering is used to determine the optimal number of clusters. The data from the contrast-adjusted and unadjusted analyses are used as inputs to the GMM algorithm.

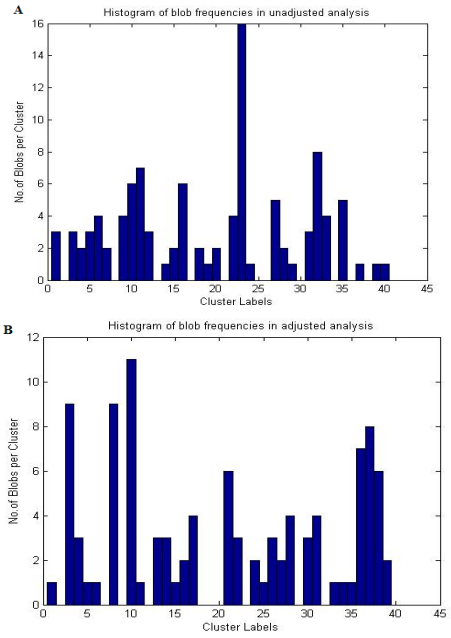

Figure (2A and 2B): *Normalized histograms for well 880 containing 40 cluster bins on the horizontal axis and the number of blobs per cluster bin in the vertical axis for the unadjusted and adjusted cases respectively.*

The algorithm sweeps between the cluster ranges of 10 to 80 for multivariate data models. The clusters are allowed to be flexible in volume and orientation. The Bayesian Information criterion (BIC) reveals that 40 is an optimal cluster size. Thus, the blobs are grouped into 40 distinct clusters by employing the kmeans clustering algorithm. Each well corresponds to a cluster membership vector –with each element representing a cluster label corresponding to a specific blob. Figure-2 represents the histograms generated for well 880 for both the contrast-adjusted and unadjusted cases.

2.5. Non-metric Multidimensional Scaling

Non-metric Multidimensional Scaling (NMDS) is a dimensionality-reduction technique that takes as input the set of pairwise distances between data points in higher dimensional space. Since the data that we choose to embed are in the form of normalized probability distributions, we choose the Earth Movers Distance as a metric of distance between data points. This is a well-known metric and has commonly been used for its applications in computer vision problems like image retrieval [6][7][8].

The Earth Movers Distance (EMD) between positive, one dimensional and equally weighted histograms can simply be calculated using cumulative distribution functions (CDF) as follows[9].

$$EMD = \sum_{i=1}^{N} |CDF(X) - CDF(Y)|$$

Where, N is the total number of histograms or cluster member ship vectors and X and Y are the cluster membership vectors.

The chi-square distance metric is also a viable metric on probability distributions[10] .It is also employed to compute the pairwise distance matrix between histograms and compared with EMD for lesser stress in embedding .

We compare the Kruskal stress values [9] of the points in the lower dimensional embedded space under both metrics (EMD and chi-squared) for the contrast-adjusted and unadjusted cases. We choose the distance metric that provides the least stress (Table 1).

Based on Table 1, we use the EMD distance metric for three-dimensional embedding, since this has the lowest stress and also yields visually interpretable embedding. Figure-3 depicts a plot obtained after multidimensional scaling with EMD as the chosen parameter to compute the dissimilarity matrix between the histogram distributions.

2.6. Clustering

Affinity Propagation (AP) is a popular clustering algorithm [11] that takes as input the real valued similarities between the data points to gauge the exemplars (similar to cluster representatives). This algorithm has the advantage over other k-center clustering algorithms in that it requires no predefined number of clusters and the outcome is not initialization-dependent. The affinity propagation algorithm assumes that every point in the dataset could be an exemplar to begin with, and by passing real-valued messages between points certain preferences or exemplars are eventually generated along with their clusters. This algorithm has also been shown to produce clusters with lesser error rate in comparison with other unsupervised clustering algorithms[12].

The scaled lower dimensional data set obtained from non-metric MDS (NMDS) is fed as the input to the AP clustering algorithm. The clustering algorithm

	Stress (With Contrast - Adjustment)	Stress (Without Contrast - Adjustment)
Three dimensional Embedding		
Earth Movers Distance	0.0664	0.0649
Chi-Squared Distance	0.1657	0.1412
Two dimensional Embedding		
Earth Movers Distance	0.0923	0.1066
Chi-Squared Distance	0.2380	0.1886

Table(1):Kruskal stress values for both distance metrics in two dimensional and three dimensional embedding are listed. Lower is better.

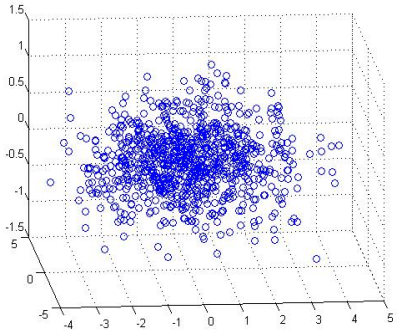

Figure (3): Plot obtained after non-metric multidimensional scaling with EMD as distance metric from the contrast-adjusted analysis.

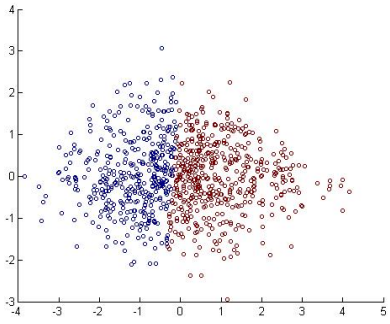

Figure (4): The plot obtained from the AP clustering algorithm with two clusters (the optimal cluster number).

is applied to the data for both contrast-adjusted and unadjusted modes of analysis. The resulting clusters contain points representing 880 different wells in the embedded space.

Figure- 4 shows a plot obtained from the affinity propagation algorithm separating the data into two optimal clusters for both the analyses. The derived clusters have 425 and 455 wells (genes) respectively.

2.7. IPA Analysis

The genes underlying the two clusters are analyzed using the Ingenuity Pathway Analysis (IPA) software[13].The IPA tool aids in the examination of the relevant canonical pathways, gene ontology, and functional enrichment underlying an input set of genes. This section contains a brief summary of the top signaling pathways and biological functions as reported by the IPA software for the two clusters obtained from the affinity-propagation clustering step.

The genes present in cluster 1 have functions in the networks that contribute to cancer and nervous disorders and growth. The biological functions represented by cluster 1 are tumorigenesis of benign and malignant tumor, immune cell trafficking, organism survival, hematological system development and connective tissue development. Immune cell trafficking and hematological system development involves genes that facilitate immune cell movement and migration, proliferation, development and activation. A large number of genes also contribute to the growth of tumor and formation of solid tumor. Several genes in cluster 1 are associated with proliferation of tumor cell lines and organism death. In totality, cluster 1 may be considered to represent a "growth-inhibition" phenotype as it contains genes that when suppressed or altered may cause inhibition of tumor growth.

On the other hand, cluster 2 predominantly contains genes that are related to cancer and developmental disorders. The genes in this cluster have relevance to tissue development, tissue morphology and hematological system development. The genes found in cluster 2 in the aforementioned categories are responsible for the quantity of immune cells present, positive selection of T-lymphocyte, disorganization of blood vessels, the morphology of endothelial cells and hypertrophy of normal and endothelial cells. There are also significant number of genes that contribute to the growth and proliferation of

normal cells, homeostasis and cell death, organization of blood vessels, proliferation of endothelial cells and fibroblasts, formation of focal adhesions, formation of cellular adhesions that are vital for cell viability and function. Therefore, broadly looking at the biological functions that are represented in this cluster we may consider cluster 2 to contain genes which when suppressed or altered may lead to invasive cancer cell formation, perhaps representing an "invasive phenotype".

For genes underlying clusters based on the contrast-unadjusted case, we observe similar functions as the clusters described above. However, those results are not presented due to ambiguity in ontological basis within those clusters.

3. CONCLUSION

In this work, we present an analytic workflow capable of processing high content images from large-scale RNAi screens. The pipeline is capable of processing blob-level image data, thereby accounting for the heterogeneity in the image fields. Using a "bag-of-words" representation, we use a combination of image-analysis, clustering and classification of the blobs into distinct cell phenotypes.

There were two predominant cell phenotypes that are recognized from the dataset - "cell death" and an "invasive" phenotype. The biological functions of these genes associate with cell phenotypes related to tumor invasion and growth- inhibition of tumor cells.

In addition, through comparison of the two modes of analyses we have shown that contrast-adjustment is a useful preprocessing step in order to obtain physiologically-relevant phenotypes.

In future work, we aim to explore other image features that might represent biologically-relevant phenotypes and to perform a systematic evaluation of multiple clustering algorithms for phenotypic classification. Furthermore, we aim to integrate hits from such HTS RNAi screens with other in-vivo data to identify gene hits that might represent viable drug targets.

4. REFERENCES

[1] Z Bailing and P Tuan, "Phenotype Recognition with Combined Features and Random Subspace Classifier Ensemble," *BMC Bioinformatics*, vol. 12:128, 2011.

[2] N. Harder, R. Eils, and K. Rohr, "Automated Classification of Mitotic Phenotypes of Human Cells Using Fluorescent Proteins," *Elsevier*, vol. 85, pp. 539–554, 2008.

[3] E. Ozdemir, T. Bilgisayar Muhendisligi Bolumu, Bilkent Univ., Ankara, and C. Sokmensuer, C. ; Gunduz-Demir, "Histopathological image classification with the bag of words model," *Signal Processing and Communications Applications (SIU), 2011 IEEE 19th Conference on*, vol. 19, pp. 634 – 637, 2012.

[4] J. Y. Shuiwang Ji, Ying-Xin Li, Zhi-Hua Zhou, Sudhir Kumar, "A bag-of-words approach for Drosophila gene expression pattern annotation," *BMC Bioinformatics*, vol. 10:119, 2009.

[5] J Freidman ,T Hastie, R Tibshirani, *Elements of Statistical Learning*. 2001.

[6] F. Wang and L. J. Guibas, "Supervised Earth Mover ' s Distance Learning and its Computer Vision Applications," pp. 1–14.

[7] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover ' s Distance as a Metric for Image Retrieval," pp. 1–20.

[8] G. Herman, "On the Earth Mover'S Distance as a Histogram Similarity Metric for Image Retrieval," *2005 IEEE International Conference on Multimedia and Expo*, pp. 686–689, 2005.

[9] P. Dollar, "Piotr's Image & Video Matlab Toolbox."
Available: http://vision.ucsd.edu/~pdollar/toolbox/.

[10] M. Kok and B. N. Chatterji, "Comparison of Similarity Metrics for Texture Image," *IEEE*, 2003.

[11] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Pyschometrika*, vol. Volume 29, no. Issue 1,, p. pp 1–27.

[12] B. J. Frey and D. Dueck, "Clustering by passing messages between data points.," *Science (New York, N.Y.)*, vol. 315, no. 5814, pp. 972–6, Feb. 2007.

[13] "Ingenuity Pathway Analysis." .
Available: www.ingenuity.com/products/ipa.