

A Web2.0 Strategy for the Collaborative Analysis of Complex Bioimages

Christian Loyek¹, Jan Kölling¹, Daniel Langenkämper¹, Karsten Niehaus², and Tim W. Nattkemper¹

¹Biodata Mining Group, Faculty of Technology, Bielefeld University, Germany

²Genome Research and Systems Biology, Proteome and Metabolome Research, Faculty of Biology, Bielefeld University, Germany

Abstract. Life science research aims at understanding the relationships in genomics, proteomics and metabolomics on all levels of biological self organization, dealing with data of increasing dimension and complexity. Bioimages represent a new data domain in this context, gaining growing attention since it closes important gaps left by the established molecular techniques. We present a new, web-based strategy that allows a new way of collaborative bioimage interpretation through knowledge integration. We show, how this can be supported by combining data mining algorithms running on powerful compute servers and a next generation rich internet application (RIA) front-end offering database/project management and high-level tools for exploratory data analysis and annotation. We demonstrate our system *BioIMAX* using a bioimage dataset from High-Content Screening experiments to study bacterial infection in cell cultures.

Keywords: Life Science, Bioimage Informatics, Data Mining, Exploratory Data Analysis, Information Visualization, High-content screening, Web2.0, Rich Internet Application, Semantic Annotation

1 Introduction

One field of research which is of growing importance regarding the development and application of intelligent data analysis is life science research, combining a multitude of fields such as molecular biology (genomics, proteomics, metabolomics), biophysics, biotechnology, biochemistry, systems biology, biomedicine, etc. The aim is to understand and model the building blocks of dynamic living systems, which are built by entities from different scales (proteins, chemical compounds, cells) and relationships of different kinds and abstraction levels (interacts-with, inhibition/excitation, co-localizes-with, ...). While most of the molecular data has been extracted for homogenized samples, i.e. without any spatial information for the molecular entities, spatial information has been identified recently as one of the last remaining open gaps in systems biology and life sciences, which has to be closed if one wants to render a comprehensive picture of living systems on all levels of biological self-organization [1]. As

a consequence, new bioimaging techniques have been developed and proposed to close this gap, like MALDI imaging or High Content Screening [1]. This new data promises to close many of the aforementioned gaps, but also trigger a new demand for new technologies to analyze this data. For instance image data produced by high-content screenings (HCS) is increasingly getting richer and more complex, since a growing number of variables is associated to each spatial element (i.e. pixel) of the sample. While this is an enormous gain in information (e.g. in pharmaceutical screenings each of the n variables encode a protein of interest or a cell compartment), it is impossible to access, quantify and extract all relevant image information in one session by one researcher. In fact, the images need to be evaluated by researchers from different fields (biophysics, cell biology, chemistry, computer science, statistics, ...) regarding different aspects (image quality/noise, semantics, cell/function classification, staining specificity, statistical significance, ...) and the result of their studies need to be integrated much earlier in research as it is done now in many projects, where researchers from different institutes in different countries meet maybe once a year.

To foster integration of results and views from different aspects of bioimage analysis a new approach is needed, that covers a large variety of bioimage analytics, ranging from manual annotation based on direct visual inspection to full automatic data mining using unsupervised machine learning. Due to the recent developments of web technology, allowing rapid dynamic integration of user generated content into new user-shaped knowledge data bases (such trends are sometimes referred to as Web2.0 or even Science 2.0 [2, 3]) we started the development of a purely web-based bioimage analysis platform which allows the user to apply different analyses to the data, share data and results with other researchers without a complicated and time-consuming act or data modeling. So the aim is not to design a web-based LIMS (laboratory information management system), but to provide a web-based work bench to interpret bioimages within a web-organized project together with a chosen group of other researchers, independent from their whereabouts condition to an internet connection.

Our fully web-based software approach to intelligent data analysis of bioimage data is called *BioIMAX* (**B**io**I**mage **M**ining, **A**nalysis and **eX**ploration) [4], developed to augment both, an easy initial exploratory access to complex high-content image data and the collaboration of geographically distributed scientists. *BioIMAX* was developed as a rich internet application (RIA), i.e. a web application whose performance and look-and-feel is comparable to a standard desktop application, but will mostly be executed in a web browser allowing for platform independency and avoiding additional installation costs. With *BioIMAX*, several types of high-content image data can be uploaded and organized in personalized projects through a simple web-based interface. This allows a rapid data search and retrieval of own datasets and easily supports sharing of data with other collaborating researchers by inviting them to own projects. With the *BioIMAX Labeler*, a graphical and textual annotation tool, the users have the possibility to annotate, discuss and comment specific image regions, e.g. by linking chat-like discussions to image coordinates. In order to initially explore high-content image

data, the *BioIMAX VisToolBox* provides general methods to get an initial visual access to the n -dimensional signal domain of high-content images. Higher level data mining applications (such as dimension reduction or clustering) which are computationally more expensive are triggered and evaluated in *BioIMAX*, but are computed on powerful external compute servers using specialized C/C++ machine learning libraries.

In recent years, several different toolboxes for bioimage informatics have been proposed and we review them briefly here. General imaging analysis tools like ImageJ [5, 6] or ITK [7, 8] aim at providing a large variety of image processing methods for tasks such as registration, filtering, thresholding or segmentation. In contrast, single purpose tools such as CellProfiler [9, 10] focus on special biological or biomedical problems as well as on data from specific imaging techniques. Another group of approaches are meant as general technological platforms to store and organize large amounts of image data in a central repository on a remote server architecture. In addition to the data management, analysis platforms can include selected methods for data visualization, annotation and analysis. One of the first tools published in this context is OME (Open Microscopy Environment) [11]. Bisque [12] is a recently introduced powerful tool, which provides a platform with an automatic 3D nuclei detection or microtubule tracking. Although tools such as CellProfiler, OME or Bisque represent great steps towards improvements in bioimage data analysis, most of them are focussed on particular well defined biological problems and provide specially adapted analysis methods to solve these problems. However, in many cases the analysis goal is vague and little a priori knowledge is available about the underlying data. Thus, it is not clear in advance, which analysis strategy should be applied. This is a general problem in the analysis of high-content bioimage data, which usually needs to be discussed by collaborating researchers from different disciplines. In the context of HCS analysis, especially pharmaceutical HCS, analysis-related decisions increasingly take place associated to a particular region of interest (ROI). Thus, discussion needs to be linked to particular (x,y) -coordinates, which leads to less trivial design issues in database and graphical user interface development. In addition to this, a successful (cross-domain) collaboration is often impeded, since the involved researchers are usually distributed across several research institutes. In the future, we assume more impacts of web technology developments for bioimage analysis. Especially the fact, that the web is getting more collaborative and user-shaped (effects referred to as *Web2.0*) and offers more and more powerful graphics applications, will stimulate new developments such as ours.

As an example, we demonstrate several aspects of *BioIMAX*, which could support the study of bacterial infection of cells with *Listeria monocytogenes*. *BioIMAX* can be accessed at <http://ani.cebitec.uni-bielefeld.de/BioIMAX> with the username "tuser" and the password "test1" for testing purposes.

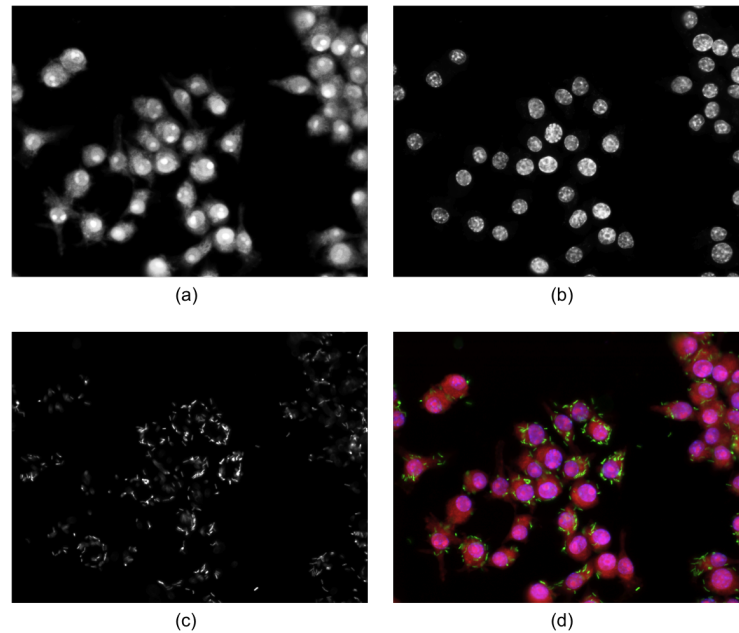


Fig. 1. Example high-content fluorescence image showing infected cells: (a) Cell channel: cytoplasm, (b) Nuclei channel, (c) Listeria channel: GFP stained Listeria and (d) RGB composition of the three channels (a)-(c)

2 Materials

Listeria monocytogenes is an intracellular pathogenic bacterium that causes a food-borne disease called Listeriosis in both humans and animals. Listeriosis is a rare but serious disease with a high overall mortality rate of 30%, most common in pregnant woman or immunocompromised individuals [13]. The bacteria is an important model organism for infection, intracellular proliferation and host-pathogen interactions. Those intracellular bacteria are protected against the host immune system and are poorly accessible for treatment with antibiotics. Therefore, the invasion of the host cells is an important and crucial step in *Listeria* pathogenesis and virulence [14]. In order to study the grade of host cell invasion with *L. monocytogenes*, a high-content screen has been set up using automated microscopy and *L. monocytogenes* expressing the green fluorescent protein (GFP). Figure 1 shows an example high-content image, obtained with the Scan^R screening station (Olympus).

3 Architecture

As previously mentioned, *BioIMAX* software was designed as a rich internet application. The usage of RIAs has several advantages, which meet the neces-

sary requirements for the development of a system like *BioIMAX*. In contrast to conventional thin-client web applications, RIAs provide a richer and more complex graphical interface, resembling desktop applications' interface interactivity and computation power. The RIA technology improves the efficiency of web applications by moving part of the computation, interaction and presentation to the client, thereby lowering the amount and frequency of client-server traffic considerably and permitting asynchronous client-server communication. As a result, the usability of web applications will be improved, annoying installation routines will be avoided and the software will be accessible from any location.

The *BioIMAX* client side was developed with *Adobe Flex* [15], which is an open-source framework for building expressive web applications. RIAs developed with Adobe Flex deploy consistently on all major browsers and operating systems by leveraging the *Adobe Flash Player*. In order to efficiently and consistently manage the data collected, *MySQL* [16] is used as a relational database management system. The communication between the Flex client and the server-side database is realized by using *AMFPHP* [17], which is one of the fastest client server communication protocol available to Flash Player developers.

Once a user has been authenticated by a username and password login procedure, she/he is presented with the *BioIMAX* start page. The start page is designed in the style of a social media platform, creating a personalized environment, which provides, e.g. access to the system-internal mail box or a navigation panel for general data handling such as image upload, project management or access to the *BioIMAX* data browser. With the data browser the user can search and browse the *BioIMAX* database. In addition to visualizing and managing the search results, the data browser serves as starting point for all data exploration tasks. In the following, we give a detailed description of the *BioIMAX VisToolBox* for visual data exploration tasks and the *BioIMAX Labeler* for semantic image annotation.

4 Visual data exploration

The *BioIMAX VisToolBox* (illustrated in figure 2) provides a set of methods to explore and analyze the signal domain of high-content images. The graphical display of the *VisToolBox* is divided into two panels. One panel contains an image viewer, especially designed for high-content or multivariate images (see figure 2(a)). The image viewer includes basic functions such as zooming or panning for image navigation purposes and allows scrolling through a stack of high-content images, image by image. The other panel (see figure 2(b)) comprises several methods from the fields of visualization, co-location analysis and exploratory data analysis (EDA), chosen by tabs, which will be described more detailed in the following.

Image comparison: This tool provides two different methods to compare up to three single image channels of a high-content image simultaneously on a structural/morphological level. The first method is called *Alpha blending* and aims

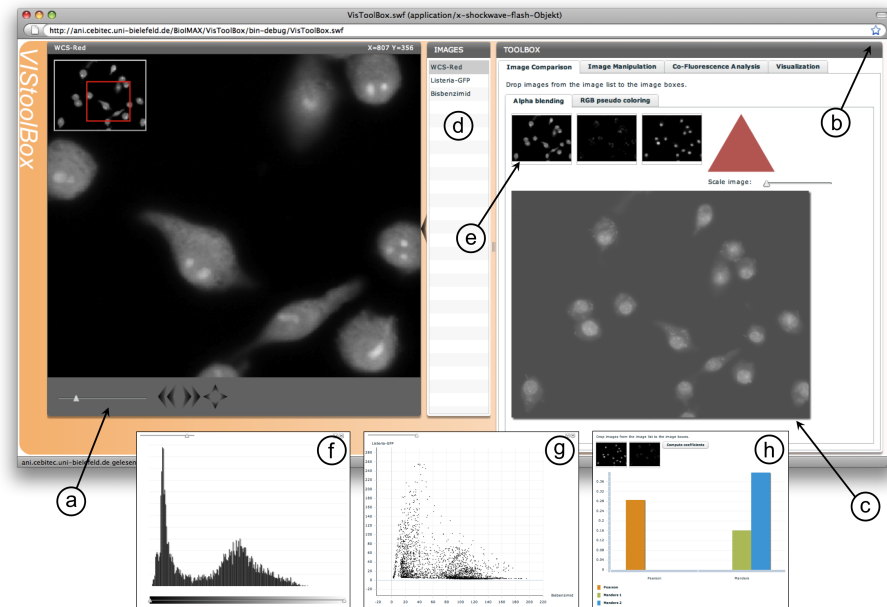


Fig. 2. Screenshot of the *VisToolBox*. This tool provides several methods to explore and analyze the signal domain of high-content images. It consists of an image viewer (a) and a panel (b) including methods from the fields of visualization, co-location analysis and exploratory data analysis (EDA), separated by tabs. In (c), three selected images can be visualized and compared simultaneously by adjusting the opacity of the respective images. By moving the mouse cursor over the red triangle, the opacity value of each single image will be adapted in real-time depending on the distance to the corners of the triangle, which represents the three selected images. Selection of images will be done consecutively per drag-and-drop from the image list (d) to one of the boxes displaying the letters A, B or C (e). The small figures below exemplary show further exploration displays: (f) Histogram and (g) Scatter plot. In (h) a co-location study of two images with statistical measurements (Manders' score and Pearson correlation coefficient) is displayed in an bar chart.

at comparing three images while superimposing them as layers and manually adjusting the opacity value of the respective layers by moving the mouse cursor over the opacity triangle (see figure 2(c)). This can be a useful tool, e.g. while evaluating analysis methods such as segmentation methods regarding their accuracy. Thus, the user can detect structural differences or similarities between the selected images. The purpose of the second method (*RGB pseudo coloring*) is to generate a pseudo color fusion image from three selected images, by interpreting each image as one color channel in a RGB image.

Image Manipulation: This tab includes two histogram dialogs, which display information about the statistical distribution of grey values in the currently se-

lected image. Both histograms are interactive, i.e. the user can manipulate the distribution and the visualization on the left is adapted in real-time. In the histogram the user can filter out irrelevant / wrong signals or study various thresholds needed for analysis tasks.

Co-Fluorescence analysis: Here, the user can compare two selected images on a statistical level by calculating (i) the Pearson correlation coefficient or (ii) the Manders' score, which is a frequently used index for co-location studies in fluorescence microscopy [18]. The results are displayed in a bar chart (see figure 2(h)).

Gating, Link and Brush for in-depth visual exploration: The last part of the *VisToolBox* allows a more detailed exploration of specific image regions. Here, the user can focus the study of *L. monocytogenes* invasion on a single cell level, e.g. to examine cell invasion in the nucleus (see figure 3). For this purpose, the user first has to select a region of interest (ROI) by drawing a rectangle on the displayed image in the image viewer. In a next step the user chooses one of the three visualization techniques at the top of the *Visualization* window, which opens a new plot dialog. Dependent on the chosen dialog, the user will be asked to drop one or more images from the image list to the dialog. After that, all pixels within the ROI will be displayed in the respective plot, i.e. a histogram, scatter plot or parallel coordinates. Selection of points in one plot triggers highlighting the referring pixels in the image on the left (for detailed description see figure 3). This process can also be referred to as "gating" or "link-and-brush" [19].

Clustering and dimension reduction: Since with the growing number of variables (i.e. grey values) a visual inspection using the above techniques will only give a limited view on the image data and the complex high-dimensional manifold given by its n -variate features, i.e. pixel values. As a consequence one is interested in using methods from unsupervised learning to reduce the complexity of the data so it can be visualized, like clustering to reduce the number of patterns to be assigned to graphical parameters (such as colour) or dimension reduction to reduce the number of variables directly. In the design of BioIMAX we created an interface that allows the integration of such algorithms so these can run on remote compute servers and write the results into the BioIMAX data base. For each algorithm at least one individual tool is integrated as well, so the user can start the computationally expensive methods from the web interface, wait for the results and can inspect these again through the web inside BioIMAX. In figure 5 we show an example result obtained with the clustering tool TICAL (Toolbox for Image Clustering ALgorithms, currently in alpha release stage) for one of our images.

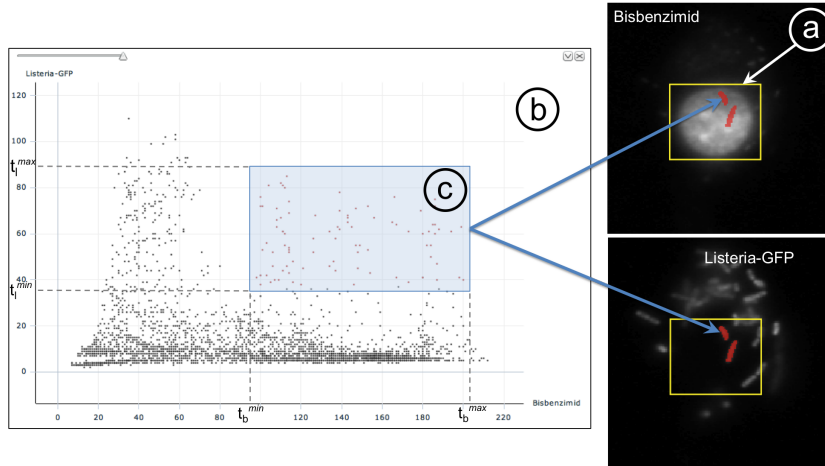


Fig. 3. Interactive exploration of bivariate data from a selected region of interest (ROI) in the image viewer (a). For the ROI, the user selects two image channels (here the nucleus channel (Bisbenzimid) and the GFP-marked *L. monocytogenes* channel) and one tool, e.g. a scatter plot (b) from the visualization tab (see figure 2). The pixel values corresponding to the same location within the ROI are displayed as points in the scatterplot. Selection of points \mathbf{x} in the plot (c) triggers highlighting the referring pixels in the image (displayed as red regions superimposing on the original image), with respect to the following criterion: $\Gamma = \{ \mathbf{x} | t_b^{min} \leq b(\mathbf{x}) \leq t_b^{max} \wedge t_l^{min} \leq l(\mathbf{x}) \leq t_l^{max} \}$, with Γ describing the selection of points \mathbf{x} in the scatter plot, t_b^{min} and t_b^{max} defines the minimum and maximum of the selection range regarding Bisbenzimid values and $b(\mathbf{x})$ is the Bisbenzimid value of point \mathbf{x} . The same applies to the *L. monocytogenes* values, accordingly. This process is often referred to as “gating” or “link-and-brush”.

5 Semantic image annotation

The *BioIMAX Labeler* tool allows one to graphically annotate image regions in single image channels. The interface provides the image viewer described before on the left and an options toolbar on the right (see figure 4(a,b)). In the toolbar the user can adjust several label properties like geometry, color or size before labeling (see figure 4(c)). Furthermore, the user can select specific semantic label types, which will be textually associated to the label. The user can choose one of the label types from the predefined semantic categories, e.g. *Cell* or *Cellular compartment* or she/he can create own label types or categories (see figure 4(d,e)). By clicking in the currently selected image, the annotations are placed as graphical objects on an invisible layer belonging to each single image, allowing for easy modification existing labels, e.g. reforming, recoloring or resizing. A set of labels can be stored into the database by saving all parameters for each single label, i.e. location, type, size, color and form, and will be linked to the respective image channel. In order to get and modify detailed information about single labels, the user can open the annotation/info window by selecting the

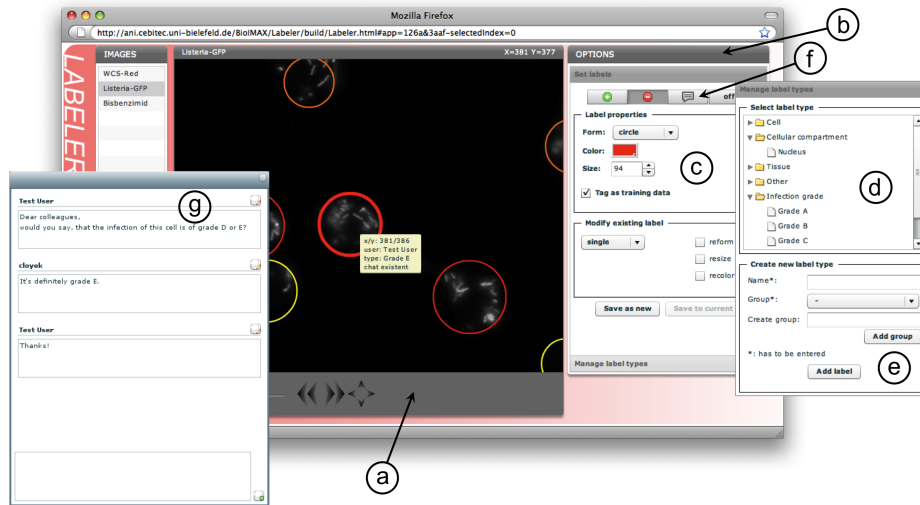


Fig. 4. Screenshot of the *Labeler*. It consists of the image viewer on the left (a), on which the user can place a number of graphical objects (called labels), to annotate specific image regions. On the right, the options toolbar (b) provides options for adjusting several label properties such as geometry, color or size (c) and the possibility to link specific semantic label types to single annotations, which can be predefined types selected by (d) or newly generated types (e), depending on the current scientific context. In order to initiate a discussion about a specific label on a higher semantic level, the user can invoke an annotation/info window by selecting the toggle button (f) at the top of the toolbar. The annotation/info window provides an option to start or open an existent chat-like discussion about an label (g).

toggle button showing the callout icon. With the *Labeler* tool we are aiming at two goals. First, user shall be enabled to label interesting image regions, which can be important in quantification and evaluation tasks. In this study, the experts have to annotate cells in a large number of images into different semantic categories. With the *Labeler* the experts can define and insert new label types representing different infection grades (see figure 4(d)) and can start labeling cells, e.g. using circles with different colors, each color representing a specific infection grade. Using the *Labeler*, the process of establishing a gold standard from several experts is speeded up and simplified, e.g. there is no need to transfer multiple copies of images to the experts. The users can easily login to the *BioIMAX* system and can immediately start labeling from any location and all label results will centrally be stored in the database and can be inspected by all collaborating researchers at any time.

Second, we want to link chat-like discussions to image regions to link high-level semantics to morphological features. Therefore, the *Labeler* provides a chat window (see figure 4(g)), which can be accessed from the annotation/info window. Here, several users can communicate about the selected label and the con-

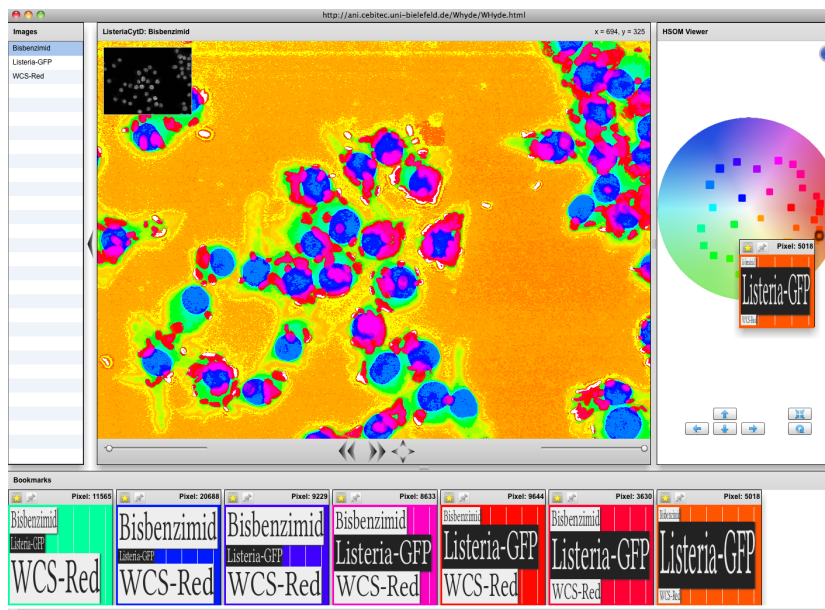


Fig. 5. A clustering based visualization of one three-dimensional HCS bioimage. Combining clustering with dimension reduction can be done by a) using a self-organizing map or b) a combination of other vector quantization algorithms (k -means, neural gas) with dimension reduction techniques (PCA, LLE, t-sne). Both approaches allow the mapping of cluster prototypes to colors which is used to colorize each pixel applying the best matching criterion to the pixel and all the cluster prototypes. In the middle of the display, one can see the result pseudo color image. On the bottom, a small number of clusters has been chosen in the overview (right panel) and displayed. The lengths of the horizontal boxes display the average signal intensity in that cluster.

versation will additionally be stored together with the label. This facilitates Web2.0 style collaborative work on one image, while the stored states of communication content are directly linked to image coordinates/ROIs. While developing new analysis strategies for high-content image data, researchers have to discuss aspects about the *original data*, e.g. the trustworthiness of signals, and about *analysis methods*, e.g. the quality of intermediate results such as registration or segmentation. Figure 6 illustrates both scenarios.

6 Discussion

In this paper we proposed a Web2.0 approach for the collaborative exploration of high-content screening bioimages in life sciences and demonstrated its application with an example dataset from *Listeria monocytogenes* cell invasion analysis. Due to the complexity of high-content image data, the extraction and quantification of all image information and the generation of analysis strategies is a

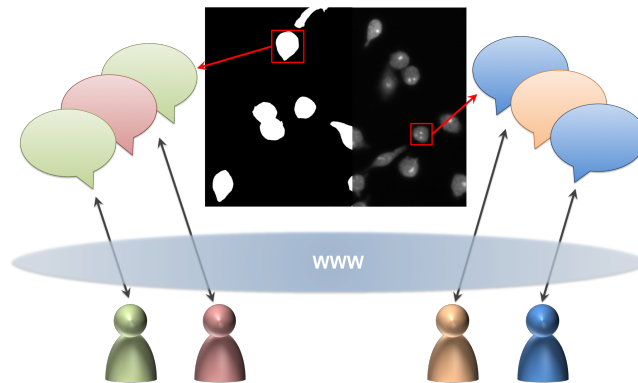


Fig. 6. Illustration of a chat-like discussion about image regions with the *BioIMAX Labeler*. The figure demonstrates two possible discussion scenarios based on the same image data: discussion about the raw image and discussion about analysis methods or results.

difficult task for researchers and different aspects need to be discussed by collaborating researchers from different disciplines. Thus, we presented fully web-based tools, which support both, exploratory analysis of high-content image data and important collaborative aspects in *Listeria monocytogenes* infection analysis.

The *BioIMAX VisToolBox* provides methods from the field of exploratory data analysis, in order to gain initial insights into the structural characteristics of the underlying data, following Ben Shneidermans information visualization mantra: Overview first, zoom in and filter, details on demand. The concept of *VisToolBox* does not include predefined analysis pipelines regarding a special biological question in the form of a black box model, which gets an image as input and the user is presented with the finalized result. With the *VisToolBox*, the user is directly involved in the knowledge discovery process, while exploring the data space themselves with specific information visualization techniques. This is an important strategy in the field of visual data mining and exploration [20]. Using the clustering tool TICAL, even higher dimensional image data can be visually explored without the need to install a machine learning toolbox on ones desktop, since *BioIMAX* allows the application of clustering independent from the users whereabouts, condition to an internet connection.

The *BioIMAX Labeler* provides tools to communicate and discuss about specific image regions, which is of great value, since analysis-related decisions are increasingly associated to particular regions of interest. The *Labeler* allows to annotate image regions with graphical objects and to link chat-like discussions representing high-level semantics to morphological features.

Since *BioIMAX* is designed as a rich internet application, one of the key feature is, that a user only needs a login and a password to get access to the *BioIMAX* platform provided an internet connection is available. Except for the installation of the Flash Player, which is available for most browsers, no ad-

ditional software packages and libraries have to be installed. The fact that all collaboration and exploration tasks will be performed within one web-based platform is of great value, since it simplifies and speeds up several aspects in the analysis process, e.g. avoiding transfer of data between researchers, since all researchers work on the same copy centrally stored in the *BioIMAX* database.

We believe, that in the age of the ongoing development of web technologies, our Web2.0 approach is an important step forward, to support complex analysis tasks regarding high-content data. Such an approach is of particular benefit to those scientific projects, where several scientists from different institutes at different locations are involved and has to collaborate.

References

1. Megason, S.G., Fraser, S.E.: Imaging in Systems Biology. *Cell*. 130(5), 784–795 (2007)
2. Shneiderman, B.: Science 2.0. *Science*. 319, 1349–1350 (2008)
3. Waldrop, M.M.: Science 2.0 - Great new tool, or great risk? *Scientific American*. (January 9, 2008)
4. Loyek, C., et al.: BioIMAX: A Web 2.0 approach for easy exploratory and collaborative access to multivariate bioimage data. *BMC Bioinformatics*. 12, 297 (2011)
5. Image Processing and Analysis in Java, <http://rsbweb.nih.gov/ij/>
6. Abramoff, M.D., Magelhaes, P.J., Ram, S.J.: Image Processing with ImageJ. *Biophoto. Int.* 11(7), 36–42 (2004)
7. Insight Segmentation and Registration Toolkit (ITK), <http://www.itk.org>
8. Yoo, T.S., et al.: Engineering and Algorithm Design for an Image Processing API: A Technical Report on ITK - the Insight Toolkit. In: Westwood, J (ed) *Proceedings of Medicine Meets Virtual Reality*. IOS Press, Amsterdam, pp. 586–592 (2002)
9. Carpenter, A.E., et al.: CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* 7(10), R100 (2006)
10. Lamprecht, M.R., Sabatini, D.M., Carpenter, A.E.: CellProfiler: free, versatile software for automated biological image analysis. *Biotechniques*. 42, 71–75 (2007)
11. Swedlow J.R., et al.: Informatics and Quantitative Analysis in Biological Imaging. *Science*. 300, 100–102 (2003)
12. Kvilekval, K., et al.: Bisque: a platform for bioimage analysis and management. *Bioinformatics*. 26(4), 544–552 (2010)
13. Ramaswamy, V., et al.: Listeria - review of epidemiology and pathogenesis. *J. Microbiol. Immunol. Infect.* 40(1), 4–13 (2007)
14. Ireton K.: Entry of the bacterial pathogen *Listeria monocytogenes* into mammalian cells. *Cell Microbiol.* 9(6), 1365–1375 (2007)
15. Adobe Flex. <http://www.adobe.com/products/flex/>
16. MySQL. <http://www.mysql.com/>
17. AMFPHP - Action Message Format PHP. <http://amfphp.sourceforge.net/>
18. Manders, E., et al.: Dynamics of three-dimensional replication patterns during the S-phase, analysed by double labelling of DNA and confocal microscopy. *J. Cell Science*. 103, 857–862 (1992)
19. Ware, C.: *Information Visualization - Perception for Design*, Morgan Kaufmann Publishers Inc. San Francisco (2004)
20. Keim, D.A.: Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and Computer Graphics*. 7(1), 100–107 (2002)