ANALYZING MULTI-TAG BIOIMAGES WITH BIOIMAX COLOCATION MINING TOOLS

Jan Kölling, Magnus Rathke, Dominic Gardner, Tim W. Nattkemper

Biodata Mining Group
Faculty of Technology
Bielefeld University, Germany

Sylvie Abouna, Michael Khan

Biomedical Research Institute
School of Life Sciences
Coventry, UK

ABSTRACT

The application of multi-tag protocols in fluorescence microscopy allows the visualization of a large number (> 10) of molecules (i. e. proteins) in a sample (like a tissue section). However, the analysis of such high dimensional bioimages is a difficult task for most of the labs, since software solutions for particular data mining steps are difficult to use or just not available. In this paper we present two new free online tools: MICOLT (Multivariate Image COlocation Tool) and MIFIST (Multivariate Image Frequent Item Set Tool). Both tools can be used via our recently proposed online bioimage analysis platform BioIMAX, so users can upload their bioimage data, apply the tools and share the results with other invited users based on BioIMAX' concept of shared virtual projects. Data mining with these tools includes the computation and visualization colocation factors well established in the microscopy community (like Mander's score) and association rule mining following the frequent item set principle, thereby supporting large and small scale analysis.

Index Terms— Bioimage informatics, fluorescence microscopy, proteomics, tissue analysis, data mining, visualization, web systems, science 2.0

1. INTRODUCTION

In recent years, new fluorescence microscopy protocols have been proposed to study the spatial dependencies of N proteins in a sample, e. g. a tissue section or a cell culture. With these new protocols, the number of N now depends just on the availability of N high quality antibodies for the N proteins of interest. Conjugating all N antibody markers to one and the same dye and applying these tags in a cyclic protocol of marking, imaging, soft bleaching allows recording a stack of N grey value fluorescence micrographs, all showing the same field of view (FOV). This particular multi-tag technique is usually referred to as MELC (Multi-Epitope Ligand Cartography) or TIS (Toponome Imaging System) and has been originally introduced in [1] (we will refer to the images as TIS images in the remaining of this paper). TIS images belong the category of so called multivariate bioimages (MBI) [2] which relate to so called high content imaging techniques. Such techniques have been proposed recently to fill some gaps in systems biology research which are left open by traditional molecular techniques, since those are based on homogenizing a sample. As a consequence these techniques neglect the topological orders of molecular self organization, i. e. the spatial dependencies in molecular networks [3, 4]. To visualize MBI, clustering and dimension reduction can be applied to render dynamic pseudo color maps of the images as we have shown in recent works [5, 6]. However, MBI can also be analyzed following a more traditional data mining approach. We consider one image stack of dimension $n_x \times n_y \times N$ (n_x: image width, n_y: image height, N number of grey values associated to each pixel/number of tags). This stack has to searched for (hidden) regularities and patterns in the fluorescence features $\mathbf{f}_{x,y} = (f^{(0)}, ..., f^{(N-1)})_{x,y}$. With $F^{(a)} = \{f^{(a)}_{x,y}\}$ we will refer to one fluorescence micrograph of the stack, recorded with the antibody / marker a (in the following, we will omit the subscripts $_{x,y}$ if possible). Large scale patterns can be expressed by large scale dependencies, covariances, colocations or correlations of signals from pairs of antibodies. Small scale patterns can be expressed by single or a cluster of particular N-dimensional signal patterns, observed at some pixels in an image. To analyze such large scale or small scale patterns, researchers usually apply a variety of software, e. g. MATLAB, software provided by the microscope manufacturer, R or in-house products.

In this manuscript we present an online, i. e. web-based approach to multivariate bioimage data mining as outlined above. The technological platform is provided by the recently introduced open online bioimage analysis platform BioIMAX (BioImage Management, Analysis and eXploration) [7]. The motivation for BioIMAX was the observation, that due to the rapidly evolving options and flexibility in internet connection (Wi-Fi, UMTS, ethernet) and the increasing bandwiths new, pure online - based tools for bioimage analysis are attractive for some life science researchers, who may want to share some of their data and results with their collaborators, independent from their whereabouts, condition to an internet connection. BioIMAX is a pure web-based approach, allowing users to upload MBI, apply exploratory data analysis to MBI, annotate regions of interests, share MBI with

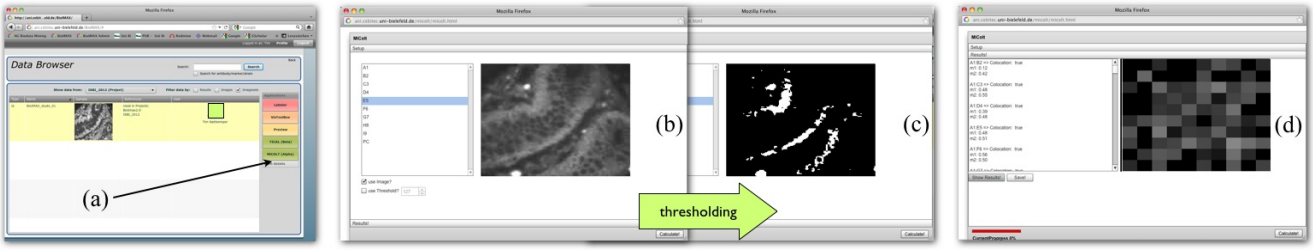

Fig. 1. MICOLT is started within BioIMAX at the project browser level after selecting one data set (a). In the first step (b) the user can deselect images in a stack and set a threshold manually after activating the box (otherwise, Otsu method-based thresholds are applied as a default). Manually changing the threshold dynamically updates the image display as a binary image. Although the image is not really binarized (see eq. (2)) a binary display is more appropriate to tune the threshold (c). The result is shown in (d), the result is shown as text on the left and as grey value matrices on the right. The result Manders score values can be saved to disc so the data can be rendered with alternative software or written to a database.

other users and apply different kinds of analytical tools to their data. Here we present two new tools, MICOLT (Multivariate Image CO Location Tool) and MIFIST (Multivariate Image Frequent Item Set Tool), which can be applied to data via the BioIMAX platform. Using the login name "test" and the password "test1", the reader can log into BioIMAX [1], choose the project `ISBI_2012` from the project list, activate the `project browser` and apply the tools as explained below.

2. MATERIALS

To illustrate the functions of MICOLT and MIFIST we will consider a data set from a TIS study. TIS imaging was applied to two tissue samples from a colon cancer patient. One tissue sample was selected from cancerous tissue, the other sample was selected from healthy colon tissue from the same patient. An antibody library of 22 tags (see [8] for details) was applied to record a stack of 22 fluorescence images from manually selected two visual fields in each sample, leading to four TIS data sets. After image registration has been applied [9, 10], a set of $N = 8$ channels, i. e. proteins were selected for a deeper analysis. In each image, a pixel (x, y) is associated to a 8-dimensional intensity vector $\mathbf{f} = (f_1, ..., f_8)$ with $f_a \in [0; 1]$. To view the data in BioIMAX, the user starts the `Preview` tool on the right [7].

3. METHODS

The BioIMAX system connects the user to a data server to upload and manage the the image data and with a powerful compute server to offer efficient analytical compute services for data mining. On the client side, BioIMAX is designed as a rich internet application using the FLEX environment. This concept allows the combination of powerful data analysis tools, efficiently implemented in C/C++ with sophisticated and visually appealing interfaces which is attractive, especially for users who do not apply data mining algorithms every day.

3.1. MICOLT

One basic question in the analysis of MBI such as TIS images (or for instance MALDI images) is, if some of the considered molecules or residues colocate across the visual field. In the context of TIS analysis, a colocation of two (or more) proteins could hint to particular molecular networks or active pathways. In other techniques, the colocation of signals would help in sorting or binning the images of one stack into groups which should be analyzed in a next step, but restricted to the colocating signals, for instance in the case of MALDI images. The most established parameter to measure colocation is Manders' score [11], which is defined for the comparison of two fluorescence micrographs F_a and F_b. To quantify the colocation of the two fluorescence signals recorded with antibodies a and b the following two coefficients are computed:

$$m_a = \frac{\sum_{x,y} C_b(f^{(a)})}{\sum_{x,y} f^{(a)}} \text{ and } m_a = \frac{\sum_{x,y} C_a(f^{(b)})}{\sum_{x,y} f^{(b)}}, \qquad (1)$$

$$\text{with } C_b(f^{(a)}_{x,y}) = \begin{cases} f^{(a)}_{x,y} & \text{if } f^{(b)}_{x,y} > 0 \\ 0 & \text{else} \end{cases}$$

and $C_a(f^{(b)}_{x,y})$ defined analogous. However, before the two scores are computed thresholding is applied to each image $F_a(a = 1, ..., N)$ of a stack to map weak signals to the value 0:

$$f^{(a)'}_{x,y} = \begin{cases} f^{(a)}_{x,y} & \text{if } f^{(a)}_{x,y} > t_a \\ 0 & \text{else} \end{cases} \qquad (2)$$

To select an appropriate threshold t_a for each image in a stack, the user can either apply the default value, which is computed

[1] `http://ani.cebitec.uni-bielefeld.de/BioIMAX/`

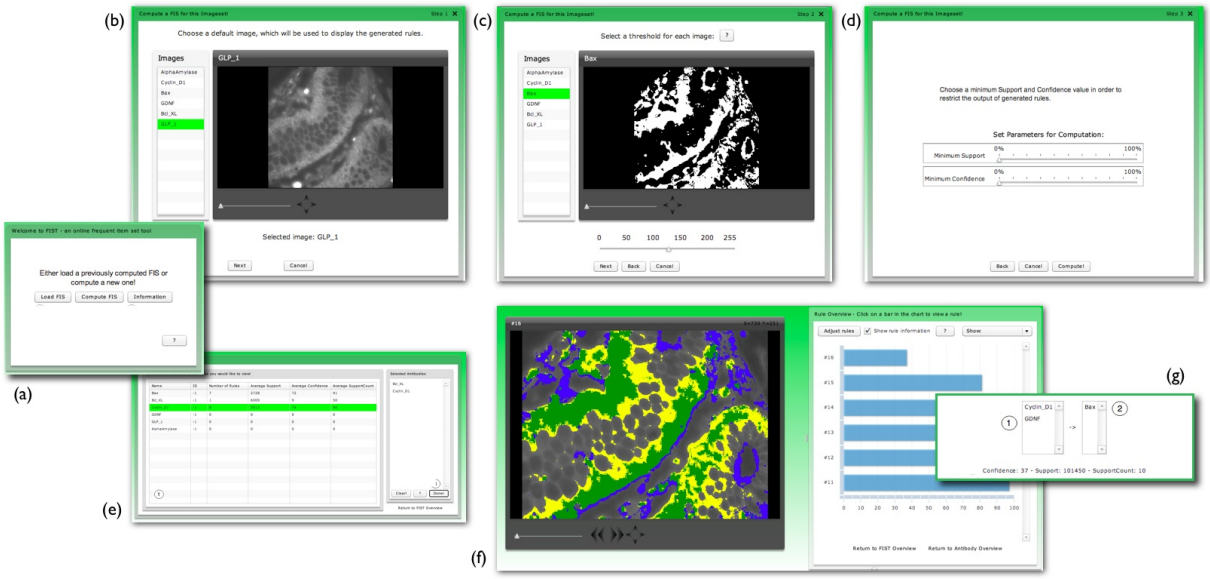

Fig. 2. MBI frequent item set analysis with the MIFIST tool: MIFIST runs basically in two modes, selected in the first screen (a). In the compute mode, a MBI (b) is searched for frequent items set rules (FIS) with a minimum support and confidence (note that in this screen, one image from the stack has to be selected as the reference image for later visualizations). In a next step the images are transformed into binary images by manually adjusting thresholds (c) and the thresholds t_s and t_c are selected (d), before finally the computation of the rules starts. In the display mode, a computed list of frequent item set rules (FIS) can be analyzed after it has been selected in screen (a). In a next step (e) the user chooses the antibodies / markers of interest, to filter out those item sets that do not show these ones. Finally, the rule overview is displayed (f). The location of a rule $I_A \rightarrow I_B$ is displayed in a special color code: I_A = yellow, I_B=blue, $I_A \cup I_B$= green. The right side displays all rules as a bar chart, each bar representing one rule. Selection of a bar triggers the display of the rule (g).

using Otsu's method [12], or select the threshold manually using a slider and inspecting the thresholding result in real time as displayed in Figure 1. MICOLT displays the Manders' scores for all pairs of images in as text and as a non-symmetric grey value matrix. The text can be saved to disc, so alternative visualization tools (such as graph / network visualizations) can be applied. In addition to Manders' scores MICOLT also computes the $N \times N$ covariance matrix C with coefficients $c_{a,b} = (f^{(a)} - E(f^{(a)}))(f^{(b)} - E(f^{(b)}))$ and display it as a grey value matrix (see Figure 1).

3.2. MIFIST

While MICOLT compares pairs of fluorescence images on a large scale, MIFIST was developed to study the patterns on a pixel level. The idea is, that in the analysis of MBI data, one could consider all $M = n_x \times n_y$ pixel as M observations of N variables and one way to analyze this data set without any prior knowledge is association rule mining [13]. One algorithm to compute association rules is the frequent item set algorithm which is applied in many data mining contexts such as market basket analysis. MIFIST applies the frequent item set algorithm to binary image data. Thus, each image of the MBI has to be thresholded, transforming the grey value images into binary images:

$$b_{x,y}^{(a)} = \begin{cases} 1 & \text{if } f_{x,y}^{(a)} > t_a \\ 0 & \text{else} \end{cases} \tag{3}$$

so the frequent item set algorithm is applied to a stack of N binary images $B^{(0)}, ..., B^{(N-1)}$ with $b_{x,y}^{(a)} \in 0; 1$. And the binary values of one pixel $\mathbf{b}_{x,y} = (b^{(0)}, ..., b^{(N-1)})$ are considered as one observation of N variables or items. An item set is a collection of variables with $b_{x,y}^{(a)} = 1$ (at least one). In the case of TIS images, one item corresponds to an observed protein, visualized with a fluorescence value $f^a > t_a$. A frequent item set is an set of items (i. e. proteins) that has met a certain condition: Its *support* has to be greater than or equal to the minimum support t_s and the support is defined as the number (or percentage) of observations showing this time set. After computing all item sets, i. e. all observed combinations proteins that have been observed with the minimum support, each item set I is transformed into rules. Rules are generated by dividing an item set I into two disjoint subsets I_A and I_B (i. e. $I_A \cup I_B = I$). For each possible rule the *confidence*, is computed. The confidence in a rule $I_A \rightarrow I_B$ is defined as the percentage of all observations of item set I_A that also showed the item set I_B. Since we deal with image data, that is always subject to distortions and noise, we allow users to filter

out rules with a confidence below a given confidence threshold t_c. Another special feature of MIFIST is, that we need to account for the fact, that in bioimage analysis not only the colcation pattern, i. e. the frequent item set rule is of interest but also its location. Thus, after computing the rules for given thresholds t_s and t_c, MIFIST allows an interactive filtering and display of rules at the anatomical site in the image, as shown in Figure 2.

4. RESULTS

Both tools MICOLT and MIFIST can be applied through BioIMAX to one example data set. Registered BioIMAX users can upload own data and apply the tools to the data without restrictions. In practice we recommend first to apply MICOLT to study the large scale dependencies between the considered molecules. In a next step, and based the results, MIFIST is applied maybe to a subset of the data and with appropriate values for t_s and t_c.

5. DISCUSSION

MICOLT and MIFIST resemble first steps into the direction of sophisticated MBI mining through the web. Next, we will extend MICOLT with other indices (such as mutual information for instance) and develop alternative filtering operations for MIFIST and its applicability to sets of MBI.

6. ACKNOWLEDGEMENT

Special thanks go to Sayan Battacharya for efforts in designing the antibody library. 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.

7. REFERENCES

[1] W. Schubert, B. Bonnekoh, A.J. Pommer, L. Philipsen, R. Boeckelmann, Y. Malykh, H. Gollnick, M. Friedenberger, M. Bode, and A.W. Dress, "Analyzing proteome topology and function by automated multidimensional fluorescence microscopy," *Nat Biotechnol*, vol. 24, no. 10, pp. 1270–8, Oct 2006.

[2] J. Herold, C. Loyek, and T.W. Nattkemper, "Data mining in multivariate images," *Wiley Interdisciplinary Reviews: DATA MINING AND KNOWLEDGE DISCOVERY*, vol. 1, no. 1, pp. 2–13, Jan 2011.

[3] 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.

[4] S.G. Megason and S.E. Fraser, "Imaging in systems biology," *Cell*, vol. 130, no. 5, pp. 784–95, Sep 2007.

[5] D. Langenkämper, J. Kölling, S. Abouna, M. Khan, and Nattkemper T.W., "Tical - a web-tool for multivariate image clustering and data topology preserving visualization.," in *Workshop on Microscopic Image Analysis with Applications in Biology (MIAAB)*, Heidelberg, Germany, 2011.

[6] D. Langenkämper, J. Koelling, A. Humayun, M. Khan, N. Rajpoot, D.B.A. Epstein, and Nattkemper T.W., "Towards protein network analysis using tis imaging and exploratory data analysis," in *Workshop on Computational Systems Biology (WCSB) 2011*, Zuerich, Switzerland, 2011.

[7] C. Loyek, N. Rajpoot, M. Khan, and T.W. Nattkemper, "Bioimax: A web 2.0 approach for easy exploratory and collaborative access to multivariate bioimage data," *BMC Bioinformatics*, vol. 12, no. 1, pp. 297, 2011.

[8] S. Bhattacharya, G. Mathew, E. Ruban, D.B.A. 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] A. Humayun, A. Raza, C. Waddington, S. Abouna, M. Khan, and N. M. Rajpoot, "A framework for molecular co-expression pattern analysis in multi- channel toponome fluorescence images," in *Proc. of MIAAB*, Heidelberg, Germany, 2011.

[10] S. E. A. Raza, A. Humayun, T. W. Nattkemper, D. Epstein, M. Khan, and N. Rajpoot, "Registration of multiplexed fluorescence images using phase contrast images," in *Proc. of ISBI*, Barcelona, Spain, 2012, submitted.

[11] S. Bolte and F. P. Cordelieres, "A guided tour into subcellular colocalization analysis in light microscopy," *J Microsc*, vol. 224, no. 3, pp. 213–32, 2006.

[12] N. Otsu, "A threshold selection method from gray-level histograms," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 9, no. 1, pp. 62 –66, 1979.

[13] I.A. Witten and F. Eibe, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, 2 edition, 2006.