TICAL - a web-tool for multivariate image clustering and data topology preserving visualization

Daniel Langenkämper[1], Jan Kölling[1], Sylvie Abouna[2], Michael Khan[2], Karsten Niehaus[3], Tim W. Nattkemper[1]

[1]Biodata Mining Group, Faculty of Technology, Bielefeld University, Germany
[2]Department of Biological Sciences, Warwick University, Coventry, United Kingdom
[3]Faculty of Biology, Department for Proteome and Metabolome Research, Bielefeld University, Germany

***Abstract*— In life science research bioimaging is often used to study two kinds of features in a sample simultaneously: morphology and co-location of molecular components. While bioimaging technology is rapidly proposing and improving new multidimensional imaging platforms, bioimage informatics has to keep pace in order to develop algorithmic approaches to support biology experts in the complex task of data analysis. One particular problem is the availability and applicability of sophisticated image analysis algorithms via the web so different users can apply the same algorithms to their data (sometimes even to the same data to get the same results) and independently from her/his whereabouts and from the technical features of her/his computer. In this paper we describe TICAL, a visual data mining approach to multivariate microscopy analysis which can be applied fully through the web. We describe the algorithmic approach, the software concept and present results obtained for different example images.**

***Index Terms*— bioimage informatics, multivariate bioimage analysis, fluorescence microscopy, high-content screening, MALDI imaging, multi-tag imaging, MELC, TIS, visual data mining, visualization, machine learning, clustering, dimensional reduction**

I. Introduction

In the last ten years, bioimage informatics has evolved as a new branch in the tree of bioinformatics. This community has made great achievements in the development of freely available toolkits to support biologists in setting up a bioimage data base for image management, retrieval, annotation and processing as well [1], [2], [3]. The main reason behind this trend is the ongoing development of bioimage technologies towards automation, higher throughput, better imaging quality and higher signal dimension. The latter can be achieved using multi tag imaging and high content screens [4], [5], [6], [7], [8] or MELC/TIS (toponome imaging system) imaging [9]. Recently, other non-optical techniques have been proposed to record high dimensional bioimage data visualizing a) co-location of proteins/metabolites (such as MALDI imaging [10]) or b) molecular interaction patterns, such as vibrational spectroscopy [11]. The analysis of such multivariate bioimage data is a non-trivial task, since two domains of information are of interest to the biologist here: the tissue/cell morphology *and* the high dimensional space of co-location patterns, where the dimension D is given by the number of molecules (or residues) which are visualized with the techniques [12]. If the number of tags or visualized molecular components is low (i. e. $D \leq 3$) the images can be explored with standard procedures such as RGB pseudocolor visualization (i. e. the single grey value images are interpreted as red-, green- or blue-channel) or link & brush techniques [13]. However, in case of larger values of D the analysis strategy is less straightforward and new approaches and tools are needed that can cope with the high dimension. However, although the fields of machine learning and image processing seem to offer a wealth of promising technologies to find the hidden regularities in high dimensional data the application on biology depends on the availability of these methods as integrated tools. This is sometimes hampered by the fact that most of the underlying algorithms are computationally expensive or the installation of the software depends on a particular platform feature like the operating system of the computer. This is a serious problem, since one reason for the rapid developments in molecular biology and life sciences in the last decade was the free availability of computational tools for the analysis of genomics or proteomics data. If a researcher wants to analyze their sequences of amino acids or some results from mass spectrometry, free tools such as BLAST, CLUSTALW or MASCOT enable a quick first look at the data since they can be applied just through an internet browser. Thus we believe, that not only the development of image analysis tools is important but also the development of technical concepts that allow users to apply these tools to their data without any time-consuming burden of installing some software on their own computers. In this paper we present TICAL (Tool for Image Clustering And anaLysis), which is a web tool to support visual data mining in any kind of multivariate bioimages. We present the technical concept and show some examples obtained for multivariate fluorescence micrographs. A test user access is provided to test the system.

Corresponding author: Tim W. Nattkemper, email: tim.nattkemper@uni-bielefeld.de

II. Material

The TICAL tool and its functions are demonstrated with two different data sets. First we apply TICAL to fluorescence micrographs from a study about bacterial infection in cell cultures [7]. The images were obtained using three stains in a high content screen to visualize 1) the nucleus (Hoechst 33342), 2) the cytoplams (WCSR) and 3) the bacteria Listeria monocytogenes (GFP), so $D = 3$. The other data set is from a study of protein co-location patterns using TIS imaging in colon cancer tissue and normal healthy tissue [14], where $D =$

9 antibodies have been selected. We show results obtained from one visual field in a tissue section from healthy tissue.

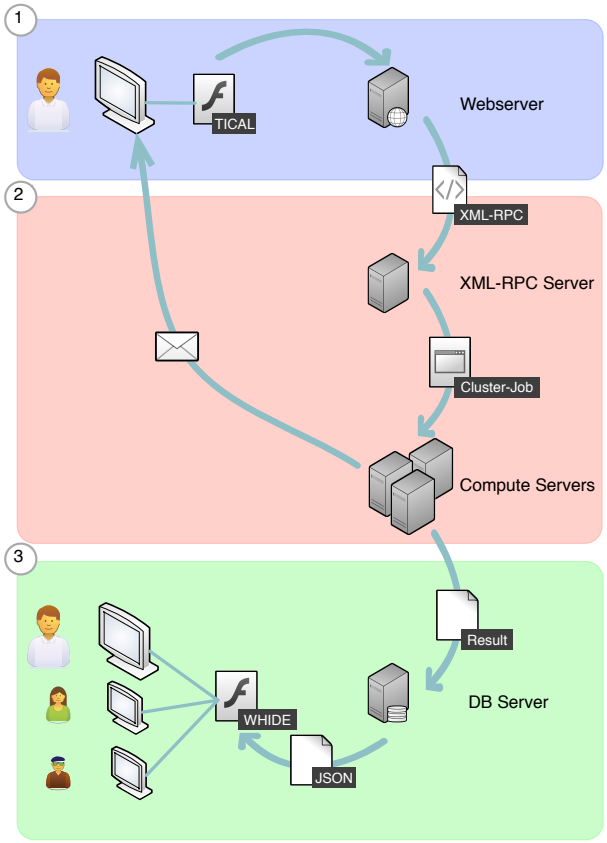

Fig. 1. The TICAL architecture consists of three layers. 1. The user submits a request to the web server, which triggers a XML-RPC call. 2. The call is received by the XML-RPC server and starts the execution of the clustering software on the compute servers using the parameters entered by the user. 3. When the algorithms have finished the user is notified by mail, the clustering result (usually a set of prototypes) is written to a file and saved in a database together with additional metainformation. By requesting to view the result in another web application such as WHIDE, the corresponding JSON file is loaded and a visualization of it is presented to the user.

III. Methods

One way to analyze multivariate image data is a combination of clustering and dimension reduction. In a first step, a vector quantization clustering is applied to the D-dimensional image data. A set of K prototype vectors (often also referred to as reference or codebook vectors) $\mathbf{u}^{(k)}_{k=1,\ldots,K} \in [0,1]^D$ is trained using the the co-location feature vectors $\mathbf{x}^{(\xi)} \in [0,1]^D$ that contain the D grey values of pixel ξ, normalized to scale $[0,1]$. Second, a cluster map is generated that shows for each pixel ξ the value $\kappa \in [1,\ldots,K]$ which is the index of the best matching prototype to its co-location feature vector $\mathbf{x}^{(\xi)}$. Third, the prototypes are mapped to a color scale and the cluster map is visualized showing for each pixel ξ the result coloring. This mapping is achieved by applying a dimension reduction algorithm to the set of all prototypes $U = \{u^{(k)}\}$ to map each prototype to a low-dimensional representation $\mathbf{n}^{(\mathbf{k})} \in [0,1]^d$ with $d \leq 3$. The aim of this reduction is to preserve as much of the important topological features of the prototype distributions as possible. These new coordinates are interpreted as the prototypes' coordinates in the chosen color scale and the cluster map is displayed showing for each pixel ξ the color of its best matching prototype. To choose the color scale (and its dimension), different strategies can be applied. Levkovitz [15] has proposed a variety of strategies to render one-dimensional color scales. However, we propose to use a two dimensional color scale for a practical reason. A two-dimensional color scale can be displayed as a plane and the positions of the cluster prototypes can be marked with icons in this plane. Moving the icons on the plane can be used to change the mapping between prototypes and colors which is a powerful mechanism to enhance the color contrast for particular prototypes, i.e. subregions in the D-dimensional co-location space. Recently, we have illustrated the power of this approach in an application to TIS images in a colon cancer study [14]. Using our approach we effectively visualized, which features normal and cancer tissue have in common and which features are different. Morphological features were visually accessible simultaneously to high-dimensional co-location features. In the next section we describe the technical concept of TICAL, which basically follows the idea of a client server architecture.

A. Implementation

The TICAL architecture consists of three layers as can be seen in Figure 1, modeling a client-server-architecture as mentioned before. First the user chooses several parameters for clustering in the TICAL user interface (see Figure 2). When submitting the job afterwards, a HTTP-POST request is sent to the web server, where it triggers a PHP script on the web server. The script executes a XML-RPC (XML Remote

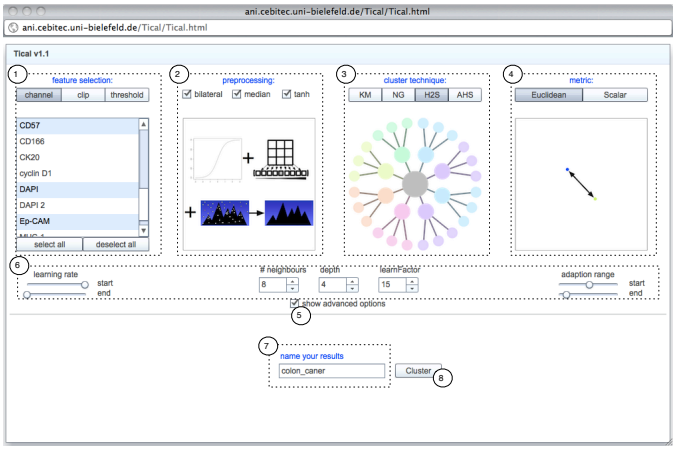

Fig. 2. The TICAL user interface: The feature selection (1) consists of channel selection, image clipping and threholding each channel individually. Preprocessing (2), clustering techniques (3) and metric selection (4) are presented in visual manner to aid the user in selection. Different techniques can be chosen here. Professional users might check the box (5) to get advanced options. Non-professionals can always stick with the defaults and are usually fine with that. The parameters under (6) control how clustering is performed. Later the user can find the result under the name provided (7) in BioIMAX. By pressing the button (8) the cluster-job is submitted.

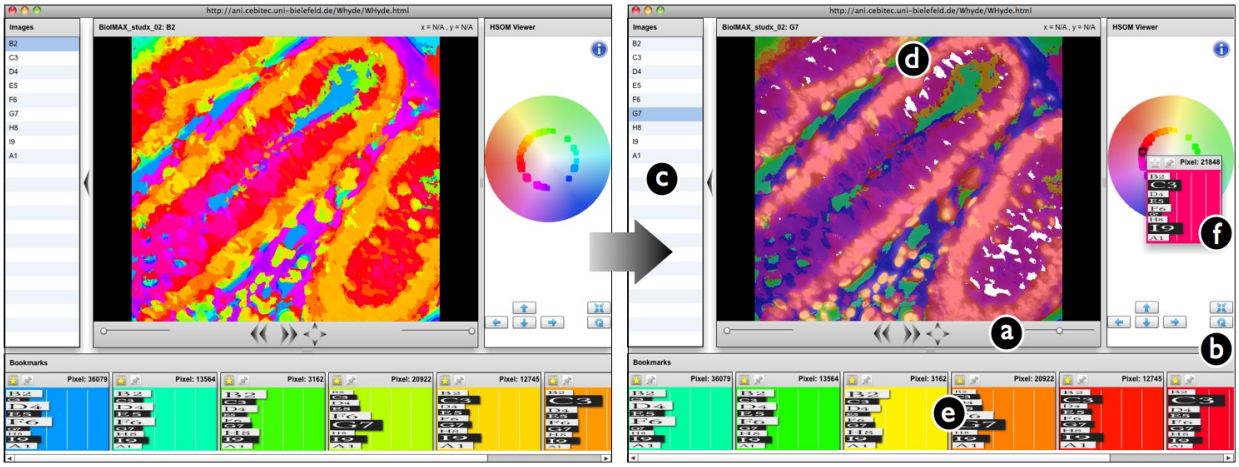

Fig. 3. A pseudo color visualization obtained with neural gas vector quantization clustering ($K = 50$). PCA was applied to map prototypes to color scale coordinates in a circular color scale varying the hue and the saturation of a color. Similar colors represent similar co-location patterns which point at similar biological functions. The whole framework supports the process of visual data mining with several functions to modify the display interactively which is shown in the figure moving from the left screen to the right screen: Using slider (a) the opacity of the cluster map is controlled to create a fusion display of the cluster pseudocolor map and one original fluorescence image. In the color scale mapping window (b), the cluster icons can be shifted using the arrows and the color scale can be rotated to change the pseudocolor map (d) according to individual criteria, i. e. considering the fact, that human observers do not have the same sensitivity for color contrast along the visual spectrum. While the color mapping is modified, the colors of the selected cluster icons in the bottom row (e) are updated accordingly (e).

Procedure Call)[1]. This allows us to separate the web server and compute server for both performance and security benefits.
In the context of TICAL, the procedures executed by the XML-RPC server are the clustering algorithms for instance. Since TICAL is designed as a web tool which can be applied by multiple users, the procedures ought to be implemented with threads to avoid blocking of the server. Because this limits the number of jobs at a time and requires manual scheduling we instead chose to submit the heavy computing tasks (such as clustering) to a grid of compute servers.
Each clustering job consists of the the following components:

1) pre-processing
2) pixelwise feature extraction
3) clustering
4) post-processing
5) saving results

These tasks are performed using the open image processing software library OpenCV [16] and our own in-house machine learning software library MLlib, which is implemented in C/C++. The clustering algorithms are all written in C++ for performance reasons. Furthermore most of our clustering algorithms (see below) use parallel computing to achieve a better runtime. For parallelization OpenMP [17] is used. As soon as one clustering job is finished the user gets an email notifying him of the availability of his results which are saved as compressed flat files to enable fast transfer to the client's computer later on. Metainformation on the clustering job (time/date, input data set, pre-processing parameters, clustering parameters) is stored in a MySQL database. If the user wants to view the results for instance via the online tool WHIDE (see below) the metainformation are read from the database and the representation file is parsed to generate a visual result of the data.

The user interface of TICAL is written in Adobe Flex because of the broad availability of flash-enabled web browsers. The XML-RPC server is written in C++ and uses the DRMAA (Distributed Resource Management Application API) library [18]to communicate with the distributed resource management of the compute grid. The grid is an aggregation of loosely coupled computers and the tasks of scheduling and load balancing is done automatically by the engine. The result files are stored in zlib compressed JSON format, which is a lightweight alternative to XML. zlib compression is used because it is the only compression which can be uncompressed by Adobe Flex.

B. How to use TICAL online

Examples from both studies can be accessed via our online bioimage analysis platform BioIMAX (BioImage Management, Analysis and eXploration)[2] using the login `miaabtest` and the password `go4miaab`. After logging into BioIMAX (please ensure you use a flash-enabled browser such as firefox), please choose the project MIAAB from the list on the left and start **project browser** below. After a few seconds the project browser appears, showing two data sets. To start TICAL, select one data set (now highlighted in yellow) and activate TICAL on the right. A new window opens, showing the TICAL user interface (see Figure 2). Here, one can choose the tag images that should be considered (since in some applications one might be interested only in a subset of tags), pre-processing options and the clustering algorithm to be applied. In the current version three different cluster algorithms can be applied: k-means, neural gas [20] and the hyperbolic self-organizing map (H^2SOM) [21]. Regarding the

[1]XML-RPCs are XML documents sent via a web protocol which are parsed at the server and contain an instruction to trigger a procedure server-side.

[2]http://ani.cebitec.uni-bielefeld.de/BioIMAX/ [19]

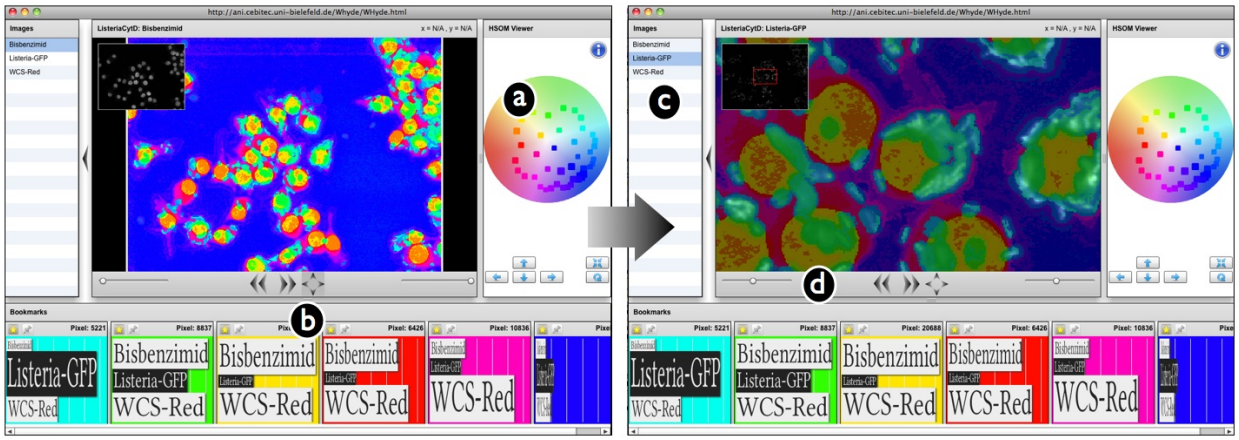

Fig. 4. The result color map for a three component fluorescence micrograph from a study about bacterial infection in cell cultures. Using TICAL, a hierarchical hyperbolic SOM (H2SOM) was trained and the result cluster map is displayed and interactively modified. In a first step, the cluster prototypes are moved across the color plate so a large number of prototypes that represent the background are squeezed into a small part of blue in the color plate (a). In the bottom row (b), some example prototypes are displayed showing the high color contrast for every co-location signal combination of interest. In a next step, one can select one of three component images (c), like the Listeria monocytogenes GFP channel here, continuously zoom into the image (d) and change the opacity to explore the local co-location features in detail.

dimension reduction step two (see above) the three algorithms are treated slightly different. While the dimension reduction for prototypes trained with k-means or neural gas are mapped using a principal component analysis (PCA) the low dimensional coordinates of the H2SOM prototypes are computed from the grid coordinates. Since the dimension reduction is computed on powerful compute servers (see Figure 1) we are planning to implement other dimension reduction algorithms as well, that have been proposed in the last years such as t-SNE [22], LLE [23] or ISOMAP [24] which are not well suited to project an entire image within reasonable computation time but are definitely applicable to a comparable low number of $n \cdot 10^2$ to $n \cdot 10^3$ prototypes.

C. Online visual exploration of the result

In the standard setting, the user is logged into BioIMAX and receives a message after the clustering and dimension reduction is completed[3]. To view a clustering result, start the **Data Browser** (or go back to the **Data Browser** if you are still logged in) and activate the **Results** box in the above filter menu. In addition be sure that the pull-down menu **show data from** shows **My Data (default)**. After a few seconds, the clustering result is displayed (with its name assigned to it using the TICAL interface see Figure 2 above) and can be selected and displayed using the tool **WHIDE** on the right (see Figure 3 and 4 for examples).

IV. Results

To demonstrate the usefulness of TICAL we applied TICAL to two data sets described above. To demonstrate TICAL's potential for the analysis of TIS data we show results obtained using the neural gas algorithm with $K = 50$ neurons, i. e. prototypes. For this setting the computation time was 10 minutes for 30 learning epochs. One learning epoch spans n training steps with n as the number of training items, i. e. number of image pixels. The result is shown in Figure 3 in pseudocolor using the approach explained above. Using a slider, the opacity is controlled so that the relation between the pseudocolors to the original image signals can be investigated (here the DAPI signal). One can see the usefulness of the topology preserving strategy in the pseudocoloring, since it preserves many morphological structures which are usually lost when using random color mapping of prototypes [9]. The Figure 3 shows a small subset of selected cluster prototypes in bottom row so the reader can see how the one dimensional changes in the color scale are mapped to changes to changes in the D-dimensional co-location feature space.

To demonstrate TICAL's value even for lower dimensional cases we show results for the listeria infection data set ($D = 3$) with a H^{2}SOM of three layers, including the central node. For this setting, the training time was approx. one hour. In this result displayed in Figure 4 one can see the advantage of using TICAL compared to a standard RGB coloring. The clustering based pseudocoloring enhances the color contrast allowing a quick assessment of different categories of bacterial invasion in cells, expressed by different color patterns.

V. Discussion

The TICAL tool shows to be an efficient interface for users to apply machine learning algorithms to their multivariate image data. Due to our flat data model TICAL is very flexible and can be applied to any stack of aligned grey value images like for instance MALDI images, hyper- or multispectral data. However, a direct visual inspection of the data is necessary to understand the data, i. e. the hidden regularities between morphological structure and co-location. Since PCA is limited to visualize linear structures other dimension reduction techniques (see above) need to be implemented in the future to capture non-linear features in the co-location feature space. One example is already applicable, which is the H^{2}SOM that

[3]Since the reviewers can only use the test account they can not be informed by email so we kindly ask them to just wait a few minutes/hours and come back to BioIMAX afterwards to look for the results.

shows for example good results for both data sets (see Figure 4 for an example from the Listeria infection study and our recent publication for results obtained for TIS data [14]).

VI. Conclusion

TICAL shows how machine learning libraries can be made applicable for a non-expert users by streamlining the process (in this case the application to multivariate bioimage data) and integrating these methods in client server data analysis frameworks. Although this strategy is generally not new and the technical concept contains some state of the art components we believe, that the integration of all the technical components into one web tool is a valuable contribution to the field of bioimage informatics with a focus on the very special group of multivariate bioimages which clearly have the potential to close serious gaps in life science knowledge in the future.

VII. Acknowledgements

We thank Nasir Rajpoot and his group (Computational Biology and Bioimaging Group, University of Warwick) for providing TIS image alignment which is a pre-processing step of crucial importance to any data mining application in multivariate bioimages..

References

[1] I.G. Goldberg, C. Allan, J.M. Burel, D. Creager, A. Falconi, H. Hochheiser, J. Johnston, J. Mellen, P.K. Sorger, and J.R. Swedlow, "The open microscopy environment (ome) data model and xml file: open tools for informatics and quantitative analysis in biological imaging," *Genome Biology*, vol. 6, May 2005.

[2] A. E. Carpenter, "Software opens the door to quantitative imaging," *Nat Methods*, vol. 4, no. 2, pp. 120–1, 2007.

[3] K. Kvilekval, D. Fedorov, B. Obara, A. Singh, and B.S. Manjunath, "Bisque: a platform for bioimage analysis and management," *Bioinformatics*, vol. 26, no. 4, pp. 544–52, 2010.

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

[5] 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.

[6] E. Barash, S. Dinn, C. Sevinsky, and F. Ginty, "Multiplexed analysis of proteins in tissue using multispectral fluorescence imaging.," *IEEE Trans Med Imaging*, vol. 29, no. 8, pp. 1457–62, 2010.

[7] M. Arif, N.M. Rajpoot, T.W. Nattkemper, U. Technow, T. Chakraborty, N. Fisch, N.A. Jensen, and K. Niehaus, "Quantification of cell infection caused by listeria monocytogenes invasion," *J Biotechnol.*, 2011, doi:10.1016/j.jbiotec.2011.03.008.

[8] A Reinert, A Mittag, T Reinert, A Trnok, T Arendt, and M Morawski, "On the quantification of intracellular proteins in multifluorescence-labeled rat brain slices using slide-based cytometry," *Cytometry A*, vol. -, no. -, pp. –, Mar 2011, doi: 10.1002/cyto.a.21047.

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

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

[11] H.J. van Manen, Y.M. 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.

[12] 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.

[13] J Herold, L Herold, S Abouna, S Pelengaris, D Epstein, M Khan, and TW Nattkemper, "Integrating semantic annotation and information visualization for the analysis of multichannel fluorescence micrographs from pancreatic tissue," *Computerized Medical Imaging and Graphics*, 2010, PMID: 19969439.

[14] D. Langenkämper, J. Kölling, 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.

[15] Haim Levkovitz and Gabor T. Herman, "Color scales for image data," *IEEE Computer Graphics & Applications*, vol. 12, pp. 72–80, 1992.

[16] G. Bradski and A. Kaehler, Eds., *Learning OpenCV: computer vision with the OpenCV library*, O'Reilly, 2008.

[17] L. Dagum and R. Menon, "OpenMP: an industry standard api for shared-memory programming," *Computational Science & Engineering, IEEE*, vol. 5, no. 1, pp. 46 – 55, 1998.

[18] P. Trger, H. Rajic, A. Haas, , and P. Domagalski, "Standardization of an api for distributed resource management systems," in *Proceedings of the Seventh IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2007)*, Rio de Janeiro, Brazil, 2007, pp. 619–26.

[19] Christian Loyek, Nasir Rajpoot, Michael Khan, and Tim 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.

[20] T.M. Martinetz, S.G. Berkovich, and K.J. Schulten, "Neural gas network for vector quantization and its application to time-series production," *IEEE Trans on NN*, vol. 4, pp. 558–69, 1989.

[21] J. Ontrup and H. Ritter, "Large-scale data exploration with the hierarchically growing hyperbolic som," *Neural Networks*, vol. 19, no. 6, pp. 751 – 61, 2006.

[22] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–605, 2008.

[23] S.T. Roweis and L.T. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–6, Dec 2000.

[24] J.B. Tenebaum, V. de Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–23, 2000.