A Framework for Molecular Co-Expression Pattern Analysis in Multi-Channel Toponome Fluorescence Images

Ahmad Humayun[1], Shan-e-Ahmed Raza[1], Christine Waddington[2], Sylvie Abouna[3], Michael Khan[3], Nasir M. Rajpoot[1,]*

[1] Department of Computer Science, University of Warwick, UK
[2] Molecular Organisation & Assembly in Cells (MOAC), University of Warwick, UK
[3] School of Life Sciences, University of Warwick, UK

[*] Corresponding author: Nasir Rajpoot <nasir@dcs.warwick.ac.uk>

Abstract **—** Bioimage computing is rapidly emerging as an important area in image based systems biology with an emphasis on spatiotemporal localization of subcellular bio-molecules, most importantly proteins. A key problem in this domain is analysis of protein co-localization or co-expression of protein molecules. Imaging techniques, such as the Toponome Imaging System (TIS) [1], with the ability to localize several different proteins in the same tissue specimen are only becoming available recently. Traditional co-localization studies and some of the modern co-expression studies have serious limitations when analyzing this kind of data. Here we present a robust framework for the analysis of molecular co-expression patterns (MCEPs) in TIS image data.

Index Terms **—** Protein co-localization, molecular co-expression, multi-fluorescence imaging, bioimage computing, clustering analysis, MELC, TIS

I. INTRODUCTION

In recent years, bioimage computing is emerging as a cornerstone of hypothesis-driven research in life sciences with an emphasis on spatiotemporal localization [2]. A major focus in the post-genomic era is on analyzing subcellular protein patterns enabled by the knowledge of spatiotemporal distribution of key proteins expressed in a given cell type [3]. As proximity of proteins located within similar compartments of a cell provides a powerful surrogate for functional complexes, functional studies involve proteins which appear to be key players in multiple cancer specific complexes. Imaging techniques with the ability to localize several different proteins in the same tissue specimen such as the Toponome Imaging System (TIS) [1], MALDI imaging [4], Raman spectroscopy [5], or multi-spectral imaging methods [6] are only becoming available recently. Of these, TIS is an automated fluorescence technique shown to have the ability to co-map hundreds of different proteins or other TAG-recognizable bio-molecules on a single tissue section. This results in a multi-tag fluorescent image stack with corresponding phase contrast images before and after incubation of the corresponding antibodies. For each antibody tag, four images of the tissue section are obtained, resulting in a stack of images for selected visual fields in a tissue specimen. These images take the form of an initial phase image followed by one under ultraviolet light for fluorescence, then a waiting time occurs for the tag attachment and specimen rinsing, before another phase image is taken with its matching fluorescent image. After these 4 images are taken, a bleaching cycle occurs before the process begins for the next tag. Before any significant conclusions can be drawn about co-localization of proteins, the image stack must be accurately aligned or registered. The phase images can be used for alignment purposes since (a) these are not expected to vary throughout a TIS run, and (b) each phase image is taken a fraction of a second before its matching fluorescence image and so it is assumed that the phase/fluorescence image pair for a particular antibody is perfectly aligned.

Traditional protein co-localization studies involve three key steps: staining the tissue specimen with two or three different dyes (such as DAPI for nuclei and a green or red fluorescent proteins, also known as GFP or RFP, binding to specific antigens), taking fluorescence images of each dye with different wavelengths of the incident laser beam, and treating the individual fluorescence channels as one of the Red, Green, or Blue channels to construct a color image. In this color image, pixel locations with yellow color, for instance, indicates the simultaneous presence of bio-molecules corresponding to red and green channels. A major shortcoming of this method of studying protein co-localization is that it does not take into account the varying levels of expression of different proteins. Another disadvantage is that often this kind of simplistic analysis is limited to the study of co-localization of 2-3 proteins at the same time.

Multi-tag bioimaging methods, such as the ones mentioned above, allow us to study combinatorial protein patterns in specific types of tissues and to characterize and differentiate between different kinds of cells within the same specimen. However, recent developments in machine learning and computer vision are yet to impart their influence on the data analysis pipeline of standard software used with most of these imaging systems. For instance, much of the reported analysis of TIS image data such as [7] is based on binarization using manually selected thresholds.

In this paper, we describe a framework to pre-process and analyze multi-channel fluorescence microscopy images obtained using TIS [1]. The framework is composed of three key components: pre-processing for robust alignment of TIS images using a modified version of the recently proposed RAMTaB algorithm [8], segmentation of nuclei, and clustering analysis of protein patterns without binarizing individual fluorescence images thus allowing for analysis and discovery of combinatorial patterns of a whole range of protein expression levels or molecular co-expression patterns (MCEPs). The proposed framework is generic in nature and should be applicable to other multi-channel imaging methods such as MALDI [4], Raman spectroscopy [5], or multi-spectral imaging methods [6].

II. MATERIALS & METHODS

The image data obtained in this study was acquired using a TIS microscope installed at Warwick. The human colon tissues were surgically removed from cancerous patients. For each cancerous tumor, distal normal section was also taken from the same patient's colon. Patient consent and appropriate ethics approval were obtained to remove and handle these tissues for research. A library of 26 antibody tags, some of which are known tumor markers and others cancer stem cell markers, were used based on a previous study [7].

The molecular co-expression analysis framework proposed in this paper is based on three major components: aligning the images corresponding to individual antibodies in a TIS stack with each other, segmentation of nuclei and surrounding pixels to consider only molecular expression in the cellular areas, and analysis & visualization of molecular patterns using a clustering method.

A. Image Registration

We employ the RAMTaB algorithm [8] for aligning TIS images. However, a limitation of that algorithm is that it does not explicitly deal with phase images being out of focus. Due to the autofocus feature of TIS microscope, it starts with the focus that was previously calculated (because the lens is already in this position), if there is variation between focusing planes of the visual fields this could cause a drift of focal plane as the run progresses. This can either show itself in the phase images as an out of focus image or blurry image, or the sample is focused at a different level causing differences in locations of the cell walls. However, where the plane of focus has changed, some cells appear smaller and some larger. This sort of focal error is difficult for the RAMTaB registration algorithm to align, as it is not a simple shift or shear combination, and so needs to be avoided. Therefore the sample needs to be very flat on the cover slip and the sample to be as thin as possible. Choosing visual fields with similar z coordinates for the plane of focus may be effective in reducing focusing errors.

Experiments have shown that such images can be aligned by using recent image de-blurring techniques which use a normalized sparsity measure [9]. An example of an out of focus phase image is shown in Figure 1(b) – a corresponding reference phase image with the right focus is shown in 1(a). Figure 1(c) shows de-blurred version of 1(b) using a standard blind de-convolution technique [10], and 1(d) shows de-blurred image with the help of blind de-convolution using normalized sparsity measure. The blind de-convolution algorithm minimizes the scale-invariant cost function, l_1/l_2 *norm ratio* to estimate the kernel blur. The kernel is estimated in a multi-scale approach from coarse to fine image resolutions. Once the kernel is estimated, the image is de-blurred using the de-blurring method of Krishnan & Fergus [11]. Images in 1(a) and 1(d) can be aligned using the RAMTaB algorithm [8].

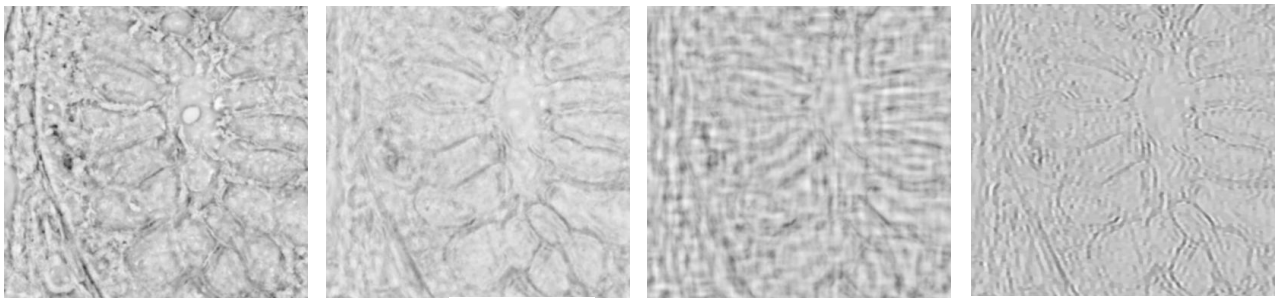
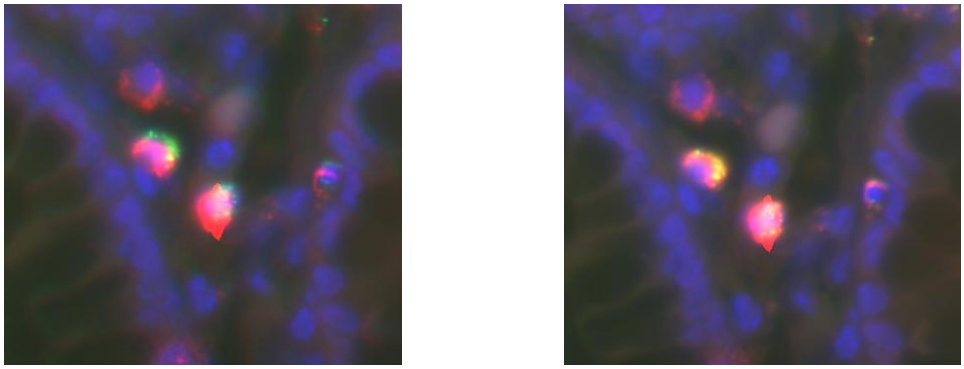


Figure 1: Phase image (a) shows a reference image with normal focus, phase image (b) shows a floating image with out of focus capture during the imaging process. Image in (c) shows a de-blurred version of (b) using the standard blind MATLAB® deconvolution of Holmes et al. [10], while (d) shows the result of blind de-convolution using normalized sparsity measure [9]. All phase images are shown with their complements here for the clarity of display.



(a) (b)

Figure 2: Pseudo-color images generated by taking fluorescence images for three antibodies Muc1, CD57, and DAPI as red, green, and blue channels: (a) before registration (b) after registration. The co-localization of Muc1 and CD57 in compartments of one of the cells is much more obvious as a yellow in (b).

B. Nuclei Segmentation

Once the image data from a TIS multi-fluorescent image stack have been aligned, we normalize the intensity values in each of the aligned TIS images to the range [0,1]. The aligned DAPI channel is then segmented in order to extract pixel locations corresponding to the cell nuclei and their immediate neighborhood only. This step ensures that only molecular patterns localized to cell nuclei and cytoplasm are considered. This removes signal from stroma and lumen in the case of colon, for example, which may add noise to the process of pattern analysis. This segmentation of pixels into nuclei and their immediate neighborhood is achieved using Gaussian mixture modeling (GMM) over the normalized intensity values of the fluorescent channel images and the Bayesian information criterion (BIC) for model selection [12] [13].

C. Clustering Analysis for Extracting Molecular Co-Expression Patterns (MCEPs)

After registration and segmentation of the stack of TIS images, we obtain protein expression vectors $\boldsymbol{I}_i$ of length *N* at each segmented pixel location *i*. Since each dimension in this vector encodes the difference in expression levels after adding a particular anti-body, the vector can be used to cluster pixels based on responses to *N* anti-bodies. Our goal is to study how well unsupervised clustering can reveal the differences in inter and intra- tissue anti-body responses. Our approach is based on a simple hierarchical clustering method, which is a bottom-up clustering method [14]. It starts with each pixel as a cluster and iteratively merges these clusters to form bigger ones. Existing clusters are merged to create new ones, reducing the number of clusters by 1 at each of the iterations until there is only one cluster containing all the data points. For instance, if initially there are n pixels (and an equal number of clusters), the first iteration merges two pixels to give $n-1$ clusters. This process can be better represented as a dendrogram tree structure, where cutting across the tree at level k would give $n-k$ clusters[1]. Like many other unsupervised methods, hierarchical clustering can also be provided with the number of clusters desired. We aim for $C = 20$ clusters, which we call molecular co-expression patterns (MCEPs). These C clusters are produced by cutting the tree at level $\hat{k} = n - C$. As mentioned above, two clusters are merged at each iteration. The criterion we employ to select these two clusters aims to minimize the increase in the variance of clusters [15]. Mathematically, at each tree level k, we have clusters $\mathrm{S}_j = \left\{\boldsymbol{I}_1, \ldots, \boldsymbol{I}_{n_j}\right\}$ where $n_j = \left|\mathrm{S}_j\right|$ and $j \in \{1, \ldots N-k\}$. Here, we can define the within-class variance of cluster S_j as follows:

$$\sigma(\mathrm{S}_j) = \sum_{m=1}^{n_j} (\boldsymbol{I}_m - \bar{\mathrm{S}}_j)(\boldsymbol{I}_m - \bar{\mathrm{S}}_j)^T \qquad (1)$$

where $\bar{\mathrm{S}}_j$ is the centroid vector for cluster S_j. To make clusters for level $k+1$, we seek to combine vectors in $\mathrm{S}_{\hat{u}}$ and $\mathrm{S}_{\hat{v}}$ such that:

$$\hat{u}, \hat{v} = \underset{u,v \in \{1,\ldots N-k\}}{\operatorname{argmin}} \left[\sigma(\mathrm{S}_u \cup \mathrm{S}_v) - \sigma(\mathrm{S}_u) - \sigma(\mathrm{S}_v)\right] \qquad (2)$$

or

[1] The algorithm starts at level $k = 0$, where there are n clusters. *Cutting the tree* at level k means truncating the tree after level k.

$$\hat{u}, \hat{v} = \underset{u,v \in \{1,\ldots N-k\}}{\operatorname{argmin}} n_u n_v \frac{(\|\bar{S}_u - \bar{S}_v\|_2)^2}{n_u + n_v} \quad (3)$$

where $\|\cdot\|_2$ is the Euclidean norm. This step will result in a new cluster $(S_{\hat{u}} \cup S_{\hat{v}})$ formed by merging $S_{\hat{u}}$ and $S_{\hat{v}}$, hence reducing the number of clusters by 1.

III. Experimental Results

Using the above clustering method, we pick the top 20 clusters localized to nuclei and their vicinities. Each of the centroids of these clusters is given a unique pseudo-color. We employ the MATLAB® `jet` colormap, a variation of the `hsv` colormap, which goes from dark blue (for the first MCEP) to dark red (for the last MCEP) passing through the colors cyan, yellow, and orange in between. A pseudo-color overlay of MCEPs on corresponding phase contrast images using the centroids of top 20 clusters for two human colon tissue specimens (cancer on the *Left* and normal tissue on the *Right*) is shown in Figure 3. It can be seen from this display of molecular co-expression patterns that there is a clear difference in tissue morphology and molecular expression at sub-cellular level in normal and cancer specimens. This approach is fundamentally different to the standard TIS visualization approaches using thresholds and random colors [1], [7]. Furthermore, by localizing the pattern analysis to DAPI-positive pixels and their surroundings we are able to filter out any noise due to non-cellular pixel locations such as lumen or stroma.

IV. Conclusions

In this paper, we have presented a robust framework for the analysis of molecular co-expression patterns in multi-tag fluorescent image stacks generated by the TIS microscope. The framework should be applicable to other multi-tag imaging systems and we hope that it will serve as a critical building block for further analysis of TIS stacks in cancer studies.

Acknowledgements

This work was partly funded by the Warwick Institute of Advenced Studies (IAS) and the HDC. The authors are grateful to W. Schubert, the inventor of TIS, who helped us establish a TIS machine at Warwick, and members of his team at ToposNomos and the University of Magdeburg, especially A. Krusche and R. Hillert. Special thanks go to Sayan Bhattacharya for contributions to design of the antibody library.

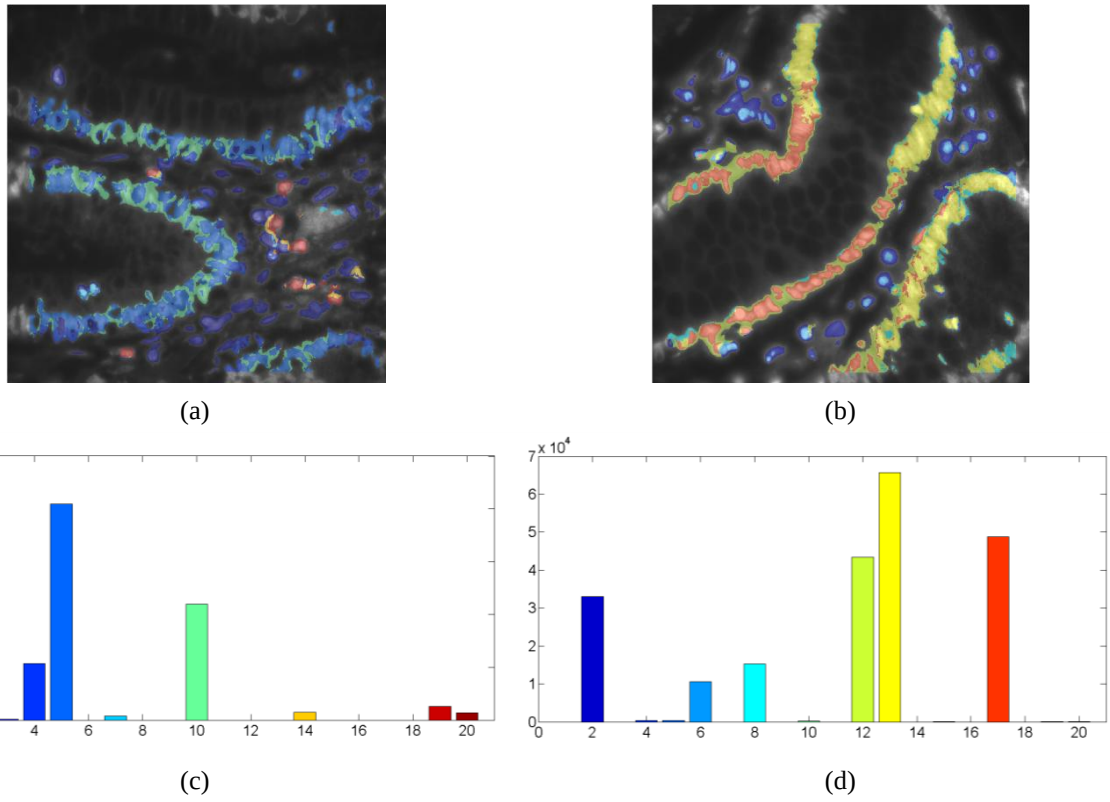


Figure 3: Pseudo-color overlay of molecular co-expression patterns (MCEPs) on corresponding phase contrast images of two human colon tissue specimens (cancer *a* and normal *b*) using the centroids of top 20 clusters. The bar charts in *c* and *d* show a histogram of the top 20 MCEPs found in the specimens.

References

[1] W. Schubert et al., "Analyzing proteome topology and function by automated multidimensional fluorescence microscopy," *Nature Biotechnology*, 2006, vol. 24, pp. 1270-1278.

[2] S. Megason and S. Fraser, "Imaging in Systems Biology," *Cell*, 2007, vol. 130, pp. 784-795.

[3] R.F. Murphy, "Putting proteins on the map," *Nature Biotechnology*, October 2006, vol. 24, no. 10, pp. 1223-1224.

[4] D. Cornett, M. Reyzer, P. Chaurand, and R. Caprioli, "Maldi imaging mass spectrometry: molecular snapshots of biochemical systems," *Nature Methods*, 2007, vol. 4, pp. 828-833.

[5] H. van Manen, Y. Kraan, D. Roos, and C. Otto, "Single-cell raman and fluorescence microscopy reveal the association of lipid bodies with phagosomes in leukocytes," *PNAS*, July 2005, vol. 102, no. 29, pp. 10159–64.

[6] E. Barash, S. Dinn, C. Sevinsky, and F. Ginty, "Multiplexed analysis of proteins in tissue using multispectral fluorescence imaging," *IEEE Trans Med Imaging*, 2010, vol. 29, no. 8, pp. 1457-1462.

[7] S. Bhattacharya et al., "Toponome imaging system: In situ protein network mapping in normal and cancerous colon from the same patient reveals more than five-thousand cancer specific protein clusters and their subcellular annotation by using a three symbol code," *J. Proteome Res*, 2010, vol. 9, no. 12, pp. 6112-6125.

[8] S. Raza et al., "RAMTaB: Robust Alignment of Multi-Tag Bioimages," *submitted to BMC Biophysics*, 2011.

[9] D. Krishnan, T. Tay, and R. Fergus., "Blind Deconvolution using a Normalized Sparsity Measure," in *IEEE Computer Vision and Pattern Recognition (CVPR) 2011*, 2011, Colorado, pp. 233-240.

[10] T. J. Holmes et al., "Light Microscopic Images Reconstructed by Maximum Likelihood," in *Handbook of Biological Confocal Microscopy*, J. B. Pawley, Ed. New York: Plenum Press, 1995, pp. 389-402.

[11] D. Krishnan and R. Fergus., "Fast Image Deconvolution using Hyper-Laplacian Priors," in *Neural Information Processing Systems*, 2009, Vancouver.

[12] C. Farley and A.E. Raftery, "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis," *The Computer Journal*, 1998, vol. 41, no. 8, pp. 578-588.

[13] N.Rajpoot and M. Arif, "Unsupervised shape clustering using diffusion maps," *Annals of the BMVA*, 2008, no. 5, pp. 1-17.

[14] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM computing surveys (CSUR)*, 1999, vol. 31, no. 3, pp. 264-323.

[15] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American statistical association*, 1963, vol. 58, no. 301, pp. 236-244.