[Preprint submitted to WCSB 2011]

TOWARDS PROTEIN NETWORK ANALYSIS USING TIS IMAGING AND EXPLORATORY DATA ANALYSIS

Daniel Langenkämper[1], Jan Kölling[1], Ahmad Humayun[2], Sylvie Abouna[3], David Epstein[4], Michael Khan[3], Nasir M. Rajpoot[2], Tim W. Nattkemper[1]

[1]Biodata Mining Group, Faculty of Technology, Bielefeld University, Germany
[2]Computational Biology and Bioimaging (COMBI) Group, Department of Computer Science, University of Warwick, Coventry, United Kingdom
[3]Department of Life Sciences, University of Warwick, Coventry, United Kingdom
[4]Department of Mathematics, University of Warwick, Coventry, United Kingdom

ABSTRACT

Identification, analysis and visualization of functional molecular networks are key objectives in systems biology and the logical extension of existing molecular profiling techniques. Here we used TIS (toponome imaging system) imaging to visualize co-location of proteins in tissue samples, thereby integrating two distinct information domains, morphology and molecular interaction. Using a library of 13 selected dye-conjugated antibodies, TIS recorded a stack of 13 fluorescence images, each showing the same visual field, with high fluorescence values indicating the presence of the corresponding bio-molecule or protein. We show first results obtained using machine learning approaches that allow the identification and spatial analysis of co-location patterns without manual thresholding. The authors believe that TIS imaging in combination with advanced visual data mining methods can contribute substantially to addressing several outstanding issues in systems biology where molecular co-location is involved.

1. INTRODUCTION

To understand cellular biology on a systems level, important relationships between intra-cellular molecular components must be understood not only at a functional level but localized in the spatial domain as well [1, 2]. As a consequence, new bioimaging techniques have been proposed recently to visualize co-location or interaction of several molecular components simultaneously. These include MALDI imaging [3], Raman microscopy [4], and TIS [5] (also called MELC) all of which generate multivariate bioimages, that pose substantial new challenges for computer scientists from the domains of data mining, bioinformatics, image processing and visualization [6,7]. TIS (Toponome Imaging System) uses a library of *M* dye-conjugated antibody markers to localize different proteins in a stack of images from the same visual field using an iterative protocol running in *M* cycles of labeling and soft bleaching. In [8] we have shown how TIS imaging can be applied in cancer research for *in situ* protein network mapping. However, the localization and extraction of relevant co-expression information and its integration in modeling and pathway analysis needs new algorithmic approaches. The standard way to analyze TIS images is to apply a threshold to each image of the stack so that each image is reduced to binary values. Although this reduction step is straightforward, a non-threshold based approach is preferred because thresholding images is bias-prone, subjective and time-consuming and relies entirely on user expertise, reducing the reproducibility of the final results.

In this paper we show how to explore TIS images for a first time on the grey value level using unsupervised machine learning methods. The problem in analyzing TIS on the grey value level is that the full data set of M grey value images, each one with ~1000^2 pixels and a minimum grey value precision of 2^8 represents a high volume of complex data. This cannot be analyzed ad-hoc, since a visual inspection of all images, or RGB visualization of three images selected from the entire set, is not possible–it would overburden the visual memory and the cognitive skills of any user. One method of data reduction can be applied in the spatial domain with the detection of cells as proposed in [9]. And for small M (<6) a combination of cell segmentation and scatter plots can be used as proposed in [10]. But TIS images are usually recorded for much larger antibody libraries (M>10) so alternative approaches are definitely needed. In this paper we show how a combination of image processing, dimension reduction, and principles from scientific/information visualization can be applied to render molecular co-expression maps (MCMs) for TIS data and we will discuss first results obtained for two samples from a cancer study.

2. METHODS

2.1. TIS Imaging

TIS imaging was applied to two tissue samples from the same colon cancer patient. One tissue sample was selected from a cancerous tumor, while the other sample was selected from healthy colon tissue obtained at the same time. An antibody library of 22 tags (see [6] for details) was applied to record 22 fluorescence images from two manually selected visual fields in each sample, leading to four TIS data sets. For each of the stacks, we automatically align all the images using their corres-

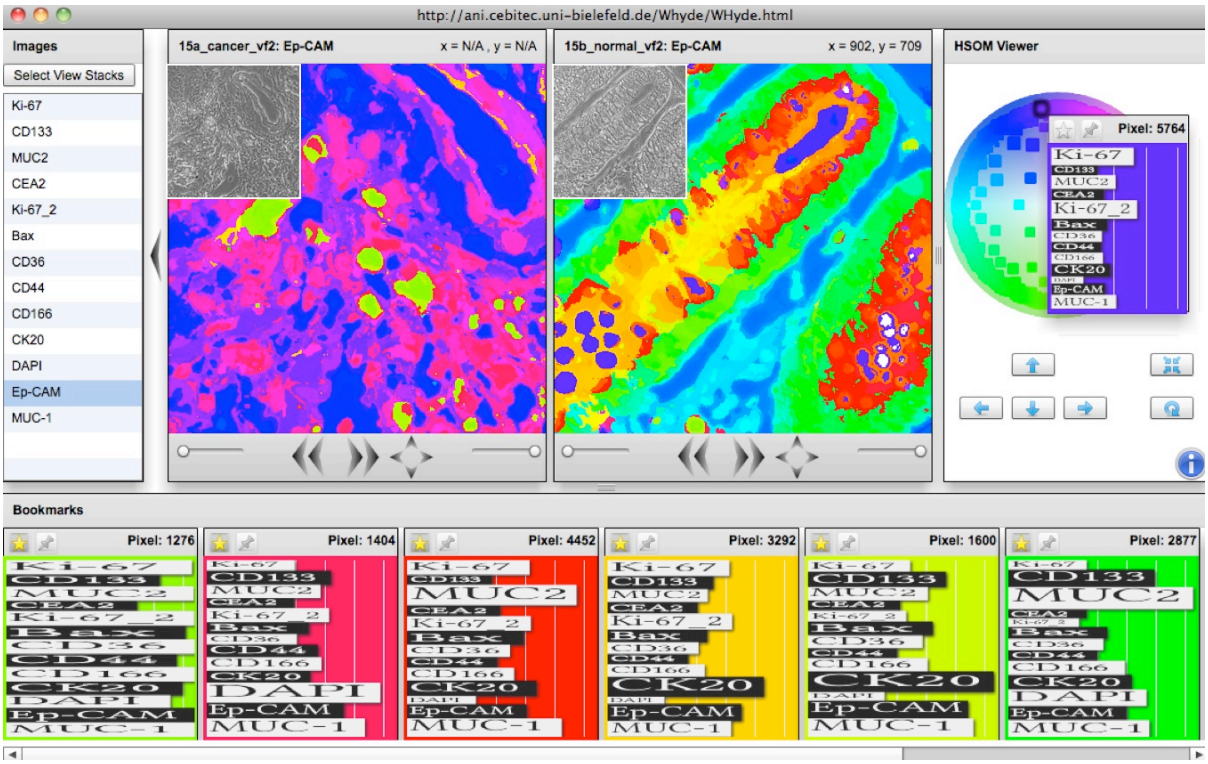

Figure 1: A pseudo color molecular co-expression map (MCM) obtained with vector quantization clustering and dimensional reduction. Similar colors at two pixels, possibly far from each other in terms of pixel distance, represent similar co-location patterns in these two pixels. That is, the same proteins are expressed in much the same way at the two pixels. This may indicate similar biological functions at the two pixels. The whole framework supports the process of visual data mining, i. e. the user explores the high dimensional image data in an interactive dynamic visualization that allows two kinds of zoom-in: First, a geometrical zoom allows the user to change the scale of the image display, resolving morphological details on different scales. Second, the user can zoom into the M-dimensional space by selecting individual clusters on the right, highlighting and bookmarking them (see bottom row). The bar glyphs represent the components in the cluster prototype, i. e. the co-location pattern. One interesting first observation is that on the left (tumor tissue) the co-location pattern heterogeneity seems to be much lower than on the right (healthy tissue). In addition, co-location clusters can be observed for the cancer sample that are totally absent in the normal sample.

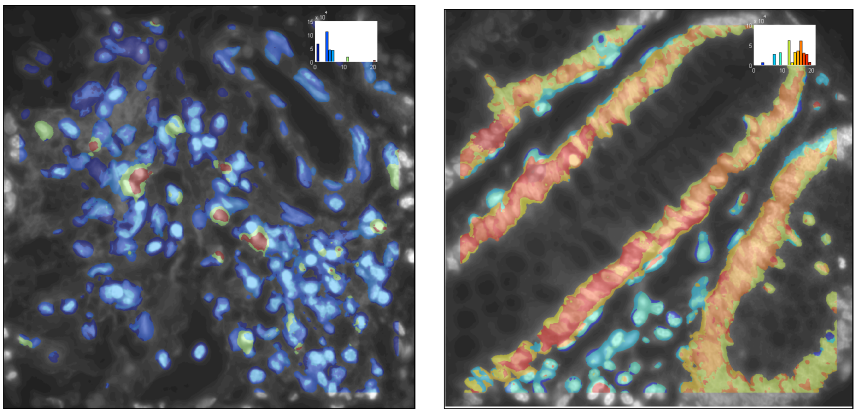

Figure 2: DAPI channel images of the two specimens shown in Figure 1 (in the same order) overlaid with the results of agglomerative hierarchical clustering method, after picking the centroids of the top 20 clusters. Note that unlike Figure 1, the pseudo coloring shown in this Figure is not topology preserving in the 13-dimensional space of intensities. The bar charts near the top right of each of the two maps above show a histogram of the top 20 Molecular Co-Expression Patterns (MCEPs) found in the sample. The pseudo coloring here is based on the MATLAB® `jet` colormap.

ponding phase contrast images. From these data, 13 channels, each usually corresponding to its own protein, were selected for a deeper analysis. These four 13-dimensional TIS images are referred to as I^{c}_{1}, I^{c}_{2}, I^{n}_{1}, I^{n}_{2} (c: cancer, n: normal). In each image, a pixel (x, y) is associated to a 13-dimensional intensity vector $i^{(x,y)} = (i_1, \ldots, i_{13})^{(x,y)}$ with $i_j \in [0,1]$ for all $j = 1,2,\ldots,13$.

2.2. Topology Preserving Pseudo-Colour Molecular Co-location Maps (MCM)

Imaging in systems biology may have the disadvantage that only a (comparatively) small number of molecular variables can be identified. However, one advantage of imaging in biology is the availability of molecular-topological information, i. e. the assignment of molecular information to the spatial domain, for instance tissue morphology. This allows a more differentiated analysis of molecular networks dependent on the anatomical site. Thus, providing a co-location visualization in the image domain which supports a simultaneous analysis of morphology and molecular networks is important. In this work color is chosen to encode co-location patterns. In contrast to previous TIS publications we use colors to represent pattern similarity. This means that when two pixels (x, y) and (x', y') show similar co-location grey value patterns $\boldsymbol{i}^{(x,y)}$ and $\boldsymbol{i}^{(x', y')}$ they are drawn in similar colors in a pseudo-color map.

This follows the idea that protein patterns which differ only in one or few proteins may contribute to related functions. One can easily obtain a visual impression of co-location pattern distribution of whether clusters with similar colors, and thus possibly with similar functions, form compact sub-regions in an image or are spread over the whole image. To map all co-location patterns $\{\boldsymbol{i}^{(\alpha)}\}_{\alpha = 1,\ldots,P}$ of one or two TIS images (with P as the *total* number of pixels) the image is first preprocessed. To this end a modified median filter was applied to eliminate outliers. Afterwards, bilateral filtering [11] was applied to smoothen homogenous regions while preserving the edge information. The grey values in each image of a stack were scaled to [0,1] using a tanh(·) squashing function which also introduces a slight contrast enhancement in the images. In the next step we apply vector quantization to the data and project the cluster centers $\{\mathbf{u}^{(k)}\}k = 1,\ldots,K$ (with K as the number of clusters) to a two dimensional space. Next we select a continuous 2D color scale and map the 2D coordinates of the centroid of each cluster to positions in this scale. For dimension reduction, one can apply for instance PCA, Sammon mapping, LLE or other techniques. In this study, we apply the self - organizing map [12] since it combines the steps of clustering and dimension reduction following the topology preservation principle (in 13-dimensional space). We applied our approach for the exploration of high dimensional MRI data by using the (hue, saturation)-disc of a HSV cone as color scale [13]. To create a pseudo color map for one image, each pixel (x, y) is mapped to a color by looking up the best-matching cluster unit using a chosen metric $d(\cdot)$, for instance using the euclidean distance or the angular metric: $\mathbf{u}(k)$ with $k = \text{argmin}_k\ (d(\mathbf{i}(x,y),\mathbf{u}(k)))$ and its corresponding color coordinates. We have integrated the whole approach in our online bioimage analysis platform BIOIMAX (BioImage Mining, Analysis and eXploration) so that users can browse the visualization independently of their operating system, using only their web browser, as shown in Figure 1. In the middle one can see two pseudo color images (shown in grey due to workshop format restriction). On the left a TIS pseudo color image from cancer tissue is shown, and on the right a TIS image from normal tissue.

Both pseudo color maps are rendered using the clustering obtained for the combined data set of all four images. Using the mouse cursor, single clusters can be selected from the color palette on the right and their components can be examined. On the left, the list of all antibodies is shown, so that a user can select one image and tune the opacity of the pseudo color map. In this way, color can directly be linked to grey value intensity.

2.3. Molecular Co-Expression Patterns (MCEPs)

In this section, we describe a slightly different approach for finding and displaying co-expression patterns found in the TIS image data. We first rescale the intensity values in each of the aligned TIS images to the range [0,1]. As a next step, we segment the aligned DAPI channel 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) using the Bayesian information criterion (BIC) for model selection [14, 15]. We then employ the standard agglomerative hierarchical clustering method and pick the top twenty clusters localized to nuclei and their vicinities. We call such a cluster a Molecular Co-Expression Pattern (MCEP). In a similar fashion to the method described in the previous section, each of the centroids of these clusters is given a unique pseudo color. However, unlike the method described in the previous section, the color allocation here is not topology preserving. Instead, 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.

3. RESULTS

Our results show readily identifiable visual differences between tumor and normal tissue in a new information domain. Due to the introduction of unsupervised learning in TIS visual data mining, the color code follows the topology preservation principle mapping the co-location information in the fluorescence values to color coordinates. Thus, differences can be observed on the level of protein co-location patterns and on the level of (x,y) pixel topology as well (Fig. 1). This can reveal relationships between protein co-location **and** tissue morphology which is fundamentally different to the standard TIS visualization approaches using thresholds and random colors [5,8].

DAPI-labelled nuclei allow an easily implementable approach to automated identification of individual cells and the extraction of cellular co-location features. An in-depth analysis of these patterns showed additional co-location patterns for proteins which were different for tumor and normal tissue. Results of cell-localized MCEPs for the same four TIS image stacks are shown in Figure 2. It is clear from these results that this approach both confirms the observations made by earlier pseudo coloring and serves as a complementary tool for exploratory analysis of the multi-variate TIS data.

4. CONCLUSIONS

The proposed approach shows how complex, high-dimensional microscopy data can be analyzed and explored for interesting co-location information that can be fused with protein network or pathway information from other sources such as pathway data-bases.

6. ACKNOWLEDGMENTS

We thank W. Schubert, who introduced us to TIS, and helped us establish a TIS machine at the University of Warwick, and members of his team at ToposNomos and the University of Magdeburg, especially A. Krusche and R. Hillert, who have provided invaluable support when we needed it. Special thanks go to Sayan Bhattacharya for contributions to the design of the antibody library.

7. REFERENCES

[1] S. Megason and S. Fraser, *"Imaging in systems biology.," Cell*, vol. 130, pp. 784–795, 2007.

[2] V. Starkuviene and R. Pepperkok, *"The potential of high-content high-throughput microscopy in drug discovery," Br J Pharmacol,* vol. 152, no. 1, pp. 62–71, Sep 2007.

[3] D. Cornett, M. Reyzer, P. Chaurand, and R. Caprioli, "Maldi imaging mass spectrometry: molecular snapshots of biochemical systems," *Nature Methods*, vol. 4, pp. 828 – 33, 2007.

[4] 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, vol. 102, no. 29, pp. 10159–64, Jul 2005.

[5] W. Schubert, B. Bonnekoh, A. Pommer, L. Philipsen, R. Böckelmann, Y. Malykh, H. Gollnick, M. Friedenberger, M. Bode, and A. Dress, "Analyzing proteome topology and function by automated multidimensional fluorescence microscopy.," *Nature Biotechnology,* vol. 24, pp. 1270–1278, 2006.

[6] T.W. Nattkemper, "Multivariate image analysis in biomedicine: a methodological review *JOURNAL OF BIOMEDCIAL INFORMATICS*, 37, (5), 380-391, 2004.

[7] J. Herold, T.W. Nattkemper, "Multrivariate Image Mining" Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1) 2-13, 2011.

[8] S. Bhattacharya, G. Mathew, E. Ruban, D. Epstein, A. Krusche, R. Hillert, W. Schubert, and M. Khan, "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.*, vol. 9, no. 12, pp. 611225, 2010.

[9] T. W. Nattkemper, H. Ritter, W. Schubert, "Extracting Patterns of Lymphocyte Fluorescence from Digital Microscope Images" AMIA 1999, Washington DC, *Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP 99), Workshop Notes*, 79-88, 1999.

[10] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *ICCV*, 1998, pp. 839–46.

[11] J Herold, S Abouna, L Zhou, S Pelengaris, D Epstein, M Khan, TW Nattkemper "Integrating automized Semantic Annotation and Information Visualization for the analysis of multichannel Fluorescence Micrographs from Pancreatic Tissue." *COMPUTERIZED MEDICAL IMAGING AND GRAPHICS*, 34 (2010), 446-52.

[12] T. Kohonen, *Self-Organizing Maps*, Springer, 3 edition, 2001.

[13] A. Saalbach, J. Ontrup, H. Ritter, and T. W. Nattkemper, "Image fusion based on topographic mappings using the hyperbolic space," *Information Visualization*, vol. 4, pp. 266–275, 2005.

[14] C. Fraley and A. E. Raftery, "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis," The Computer Journal, vol. 41(8), pp. 578-588, 1998.

[15] N. Rajpoot and M. Arif , "Unsupervised shape clustering using diffusion maps," Annals of the BMVA, 2008(5), pp. 1-17, 2008.