BioIMAX

A Web2.0 Approach to Visual Data Mining in Bioimage Data

Der Technischen Fakultät der Universität Bielefeld

vorgelegt von

Christian Loyek

zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften

Bielefeld, Germany, January 15, 2012

Abdruck der genehmigten Dissertation zur Erlangung des akademischen Grades *Doktor der Ingenieurwissenschaften* (Dr.-Ing.)

Der Technischen Fakultät der Universität Bielefeld

- am 15.01.2012 vorgelegt von Christian Loyek
- am 17.08.2012 verteidigt und genehmigt

Gutachter

- Prof. Dr. Tim W. Nattkemper, Universität Bielefeld
- Prof. Dr. Karsten Niehaus, Universität Bielefeld

Gedruckt auf alterungsbeständigem Papier nach ISO 9706

Acknowledgments

This work was carried out in the Biodata Mining and Applied Neuroinformatics Group headed by Prof. Tim W. Nattkemper, at the Faculty of Technology, University of Bielefeld. It was supported by a grant of the rectorate of the Bielefeld University and a scholarship of the Genome Informatics Group, headed by Prof. Jens Stoye.

First of all, I would like to thank Tim for his constant and encouraging support and excellent supervision throughout all stages of my work. His inspiring ideas, coupled with his humour and enthusiasm, had an invaluable impact and pushed this work forward.

The Biodata Mining & Applied Neuroinformatics Group, as well as the Neuroinformatics Group and Genome Informatics Group, were always excellent and pleasant places to work, with a friendly atmosphere amongst all colleagues. I especially want to thank Julia, who shared an office with me for a long time. I always enjoyed our scientific conversations and our amusing collaboration! I also would like to thank Jörg, Timm, Niko, Daniel, and Jan for numerous and fruitful discussions on *BioIMAX* and related topics. Also to all students who, under my supervision, made important contributions to this work as part of various student projects - thank you.

A special thanks goes to Berni, Jochen, Johannes, Kolja and Sebastian for proofreading parts of this manuscript and to the whole PAX for simply being the PAX.

Finally, I would like to thank my family and my wife Jenny. Without their invaluable support and, most of all, their unconditional patience over the years, this work would not have been possible.

Originality of the work

The work and results presented in this thesis were conceived and carried out by myself under supervision of Tim W. Nattkemper. Some of the modules of the *BioIMAX* software have been implemented within student projects (392200 Cell Screener (Pj) (WS 2009/2010)) under my supervision. Afterwards, I have corrected, optimized, and redesigned substantial parts of these modules, extended them with further functionalities, and merged them to a complete software solution, as it is presented in this thesis.

To my parents and my wife.

Contents

Publications

Parts of this thesis have been published in:

- Loyek C, Rajpoot NM, Khan M, Nattkemper TW. *BioIMAX: A Web2.0 approach for easy exploratory and collaborative access to multivariate bioimage data.*
  BMC Bioinformatics, 12:297, 2011
- Loyek C, Kölling J, Langenkämper D, Niehaus, K, Nattkemper TW. *A Web2.0 Strategy for the Collaborative Analysis of Complex Bioimages.*
  Advances in Intelligent Data Analysis X, 258-269, 2011
- Loyek C, Bunkowski A, Vautz W, Nattkemper TW. *Web2.0 paves new ways for collaborative and exploratory analysis of chemical compounds in spectrometry data.*
  Journal of Integrative Bioinformatics, 8:158, 2011
- Loyek C, Bunkowski A, Vautz W, Nattkemper TW. *A Web2.0 Collaborative Analysis of Ion Mobility Spectrometry Data.*
  International Symposium on Integrative Bioinformatics, Wageningen, 2011
- Herold J, Loyek C, Nattkemper TW. *Multivariate Image Mining*
  Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011
- Loyek C, Woermann FG, Nattkemper TW. *Detection of Focal Cortical Dysplasia Lesions in MRI Using Textural Features*
  Proceedings of BVM, 61-65, Berlin, 2008

CHAPTER 1

Introduction

Life science research aims at understanding the relationships in genomics, proteomics and metabolomics on all levels of biological self organization, combining a multitude of disciplines such as molecular biology, biophysics, biotechnology, biochemistry, systems biology or biomedicine. The major goal is to understand and model the building blocks of dynamic living systems, which are built by entities from different scales (chemical compounds, proteins, cells) and relationships of different kinds and abstraction levels (interaction, inhibition/exication, co-localization).

In the last decades, an enormous gain of knowledge about cellular components and their functions has been obtained by established molecular techniques, ranging from genomics to proteomics and metabolomics. However, these techniques lack the ability of revealing the spatial and temporal organization of the molecular components, which has been identified recently as one of the last remaining open gaps, which has to be closed, in order to get a comprehensive picture of living systems on all levels of biological self organization (Megason and Fraser, 2007).

As a consequence, innovative biological and biomedical imaging technologies, like MALDI imaging or various multi-tag fluorescence imaging modalities, have been developed and proposed in the last two decades offering substantial insights into the spatial organization of molecules forming existential complex regulatory networks in living systems (Megason and Fraser, 2007). This data acquired by such new bioimaging technologies promises to close the aforementioned gap, but also poses new demands on information technology approaches to analyze this data. Since this image data is getting richer and more complex, e.g., a growing number of variables is recorded for each spatial location of the sample, which is an enormous

gain in information, researchers are faced with the open question of how to discover and extract the valuable knowledge, in order to generate new hypotheses and come to scientific findings. In addition to the existing vital and expanding field of image processing, novel information technology approaches have to be developed to extract, compare, search and manage the data and the knowledge inherent in complex bioimage data, resulting in an emerging new engineering area called *bioimage informatics* (Peng, 2008).

Images by their very nature are semi-structured, since single image elements, i.e., pixels, themselves have no direct semantics. They only represent a gray value at a specific position and it is generally the objects consisting of a group of adjacent pixels that contains semantics. Due to the challenge of extracting image semantics in combination with the increased complexity of modern bioimage data the analysis goal is often vague and little *a priori* knowledge is available for the underlying data. Thus, researchers need an initial exploratory access to the image information, e.g., image regions, in a fast and intuitive way to aid the process of early steps in analysis and knowledge discovery, i.e., forming a mental model for the data.

Additionally, the interpretation and exploration of bioimage data poses great challenges at different levels in data analysis. Due to the increased information content inherent in bioimages, it is impossible to access, quantify and extract all relevant image information in one session by one researcher. In fact, images need to be evaluated by researchers from different disciplines, like biophysics, cell biology, chemistry, computer science or statistics, regarding different analysis aspects such as image quality/noise, semantics, cell classification or statistical significance. The results and findings of these experts need to be integrated much earlier in the research process as it is done nowadays in many scientific projects, since the experts are usually spread across several geographically distributed research institutes. Thus, collaboration plays an increasing role in bioimage research projects.

As a consequence, a new approach is needed, that first, covers a large variety of bioimage analytics and second, fosters the integration of results and perspectives from different aspects in bioimage analysis. Due to the recent developments in Web technology, the Web is getting more collaborative and user-shaped, allowing for rapid integration of user-generated content into large knowledge databases, which is a trend referred to as *Web2.0*. Internet users more and more push data to the Web, in order to share data and information with other users, e.g., through *Social Networking platforms*, or to manage and organize personal or business data such as documents, calendar, address book, presentations and many other information on remote server architectures, a trend recently termed *Cloud computing*, so their data is accessible via the Web from any location at any time with various network devices such as smart phones, tablet PCs, netbooks or laptops.

Furthermore, current trends regarding the development and deployment of powerful and rich Web applications can be observed, that are often called *Rich Internet Applications* (RIA). RIAs are Web applications whose performance and look-and-feel is comparable to standard desktop applications, offering powerful graphical user interface capabilities with sophisticated visualization and interaction functionalities. RIAs are running in a Web browser allowing for platform independency and avoiding installation and maintenance costs. In recent years, many RIA frameworks have been proposed such as Adobe Flash/Flex, Microsoft Silverlight or HTML5 that fosters the development of such desktop-like Web applications.

In view of these observations, a novel fully Web-based software approach for an intelligent data analysis of bioimages, called *BioIMAX* (**BioI**mage **M**ining, **A**nalysis and e**X**ploration) (Loyek et al., 2011c) is presented and discussed in this thesis. *BioIMAX* is the attempt to explore the potential of social network technologies in the context of bioimage analysis by combining the Web's lightweight collaboration and distribution architecture with the interface interactivity and computation power of desktop applications using modern RIA technologies. *BioIMAX* was developed to augment both, an easy initial exploratory access to complex high-content bioimage data through a large variety of bioimage analytics, ranging from manual annotation based on direct visual inspection to fully automatic datamining methods, and important collaborative aspects in data analysis of geographically distributed researchers from different disciplines. Such an Internet-based scientific collaboration is referred to as *Science2.0* (Shneiderman, 2008; Waldrop, 2008).

1.1 Organization of the Thesis

Chapter 2 provides a brief introduction to bioimaging focussing on multivariate imaging techniques and presents the basic cornerstones of the emerging new research area of bioimage informatics. Chapter 3 gives an overview about state-of-the-art bioimage informatics tools and software solutions followed by a discussion about their advantages and drawbacks, which leads to the motivation and the goals for the realization of the *BioIMAX* system presented in this thesis. In Chapter 4 the necessary requirements posed on the design and implementation of the *BioIMAX* system are described and discussed. Based on these requirements the architectural aspects of the *BioIMAX* system are depicted in Chapter 5, starting with the definition of an appropriate data model in Chapter 5.1, followed by the characterization of the technical design of the system in Chapter 5.2 including a section, which outlines current trends of Web technologies and Rich Internet Applications. Chapter 6 focusses on the realization and implementation of the different *BioIMAX* components showing several screenshots of the graphical user interface. This chapter generally follows a possible workflow from a potential user perspective. As an illustration of the applicability and usefulness of *BioIMAX* Chapter 7 outlines three potential application cases in recent bioimage analysis problems. The thesis ends with a discussion about the essential cornerstones and their benefits and drawbacks in Chapter 8 and a short conclusion and outlook in Chapter 9.

CHAPTER 2

Bioimaging

Historically, biologists have tried to understand organisms, starting with the investigation of smaller entities of organisms to gain an understanding of the larger concepts. With the sequencing of the first whole genome in the mid-nineties, the era of genomics had started. Since then, a large number of high-througput technologies, together with transcriptomics, proteomics and metabolomics methods (the so called "omics" for short) have paved the way for identifying cellular components and their functions on a large scale. However, these approaches only provide the parts list (i.e., DNA sequence, mRNA and proteins), but the next challenging question is, how these parts interact as a system and how does this system function to create an organism (Megason and Fraser, 2007). Due to this question, a shift of focus in molecular biology could be observed in recent years, from molecular characterization to the understanding of functional activity (Wolkenhauer et al., 2003). This was the birth of a growing new research area in the post-genomic era, which is referred to as *systems biology*. The aim of systems biology is to understand and to model the dynamics and structure of complete biological systems by determining how the single components interact with each other to form the complex regulatory networks underlying fundamental processes of life and disease (Zimmermann et al., 2003).

The conventional omics fields are still important in systems biology, since they continuously producing a wealth of basic molecular knowledge such as discovering new proteins and their functions. However, they only provide a rough idea of the components functions and interactions within the living organism, i.e., they do not directly show whether the components have a functional role in the cellular process that is under investigation (Pepperkok and Ellenberg, 2006). They largely suffer from the problem of not being able to provide complex temporal

or spatial information of molecular inter-relationships as they include important parameters to identify functional molecular networks in intact cells or organisms (Starkuviene and Pepperkok, 2007). Thus, novel techniques complementary to the traditional omics approaches are required, which capture data approaching the same large scale as omics approaches, but with enhanced temporal and spatial resolution.

This is the place where imaging comes in and can play an important role in systems biology (Megason and Fraser, 2007). The recent developments of modern molecular bioimaging techniques allow insight into the spatiotemporal organization of individual molecules. In particular, quantitative fluorescence microscopy has become one of the tools in large scale systematic analysis of protein function, not at least due to the availability of the green fluorescent protein (GFP) and its spectral variants (Pepperkok et al., 2001). Fluorescent imaging techniques provide information about protein localization and dynamics as well as protein-protein interactions at the single cell or even sub-cellular level imaged and quantified in living cells and organisms. Fluorescence microscopy and related techniques have proven to be complementary to traditional omics approaches to gain insights into the regulatory activity of cells, tissues, and whole organisms as well as to study diseases and to identify drug targets.

Data acquired by molecular bioimaging techniques is of high information content, often represented by multidimensional and multiparameter datasets, which are usually referred to as *high-content images* (HCI) or *multivariate images* (MVI) data. Since the focus of this work is placed on the handling of these types of image data, a short definition and an introduction in basic multivariate imaging modalities will be given in the following. Afterwards, Section 2.2 gives an overview about the basic cornerstones regarding the development of software solutions in the emerging field of bioimage informatics.

2.1 Multivariate Images

For the construction of a multivariate image (MVI) an arbitrary number of different signals is recorded for each spatial location (pixel) of the sample reflecting different characteristics of the sample. This results in n single images or channels that are combined to a stack of images $\mathbf{I}$ as illustrated in Figure 2.1. Images I_k in one stack are congruent, i.e., that for each pixel in one image there is a corresponding pixel in the other image(s) that can be referred to as the same position in the sample (Geladi and Grahn, 1996). In order to achieve congruence, images in $\mathbf{I}$ often have to be aligned to each other regarding a reference image by applying registration algorithms. The exact alignment of images is a prerequisite for meaningful MVI analysis. A pixel $\mathbf{p} = (x, y)$ in an MVI is associated to a multivariate signal vector $s(\mathbf{p}) = (s_1, s_2, s_3, ..., s_n)$ that can be considered as a point in the n-dimensional signal space. In addition to the signal domain, MVIs also describe structures in the spatial domain, e.g., morphological or geographical shapes (Herold et al., 2011). In the literature, several imaging setups and techniques are proposed to generate MVIs allowing for observing different characteristics of a sample and hence, addressing different biological questions. In the following, a short overview about basic multivariate imaging strategies, i.e., multispectral,

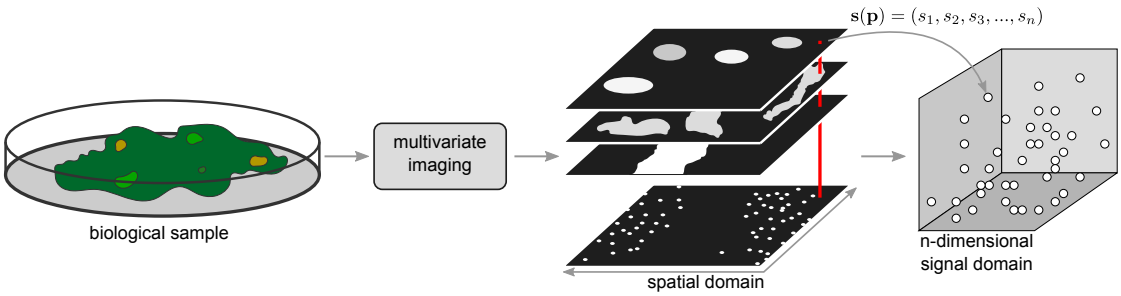


Figure 2.1: Multivariate image acquisition of a biological sample. A multivariate signal vector $s(\mathbf{p}) = (s_1, s_2, s_3, ..., s_n)$ is associated to each pixel $\mathbf{p} = (x, y)$ in the spatial domain, which is usually a regular grid. The signal vectors s are considered as data points in the n-dimensional signal domain.

multifluorescence and multimodal imaging, will be given.

2.1.1 Multispectral imaging

Multispectral imaging records image data at specific frequencies across the electromagnetic spectrum and has been widely used for satellite and air-borne remote sensing in astronomy and geology, but is a relatively novel technique in microscopy. Multispectral imaging techniques produce a set of images or channels from a biological sample representing intensity at each pixel as a function of wavelength (Levenson and Hoyt, 2000). The signal s_i ($i \in [1, n]$) reflects the intensity of a pixel at a desired wavelength λ_i and the signal vector $s(\mathbf{p})$ of a pixel can be regarded as its spectral signature. The resulting MVI stack contains spectral as well as spatial information. The spectral information, for example, can be used for classifying each pixel in an image according to its spectral signature, e.g., for the differentiation of cell types (Angeletti et al., 2005). In general, a standard RGB image can be considered as an elementary spectral image, containing only three bands and is comparable to the spectral resolving power of the human visual system. In most modern applications spectra are measured over a wide wavelength range with small increments capturing differences in color of the sample that may be overlooked by the naked eye. Multispectral imaging is applied in a large variety of biological experiments, especially in live cell imaging (Zimmermann et al., 2003; Hiraoka et al., 2002).

2.1.2 Multifluorescence imaging

In contrast to multispectral imaging, which records one sample over multiple wavelengths, in order to get detailed information about the spectral characteristics of a specimen, multifluorescence imaging aims at precisely visualizing the location of molecules in a sample (Herold et al., 2011). Multifluorescence imaging techniques make use of antibodies that are specific for biological molecules, i.e., antigens, of interest. Using specific antibodies conjugated with a fluorescent dye ensures that staining is limited to the location of the corresponding antigen in the specimen imaged with fluorescence microscopy. Such highly selective data allows the

study of the spatial distribution of molecules, their co-location characteristics and in this way molecular networks, e.g., to survey cancer versus normal cellular phenotypes in pathological tissues. The resulting signals s_i ($i \in [1, n]$) represent the intensity of a pixel for one specific molecule m_i. Due to the ever-growing variety of specific antibodies, the number of molecules that can be selectively labeled has been increased in recent years. However, only few molecules can be detected simultaneously in one sample due to the spectral limitation of common microscopy techniques (Zimmermann et al., 2003). Spectral limitation depends on the fluorescence excitation and emission spectra of each fluorophore, which often results in spectral overlap between multiple fluorophores and makes them difficult to separate. According to (Murphy, 2006) the maximum number of distinct fluorophores that can be distinguished in one specimen is around ten. In order to avoid these limitations, novel imaging technologies and automated microscopy devices have been developed in recent years that can sequentially label hundreds of distinct molecules in the same sample (Schubert et al., 2006; Micheva and Smith, 2007).

2.1.3 Multimodal imaging

The multivariate imaging setups discussed so far apply the same imaging modality to capture different characteristics of the sample. Changes in their parameterization such as recorded wavelengths in multispectral imaging or labeled molecules in multifluorescence imaging leads to a set of n different signals. These imaging setups can be summarized as *intramodular* imaging techniques (Herold et al., 2011; Nattkemper, 2004). Another approach to obtain intramodular MVIs is combining a univariate image with copies of it that are derived by calculating local image characteristics such as texture, shape or statistical features of image regions. Resulting images in the MVI stack are referred to as *feature maps* and are important for segmentation and classification tasks. In contrast to intramodular imaging techniques, multivariate image data can also be acquired by combining images from different image modalities or instruments, which is referred to as *intermodular* imaging. As an example, different microscopy techniques can be mixed like bright field imaging with images obtained by dark field imaging, spectroscopy imaging, and fluorescence microscopy (Cottrell et al., 2007). Another example of combining images from multiple optical imaging modalities is proposed in (Vinegoni et al., 2006). Both, intramodular and intermodular imaging strategies mentioned so far hold the same spatial resolution facilitating the task of aligning and registering the channels of the MVI to each other. However, it is also possible to combine images acquired by different imaging modalities, which are usually based on different physical effects, e.g., optical, electron and ion microscopy. The acquired images often differ in their spatial resolution, so directly mapping the spatial location of the images is hardly feasible and requires sophisticated resampling and interpolation methods to achieve congruence.

2.2 Bioimage informatics

Modern bioimaging systems produce a deluge of complex high-dimensional image data and linked meta data. In particular, large-scale imaging modalities such as high-content screenings

considerably increase the complexity of microscopy datasets. The large amount of data and its high information content pose great challenges for the image computing and bioinformatics community. To be useful to the scientific community, data has to be transformed into appropriate representations that allow the scientists to extract, analyze, manage, search, and share the biological knowledge of the images under investigation. Therefore, in addition to the existing vital and expanding field of image processing, novel information technology approaches have to be developed to fulfill these requirements, spawning an emerging new area of bioinformatics, which is called *bioimage informatics* (Peng, 2008).

Compared to the traditional omics approaches, bioimage informatics is a relatively young research field in bioinformatics. In genome, transcriptome, proteome, or metabolome research, scientists routinely retrieve their data from centralized open access databases for their experiments. As an example, there exist several databases providing genomic or proteomic datasets obtained from various biological organisms and specimens, e.g., GDB (The Human Genome Database) (Letovsky et al., 1998) or UniProt (Universial Proteine resource) (Bairoch et al., 2005). The key to the successful use of these datasets was the development and deployment of Web-based software applications, designed for biologists to search and analyze omics data and to share their discoveries and results with the respective scientific community, e.g., *GenDB*, a genome annotation system for prokaryotic genomes (Meyer et al., 2003), *QuPE*, an integrated bioinformatics platform for the storage and analysis of quantitative proteomics data (Albaum et al., 2009), or *MeltDB*, a software platform for the analysis and integration of data from metabolomics experiments (Neuweger et al., 2008). The question is then raised as to whether this concept is applicable for the bioimage analysis field. In principle, it should be possible. However, omics data is generally well-structured with defined formats, so software tools could be developed based on these standards. In contrast, in bioimaging, standards cannot easily be defined, since the amount of different imaging modalities produce highly unstructured image data in two or more dimensions with varying formats. Another major difference between both fields is, that omics data consist of known identifiers, e.g., nucleotide base pairs in genome sequences, representing unique semantics, whereas the image elements, i.e., pixels, contain no direct meaning, since a pixel initially represents only the intensity of a recorded signal at that spatial location. Semantics in images is given by grouping of a set of adjacent pixels forming objects or regions that represents specific biological structures of the imaged sample such as cells or tissue.

Due to these facts, developing bioimage informatics tools is much more complicated and challenging, as it requires a more sophisticated and complex data handling compared to, for example, genome sequence analysis. In general, while developing novel bioimage informatics algorithms and systems, several informatics aspects and basic requirements have to be taken into account, which will be summarized in the following.

2.2.1 Data management

In general, bioimaging experiments involve several types of data with varying degrees of complexity and format. In addition to the raw images, which by their very nature are complex and multidimensional, single experiments include many additional derived data such

as (pre)processed images, output from specific analysis procedures, and lots of meta data describing different parameters of the data or experimental setups. Typically, these highly interlinked datasets are stored on a hard disk using arbitrary directory structures. Using such kind of data storage often results in difficulties regarding several aspects in data handling. First, searching and retrieval of specific datasets in these directories is a laborious and uncomfortable task, since the scientists have to browse several directories and files by figuring out, which cryptic filenames describing the data content leads to the requested data. Second, adding new data is associated with a considerable amount of work and time. Scientists have to take care that their data is stored at the right place, so that it can be retrieved by themselves and by other scientists involved at any time. Finally, detailed text or spreadsheet files have to be prepared serving as a guideline how these diverse datasets have been generated and in which way they are linked to the original image data and how they are interconnected.

To overcome these problems and limitations, it will be of great value to use a database management system (DBMS) in connection with a centralized data repository, serving as the backbone for the dynamic integration and search of highly interlinked experimental bioimage datasets. Therefore, a considerable effort has to be made, in order to develop an appropriate data model representing the variety of structured and unstructured data and its interconnections. In addition, user interfaces should be provided to support scientists in the task of browsing and retrieving specific datasets. This implies that data in the database is indexed and thus searchable, e.g., by using text descriptors based on a controlled or standard ontological vocabulary.

The quality of an appropriate data model in combination with a DBMS greatly influences all further aspects in bioimage experiments and can be considered as the essential basis for the development and implementation of bioimage informatics tools.

2.2.2 Image processing and analysis

Extracting knowledge from the large amount of complex image data is the key issue in bioimage analysis. Up to the present, a substantial effort has been made to develop a myriad of different methods for image related analysis. Methods from the field of image processing such as registration, filtering, segmentation, and feature extraction can be considered as the classical image analysis part. Image registration is an essential application of digital image processing used to align a number of images with respect to a reference image, e.g., to compare images acquired under different conditions, as it is the case regarding all multivariate imaging modalities. Image filtering is applied to enhance image details or to suppress irrelevant image signals like signals belonging to the background or outliers. Image segmentation is the most basic processing procedure in bioimage analysis. Segmentation methods aim at detecting meaningful regions of interest, e.g., cells or tissue, in the respective images. Segmentation of relevant objects is in many applications a non-trivial task, since the accuracy of the segmentation result depends on several factors such as a low signal-to-noise ratio or the degree of objects variability. Finally, image features describe images, image regions or pixels at a higher level of abstraction by quantifying specific characteristics like statistical, geometrical, or morphological properties. Classical image analysis methods

will be nowadays enriched by analysis techniques from the fields of pattern recognition, data mining, and exploratory data analysis. Automatically phenotyping cells or determination of subcellular locations are only two prominent examples of the amount of applications that requires classification or clustering methods. The aim is to find structures or rules in typically high-dimensional feature spaces. The selection, application and combination of such analysis methods highly depends on the type of image data under investigation and is usually based on defined biological questions and hypotheses. Tools designed to select and initiate image analysis methods should provide easy and transparent user interfaces that allow scientists, even from other fields of expertise, to learn quickly how to use them and to understand the outcomes.

2.2.3 Data visualization and interaction

Visualization in bioimage experiments is closely linked to the image analysis category mentioned before. Scientists need tools, which allow them to navigate efficiently through the wealth of data, in order to identify meaningful characteristics and to explore and interpret relevant relationships between original image data and image-based analysis results. Visualization tools display data types at different levels, ranging from raw and processed images to image-derived quantitative data displayed by several graphical representations typically from the field of information visualization such as scatter plots, histograms or visualization of cluster prototypes. To gain insight into potential correlations between these data types, visualization tools should provide additional functionalities allowing for browsing one data type linked to the others in an interactive and intuitive way, e.g., by selecting interesting subsets of data in one visualization, which triggers highlighting of corresponding data in other visualizations. This process can be referred to as "gating" or "link-and-brush" (Becker and Cleveland, 1987) and is frequently used to filter data for further analysis. Linking to the original data is of particular importance, since there is often no obvious a priori link between quantitative image descriptors and biological meaning (Walter et al., 2010). Finally, if available, image acquisition parameters or biological information about the imaged biological sample should also be visualized in the form of additional meta data.

2.2.4 Data sharing and scientific collaboration

Besides the aspects mentioned so far, data sharing and collaboration is getting more and more important in bioimage informatics and in life science projects (Vicens and Bourne, 2007), since community-driven and distributed research leads to new discoveries and knowledge, as it has been the case in genomics through the public availability of genome sequences. Much effort has been made to develop analysis and visualization tools, solutions for distributing image data and analysis results to the scientific community in collaborative environments are lagging behind (Walter et al., 2010). Bioimaging and analysis incorporate different specialized disciplines, e.g., from the fields of biology, medicine, pharmaceutics, statistics, physics, or computer science, each contributing to the analysis process regarding its own domain of expertise. Thus, scientists regularly needs to have access to data and results of collaborating

groups of experts. For example, groups specialized in data mining and modeling need access to data produced by image processing groups. In addition to exchanging image and analysis data, collaborating experts frequently need to communicate about specific aspects of the data at hand, e.g., discussing conspicuous image regions or intermediate results concerning further analysis tasks. This kind of "abstract" information also needs to be transferred from one expert to another. Up to now, one group of experts has to prepare the data regarding the problem that needs to be discussed, in order to send it per email or per CD/DVD to other experts, who in turn have to repeat this task. The exchange of data with linked information about specific aspects in this way is a tedious procedure and is actually no longer contemporary, however, it is common practice in scientific projects.

A successful cross-domain collaboration is often impeded, since the involved researchers are usually spread across several research institutes. Thus, modern bioimage informatics tools should provide frameworks that allow the scientific community to easily share bioimage data and results and functionalities that support interdisciplinary communicative tasks. For the task of data sharing, modern bioimage informatics tools require an advanced data management in combination with a centralized data repository ideally accessible for collaborating scientists from external institutes or labs. The realization of communication functionalities increases the requirements concerning the development of an appropriate data model as well as improved interactive visualization interfaces.

2.3 Summary

In the future, imaging will come to play an increasing role in modern biology, in particular in systems biology. Bioimages are of a high qualitative and quantitative information content regarding the biological system in many dimensions and across many scales. The aim of bioimage informatics is to provide the scientific community with tools to extract knowledge from the deluge of complex bioimages, in order to gain a systematic and unprecedented insight into biological processes. Bioimage informatics is a relatively young, but highly active research field and has already had a major impact to the solution of many biological questions (Peng, 2008). In the next chapter, an overview about existing bioimage informatics tools and systems will be given and discussed, followed by the motivation and a description of the goals of this thesis.

CHAPTER 3

State-of-the-art in bioimage informatics

In the previous chapter the computational cornerstones regarding the emerging bioimage informatics research field have been pointed out. In recent years, many approaches have been proposed and developed tackling various aspects of these requirements for bioimage analysis. The proposed solutions differ greatly regarding their biological question, the image data under investigation, and thus, the degree of flexibility or specialization. Therefore, recent bioimage informatics approaches can be divided basically into three different categories: *general purpose analysis*, *single purpose analysis*, and *analysis platforms*. In the following, some popular open source approaches and tools grouped according to these categories will be summarized and discussed.

3.1 General purpose analysis

This category subsumes basically no ready-made bioimage analysis applications or systems but rather general purpose toolboxes providing comprehensive collections of methods and algorithms for several image analysis aspects. Such collections can be considered as programming libraries containing classes and functions that can be applied for the efficient implementation and development of customized analysis software designed to solve specific bioimage analysis problems.

The first prominent example in this context is ImageJ[1] (Abramoff et al., 2004). ImageJ is a Java-based platform-independent program for biomedical image processing and analysis.

[1] http://rsbweb.nih.gov/ij/

It includes many imaging core capabilities to display, edit, analyze, process, save and print images with varying formats. The ImageJ program was designed with an open architecture that provides extensibility via user-written macros and Java plugins. This allows the users to develop custom acquisition, analysis and processing plugins using ImageJ's built in editor and Java compiler. Another way of extending ImageJ is to use its core functionalities and custom plugins as application programming interface (API) in order to generate external standalone software applications. Such a flexible application framework enables the scientific community to solve almost any image processing and analysis problem. A number of groups have developed several bioimage analysis tools using ImageJ that focus on different biological questions (Unser, 2008).

The second powerful general purpose toolbox is the Insight Toolkit (ITK)[2]. It is an open source and cross-platform software toolkit for performing a large variety of image processing tasks such as registration, filtering, and segmentation. ITK provides a C++-API implemented using generic programming principles based on templated code, which permit efficient algorithm development. ITK supports native and generic data types and a large variety of file formats. Additionally, this toolkit allows multiple language bindings, including programming languages such as Tcl, Python, and Java. ITK has been developed with the aim to provide a software foundation for future research, an archival repository of image processing algorithms and a collection of validation techniques, serving as platform for advanced software development (Yoo et al., 2002). Although ITK was originally developed for 3D segmentation and registration purposes of macroscopic medical data types, it can greatly be applied for bioimage data originated from microscopic imaging techniques.

The tools mentioned so far only address image processing issues. However, bioimage analysis requires a substantial amount of data analysis methods from the fields of datamining or machine learning to capture and reveal the biological information hidden in the image data. There exist several general purpose toolkits and environments providing a wide range of methods that can be applied in the same fashion as ImageJ or ITK. Some of the popular examples are JavaML[3] and the WEKA software[4] implemented in Java or the MLC++ library implemented in C++. These toolkits provide collections of machine learning, datamining or statistical analysis algorithms that can be applied to arbitrary types of datasets. Another well known and frequently used environment for data analysis is the R project[5]. The R language represents an integrative and highly extensible suite of software facilities for statistical computing and graphics.

Finally, since such analysis toolkits and libraries generally provide no appropriate visualization or user interface capabilities, except for ImageJ, which includes a basic graphical user interface (GUI) and plugins for image display and manipulation. However, as mentioned in Chapter 2.2.3, interactive visualization of image data and results is an important aspect in the bioimage analysis process. Therefore, the integration of external graphics libraries,

[2] http://www.itk.de
[3] http://java-ml.sourceforge.net/
[4] http://www.cs.waikato.ac.nz/ml/weka/
[5] http://www.r-project.org/

such as QT[6], VTK[7], or several Java libraries, and the design and implementation of suitable graphical components is a necessary prerequisite for meaningful bioimage analysis and exploration and calls for additional and considerable programming expertise.

3.2 Single purpose analysis

Single purpose analysis tools subsumes a category of ready-made standalone software solutions, developed using general purpose toolboxes and APIs, that can be installed and subsequently used by scientists not specialized in software development or programming. In the context of bioimage analysis single purpose tools are often focussed on well-defined biological problems or designed for a specific type of image data, i.e., either data acquired by a specific imaging modality or even data from a specific microscope. Single purpose tools provide dedicated pools of specially adapted analysis methods and interfaces to solve these problems. In the following, three different single purpose tools are outlined.

3.2.1 CellProfiler and CellProfiler Analyst

The CellProfiler software has been developed to provide biologists with an easy-to-use open source and platform independent software for automated analysis of cell images (Carpenter et al., 2006; Lamprecht et al., 2007). Due to its modular design and its graphical user interface, the software allows scientists without knowledge and training in computer vision or programming to quantitatively measure a large number of cell characteristics from thousands of images automatically. The software bridges the gap between the powerful general purpose analysis methods and their practical application in biological laboratories or projects. CellProfiler contains advanced built-in analysis algorithms and methods for many cell types and assays and offers flexibility for image analysis experts to extend the software by developing novel routines. CellProfiler has been designed and optimized for high-content screenings of two-dimensional images. It only provides limited support for time-series or three-dimensional image data analysis, however, scientists focussed on these type of data could develop and integrate compatible modules.

In addition to the CellProfiler, a second software tool has been developed as part of the CellProfiler project: the CellProfiler Analyst (Jones et al., 2008). While the CellProfiler provides methods to calculate and extract cell-based features, the CellProfiler Analyst has been designed for the interactive exploration and analysis of these measured cell features from high-throughput image-based experiments. Therefore, the system provides various visualization capabilities, in particular several types of plots such as histogram, scatter plot or parallel coordinates plot in order to be able to explore and compare cells or cell populations in one or more images based on their descriptive features. Different data plots are interlinked and interactive, which allows data points selected in one plot to be highlighted in all other plots. This filtering technique is often referred to as "Brushing" (Becker and Cleveland, 1987) and

[6] http://trolltech.com/products/qt
[7] http://www.vtk.org

helps the scientists to examine relationships in the high-dimensional data space. It also includes machine learning methods allowing for automated scoring of complex and subtle cell phenotypes usually represented by high-dimensional feature combinations. The CellProfiler Analyst, just like the CellProfiler, is a free and open source software and provides extensibility for experienced scientists to add new plots or analysis tools.

3.2.2 VANO

The volume-object annotation system (VANO) is a QT-based cross-platform annotation system for three-dimensional bioimages (Peng et al., 2009). It has been developed to create image annotations manually and to correct or refine the output of automated image annotation methods, e.g., image segmentation methods. The users can conveniently add or edit a label to a given volume object with a textual annotation such as cell name or property with a simple and intuitive graphical user interface. Textual labels are stored and visualized in a spreadsheet connecting the raw image data and the segmentation result usually referred to as a segmentation "mask". VANO has been applied to build high-resolution digital atlases of the nuclei and cells of specific biological specimens, e.g., *C. elegans* and fruit fly (Long et al., 2009). The major goal is to provide a software tool enabling the scientific community to create segmentation results as training data more accurate than automated methods.

3.2.3 CATMAID

The collaborative annotation toolkit for massive amounts of image data (CATMAID) is a Web-based interface, implemented in Javascript, for annotation of high-resolution multi-dimensional image data from large biological specimens (Saalfeld et al., 2009). This system allows the user to navigate arbitrarily large image stacks and to collaboratively annotate and share images and regions of interest (ROI). CATMAID provides an interface allowing for rapid browsing images at multiple scales via a tiled scale pyramid inspired by GoogleMaps^{TM} enhanced for 3D image navigation. Annotations are placed as point locations at the top of the images associated with textual labels or other semantic references such as ontology terms. CATMAID provides a partially decentralized architecture, i.e., image data resides on a user controlled server, whereas the references to these images and meta data about images, user information and annotations are stored in a central database. In this way, image data and annotations can be shared easily, while the owners of images retain full control over their data.

3.3 Analysis platforms

In recent years, a trend towards the third category of bioimage informatics approaches can be observed, which can be referred to as *analysis platforms*. In contrast to single purpose tools, the development of analysis platforms spends substantial effort in data management and organization. The basic idea of analysis platforms is to provide a framework that allows scientists to efficiently store, organize, search, and retrieve the large amounts of image and

image-related datasets. Apart from data management, such frameworks also include a user management that controls access to data usually stored in a central repository on a local or remote server and thus, allows data to be shared by collaborating scientists. Based on an intelligent data and user management, an analysis platform serves as a backbone for large-scale image analysis tasks. Analysis platforms can include selected methods or tools for data visualization, annotation and analysis or provide interfaces for custom external analysis in a client-server fashion. As an illustration, the two most recent best-known analysis platforms in the context of bioimage informatics will be introduced and described in the following.

3.3.1 Open Microscopy Environment

The Open Microscopy Environment (OME) is a comprehensive informatics solution for storage, management and analysis of optical microscopic image and meta data (Swedlow et al., 2003). OME is developed within a consortium of different collaborating research groups and laboratories, aiming at producing various open software tools focusing on (multi-dimensional) biological and biomedical imaging. One major focus of the OME lies on establishing standards in software and protocols that allow image data from different microscopes with different microscopy file formats to be stored, managed and shared. In (Goldberg et al., 2005) the OME data model has been described, which represents image data and all information regarding an imaging experiment, i.e., image acquisition and processing parameters and results created during data analysis. Based on this data model, OME also provides standard file formats (OME-XML and OME-TIFF) that reflect the OME data model in order to exchange OME files between different OME databases and software tools. Due to the fact, that data produced by commercial and academic imaging acquisition procedures is stored in a proprietary file format, new open imaging software has to be able to deal with these specific file formats for processing or visualization tasks. For this reason, the OME consortium has developed and released a Java library (Bio-Formats) designed to support the conversion of proprietary file formats to the OME-XML data structure (Swedlow et al., 2009).

Based on the OME specifications, the OME group has developed two data management tools, the OME Server (Johnston et al., 2006) and the recently released OMERO (OME Remote Object) software project. Although both applications are frequently used worldwide, future OME development is almost exclusively focused on OMERO because of some essential drawbacks of the OME Server architecture (Swedlow et al., 2009). OMERO is an open-source and cross-platform Java-based client-server software platform for visualization, management, annotation, and analysis of microscopic images[8] (Swedlow et al., 2009). The platform basically consists of the OMERO.server and several OMERO.clients. The OMERO.server is responsible for storage of image and metadata and provides processing facilities such as image rendering, analysis, and further programming interfaces. It can be run either on a local machine for personal data storage and management needs or site-wide allowing for a large-scale access for entire research departments or laboratories. The OMERO.server uses a relational database management system (PostgreSQL) that implements the OME data model

[8] http://www.openmicroscopy.org/site/products/omero

mentioned above. The OMERO platform provides access to the OMERO.server via remote Java client applications summarized as the OMERO.clients, which run on the user's desktop. These clients are developed for various basic data handling tasks. The OMERO.importer application allows users to upload proprietary image data files from a local file system to a running OMERO.server. This is achieved by using the Bio-Format library to prepare proprietary file formats for the import to the OMERO.server. With the OMERO.editor biologists are able to define and view experimental meta data and workflow protocols that is associated to image files within OMERO. This tool can be run as standalone application and is also part of the OMERO.insight. The OMERO.insight application includes tools for managing, searching, browsing, and visualizing data stored in an OMERO.server. It also provides facilities to administrate users and their access to the OMERO data. In addition to the standalone client applications mentioned so far, OMERO also provides a Web browser based client (OMERO.web) that allows the user to have remote access to the OMERO.server from any location without using a previously installed client on a specific local machine. OMERO.web includes all basic functionalities such as management of users, groups and server options and importing, browsing and viewing image data.

The OMERO platform is basically not designed to build novel image analysis algorithms, but instead to provide a structural framework that allows almost any application to read, use, and store images and associated data from microscopic imaging (Swedlow et al., 2009). However, the OMERO system supports analysis tasks via a rich API available in several programming languages, i.e., Java, C++, Matlab, and Python. In this way, users are able to implement their own analysis applications or clients, while using the same functionalities of the OMERO.clients to access data stored inside OMERO.

3.3.2 Bisque

The Bisque (Bioimage semantic query user environment) system is a recently introduced Web-based environment for management, sharing, annotation and custom analysis of biological images and its meta data (Kvilekval et al., 2010). Bisque allows users to upload and to securely store image data and associated meta data to its centralized database by using either the Web-based interface or custom external tools or scripts. Bisque also provides a Web image browser including functionalities for organizing, searching, browsing and filtering of images. The Web interface includes an image viewer designed to view single images or series of images and allowing for various visualization options like channel mapping, image enhancement, and projections or rotations. Furthermore, the viewer enables the users to set, edit and delete graphical image annotations. In Bisque, image annotations are either graphical labels such as object outlines or textual annotations, e.g., for the description of diverse experimental information. Textual information are stored in the Bisque database as meta information following a flexible and hierarchical tagging principle. The Bisque platform supports several types of integrated image analysis facilities developed for well-defined biological problems. Scientists have the option to apply internal Web-based analysis and visualization tools represented by an HTML interface enriched by Javascript widgets or alternatively they can use external user interfaces that are connected to the Bisque database

via HTTP requests. For analysis problems that could not be solved by existing Bisque tools, scientists can build and integrate custom internal or external analysis modules that reflect new analysis workflows. Due to the Web-based approach of Bisque, sharing of data is possible supported by export functionalities for images, analysis results and meta data. Bisque can manage a multitude of different types of biological image data, ranging from single two- or three-dimensional images to time series and multi-channel image sets.

3.4 Discussion

The approaches presented in this chapter provide an overview of the current state of development of applications in the field of bioimage informatics. The selected list of examples is not intended to provide a full overview of all existing bioimage informatics approaches. It should mainly illustrate the differentiation between the three categories, *general purpose analysis*, *single purpose analysis*, and *analysis platforms*. Therefore, some of the latest approaches or approaches that are often mentioned in recent literature that include relevant aspects regarding this thesis are introduced. Furthermore, since the focus of this thesis is on a free and open source software solution available for public academical use, commercial and vendor-specific software systems have been explicitly excluded in the overview of current bioimage informatics tools.

The existing tools introduced in this chapter represent great steps towards crucial improvements regarding specific aspects and analysis problems in bioimage data analysis. The different categories basically emphasize that bioimage informatics approaches are focussed on different concepts in data analysis, which has direct impacts on the degree of usability for the user community, e.g., biologists, clinicians or computer scientists. This results in category specific strengths and drawbacks, which will be discussed in the following. Based on this discussion, the chapter concludes with the motivation and the goal of this thesis.

General purpose analysis tools are intended to provide flexible and powerful frameworks, which can be considered as the necessary basis for developing and evaluating novel image analysis strategies and workflows. Such tools allow users to generate software prototypes rapidly that meet precisely the requirements regarding a specific biological or data analysis problem. However, the application of such toolkits is not a trivial task for scientists without training in computer vision or programming, since the adaption of routines often requires substantial knowledge in computer science and programming skills. The algorithms and functions available in these toolkits usually have to be implemented and combined in own software solutions, especially in the case of complex or specialized tasks. Furthermore, these toolkits generally provide no appropriate visualization capabilities or user interfaces except for ImageJ, which includes a basic graphical user interface (GUI) and plugins for image display and manipulation. As mentioned in Chapter 2.2.3, interactive visualization of image data and results is an important aspect in the bioimage analysis process. Therefore, the integration of external graphics libraries and the implementation of suitable graphical components for the visualization of image data and analysis results is a necessary issue and again calls for additional and considerable programming expertise.

In contrast, the single purpose analysis category and the category of analysis platforms both describe a group of software tools, whose aim is to provide ready-made software solutions, either as standalone bioimage informatics applications that have to be installed on a local machine such as the CellProfiler or the OME system or even Web-based solutions such as CATMAID or Bisque. These tools include selected methods or algorithms, which are focussed either on predefined bioimage informatics aspects such as VANO or CATMAID do or on providing a set of methods regarding a specific biological problem, e.g., methods for analyzing cell images offered by the CellProfiler or CellProfiler Analyst. Single purpose tools and analysis platforms usually integrate methods in a user-friendly graphical interface that allows users without programming skills and knowledge in software development to apply analysis methods to their data and navigate their results. Since tools often cannot cover all analysis aspects or in some cases include virtually no analysis methods, e.g., the OMERO platform that rather focusses on other bioimage informatics aspects like data management and visualization tasks, many tools offer interface capabilities to extend the tools with custom analysis applications and methods. Although such an extensibility property represents a powerful feature it implies the same drawbacks as with the general purpose tools: developing new applications as an extension requires substantial programming knowledge. In general, analysis strategies in current bioimage informatics tools are designed to solve particular well-defined biological problems or to manage data from specific imaging modalities. Thus, such tools are especially valuable and play an essential role when the analysis goal is known. As an example, the Bisque system provides external tools designed for specific biological problems such as microtubule or retinal studies (Kvilekval et al., 2010).

3.4.1 Motivation and goal of this thesis

However, in many cases the analysis goal is vague and little *a priori* knowledge is available for the underlying image data. In such cases, the application of predefined analysis methods or workflows is generally not suitable, since it is often not clear in advance, which aspects of the data analysis should be focussed on and which analysis strategy leads to meaningful results. This applies, for example, to data acquired with novel imaging modalities or to data where a biological sample was imaged for the first time with a given imaging technique or to image data acquired under special conditions, e.g., investigating the effects of drug treatment. However, the by far most challenging problem in this context relate in particular to those types of data, where the valuable information is not directly accessible. This is especially the case regarding high-content images or multivariate images. In multivariate images, the multi-dimensional signal domain is highly linked to the spatial domain, which is the special gain in this imaging modality and is of particular biological value in systems biology, e.g., in the analysis of protein co-location and protein-protein interactions, in order to identify functional molecular networks and to understand complete biological systems. The spatial information in individual images or channels within an MVI can be inspected and determined visually, whereas a sole manual evaluation of the complex information hidden in the signal domain is unfeasible. Signals belonging to single channels have to be considered in correlation with other channels and in combination with the spatial information, in order

to extract and quantify meaningful biological knowledge (Herold et al., 2011). Due to this increased data complexity of MVIs, novel and appropriate analysis strategies still have to be developed and evaluated. Therefore, based on their specific expertise, scientists need an initial exploratory access to the image information to gain insights into the structural characteristics of the data in a fast and intuitive way that aids the process of early steps in analysis and knowledge discovery, i.e., forming a mental model for the underlying data and developing hypotheses. In this context, methods from the fields of exploratory data analysis (EDA), visual datamining (VDM) or information visualization are ideally suited to cope with such image analysis problems. Here, the basic idea is to present the data in some visual form, allowing the human to directly interact with the data by adjusting and manipulating visual data displays, so that visualization is rather becoming an analysis and exploration tool than an end product of automatic analysis (Fox and Hendler, 2011).

Furthermore, the process of developing analysis strategies or searching for decision making criteria involves substantial communication and collaboration aspects, i.e., scientists usually have to share and discuss their data, analysis results and possible findings with collaborating scientists from other disciplines to develop concrete analysis strategies or workflows. Due to the increased complexity of high-content and multivariate bioimage data, it is virtually impossible to access, quantify and extract all relevant image information in one session by one researcher. In fact, image data needs to be evaluated by researchers from different fields (biophysics, cell biology, chemistry, computer science, statistics, etc.) regarding different aspects (image quality/noise, semantics, cell classification, staining specificity, statistical significance, etc.) and the results of their studies need to be integrated much earlier in the research process as it is done nowadays in many projects, where researchers from different institutes in different countries meet maybe once a year. Since collaborating scientists are usually spread across several research institutes, often worldwide, a successful joint development and evaluation of data and analysis strategies is a time-consuming and tedious procedure that unnecessarily prolongs the analysis process. Thus, in addition to the initial exploration of MVI data, scientists need new and efficient collaboration facilities to exchange information with other scientists, i.e., sharing scientific data and image related information, e.g., by free graphical and textual annotations, which might be linked directly to image regions and coordinates as it is done in the VANO or CATMAID tools, in order to simplify and speed up important communication tasks regarding MVI data analysis.

Although desktop solutions such as CellProfiler or OMERO provide sophisticated interactive data displays, they lack substantial collaboration abilities for geographically distributed scientists, e.g., sharing of data and results. In contrast, Web-based bioimage analysis solutions like Bisque or CATMAID offer far better collaboration and data sharing capabilities, since recently the Web is getting more collaborative and user-shaped (effects that are referred to as Web2.0), but they only include rudimentary Web-based data visualization and interactivity facilities.

In view of above observations and problems the question of how to efficiently foster these aspects in the analysis of complex multivariate image data is raised. In this thesis a novel bioimage informatics software approach *BioIMAX* is presented, which embraces this question. *BioIMAX* is a fully Web-based platform designed to augment both an easy initial exploratory

access to a large variety of complex high-content and multivariate image data and convenient collaboration facilities allowing for long distance and cross-discipline collaboration and communication of scientists via the Web, which is not covered by existing bioimage informatics solutions.

Due to recent developments in modern Web technology, offering more and more powerful graphics applications, the Web is getting more collaborative and user-shaped, which are effects referred to as *Web2.0*. For this reason, *BioIMAX* has been developed as a Rich Internet Application (RIA), which is a Web application whose performance and look-and-feel is comparable to a standard desktop application, but will be usually executed in a Web browser allowing for platform independency and avoiding annoying installation and maintenance costs, which are important advantages in comparison to standalone desktop applications. It can be observed, that the application of RIAs as part of the change of the World Wide Web towards Web2.0, recently called *Social Media* is becoming more frequent and more important, especially for the collection of user-generated content. *BioIMAX* is an attempt to investigate the potential of social network technologies in the context of the bioimage analysis by combining the Webs lightweight collaboration and distribution architecture with the interface interactivity and computation power of desktop applications.

The main objective of *BioIMAX* is not to design a Web-based LIMS (Laboratory Information Management System), but to provide a user-friendly Web-based work bench for collaborating researchers, which enables scientists to easily explore, interprete, share, and discuss multivariate bioimage data and results, independent from their whereabouts (condition to an Internet connection), and without a complicated and time-consuming act, such as data modeling or annoying installation of software packages. Following the idea of Web2.0, the ability to create scientific content that is stored on a central server and can easily be accessed by other scientists via the Web, fosters the community-driven research significantly. Such an Internet-based research and scientific collaboration in the age of *Web2.0* is referred to as *Science2.0* (Shneiderman, 2008; Waldrop, 2008) and has already been an active research area in recent years, e.g., in the field of health care and medical or clinical research[9].

[9] Journal of Medical Internet Research (http://www.jmir.org)

CHAPTER 4

Requirements

In view of the motivation and goals for a free Web-based platform for collaborative exploration of MVI data, several information technology aspects regarding the design and realization of the *BioIMAX* architecture have to be considered. This leads in the first instance to a list of general requirements the development of the *BioIMAX* system has to take into account. This chapter points out details about the different requirements for the realization of the *BioIMAX* platform and highlights the challenges and problems concerning these requirements. As an overview, the following list summarizes the general requirements mentioned in this section:

- User management
- Project management
- Analysis data management
- Rights/privilege management
- Tools for exploration and analysis of multivariate image data
- Integration of advanced collaboration facilities
- Platform usability

4.1 User management

The *BioIMAX* platform should be freely available to all scientists, who are occupied with any research question regarding the analysis of multivariate bioimages. Since *BioIMAX* provides a centralized data repository, which manages data owned by different users, the *BioIMAX* system has to include a suitable user management. Regarding *BioIMAX*, a user management is essential for two reasons. First, with a user management the multitude of datasets stored in the data repository should be associated to particular users, who can be considered as the owner of the respective data. This is especially important for security and safety relevant reasons, since in many cases research data should not be available to the public. Another advantage is, that users can easily and quickly search, retrieve and manage their own datasets, like it is the case with common social media platforms. The second reason for integrating a user management refers to the collaboration aspects, which form one of the major parts of the *BioIMAX* system. Collaboration issues, e.g., sharing data or communicating through the internal messaging system, are hardly feasible without having an appropriate user management.

For a registered user, a user management has to cover several user account specific information such as name, login, password, e-mail address or user-defined avatars. These information should also be used to design a user-friendly and personalized environment, which resembles aspects of known social media platforms.

4.2 Project management

As mentioned before, collaboration is a key issue of the *BioIMAX* system. In first instance, a prerequisite for a successful collaboration is the ability to share a specific subset of data and other information. For this reason, *BioIMAX* should provide a facility that allows users to establish small communities with the objective to group collaborating *BioIMAX* members, e.g., experts from different disciplines. This should be realized by a *project* entity, which can be created by any user allowing for inviting other members to join their projects. A project should enable a group of users to collect and organize an arbitrary number of MVI data or analysis results within a defined context, e.g., data regarding a particular biological or analytical question. Using user-coordinated projects should allow a clear organization of project relevant data associated with rapid access to the data. Project members should be able to read and process project related data, whereas the data remains hidden for non-members. Technically, a project has to manage memberships of collaborating users and has to organize a collection of data from the *BioIMAX* data repository.

4.3 Analysis data management

The *BioIMAX* platform has to manage different types of analysis data. In general, multivariate images constitute the core analysis datasets within the *BioIMAX* system. As described in Section 2.1 a multivariate image describes a stack of an arbitrary number of mutually

aligned images. Thus, the first question is how to store and organize a set of linked images ordered in a virtual stack of images in a centralized data repository, so that searching and retrieval of both, single images and whole MVIs is possible. Image storage and retrieval is a non-trivial task and is still an active research question in many scientific and non-scientific applications.

Image search and retrieval

For searching and retrieving specific data from large image repositories there exist mainly two strategies: *content-based image retrieval* (CBIR) (Lew et al., 2006) and *text-based image retrieval* (TBIR). In the case of CBIR, searching for images in a data repository is based directly on the image content. To that effect, the image content is represented by specific numerical image features. Given an example image, its image features are compared to those stored in the data repository and the search result lists all images, whose image features are most similar to the query example. In contrast, according to the TBIR strategy, images are indexed with descriptive textual tags or keywords often given by their respective filenames. The *BioIMAX* system should follow the TBIR strategy, since a CBIR is difficult to realize due to the complexity of the MVI data, which makes it virtually impossible to define appropriate and meaningful image features. As filenames are often inaccurate descriptions of the image content, users should be able to define own descriptive tags for their images. Indexing images with user-defined custom tags during the upload process should facilitate a more specific textual description of image content. Thus, the *BioIMAX* system should integrate an appropriate storage of textual tags associated to single images of an MVI.

The study or analysis of biological content based on bioimages often comprises large collections of MVIs acquired with the same imaging procedure. For this reason, users should be able to store tags selected for one MVI and reuse it for the upload of further MVI of the same type. This should speed up the upload process of large collections of MVIs significantly. The freedom of assigning images with arbitrary descriptive tags involves the risk, that images including the same content, e.g., images showing signals of the same protein, are assigned with different tags due to different users. In order to avoid a unnecessarily redundant list of image tags, *BioIMAX* should provide predefined libraries of tags that are commonly used in the context of bioimaging, e.g., a list of common antibodies used for fluorescence imaging techniques. This offers the opportunity for users to select a unified tag for all images showing the same content. As a result, using predefined image tags should increase the accuracy of the search results and avoids ambiguities, since the search space is limited to a minimum set of tags.

Variants of original MVI data

In addition to the original MVI data, the *BioIMAX* system further requires two special copies or variants of the original MVIs, which should automatically be generated during the image import process. First, all original images of an MVI have to be converted to an appropriate file format, which is supported by standard Web browsers, since *BioIMAX* should realized

as a Web-based application and the visualization of images is one of the essential parts of the system. In the same step, the intensity range of all images of an MVI should be scaled to a maximum range of 256 (8 bit) grey values, otherwise it is not possible to display images in a Web browser. Bioimages are usually acquired with much larger intensity scales, in order to capture the complete range of signals. Second, small thumbnail images has to be generated from any original image for preview purposes. This is especially important for displaying search results, which is one of the core functionality of the *BioIMAX* system. Using thumbnail images for certain purposes should considerably reduce memory costs as well as computing costs at runtime. Thus, the system has to incorporate three versions of MVI data, which have to be linked to each other.

Result data

Potential *BioIMAX* analysis or exploration tools will produce several types of result data. This data is usually characterized by complex data structures, which depend on the individual application or method that has generated the data. Analysis result data often includes a combination of different data structures, e.g., images, parameter sets or annotations. This poses significant challenges to the data management, in order to link these diverse analysis results to the respective MVIs.

In sum, the *BioIMAX* system has to incorporate several types of data with different data structures showing varying degrees of complexity. Therefore, the task of a data management is to efficiently store and manage the multitude of highly interlinked data allowing for searching and retrieval of desired data. Furthermore, the data management is responsible for providing the right data at the right place according to the application the user has selected.

4.4 Rights/privilege management

BioIMAX is intended to be an open platform for scientific image data and its analysis results. In science, data, analysis results or findings are in most cases highly confidential. Thus, it should provide privilege facilities and mechanisms, which guarantee that scientific data being confidential remains confidential within *BioIMAX*. Therefore, users should be able to decide, which of their data should be made accessible for particular users and to what extent it is available. For this reason, the *BioIMAX* system should provide a rights or permission management for all kinds of data stored to the *BioIMAX* data repository. In general, data should be associated with *read* and *write* privileges adjustable by the owner, who controls the access to the data for all users of the system. If the read permission is given for a dataset, all users should be able to read this data, otherwise it should remain invisible to all other users. The write permission controls, whether datasets can be used for analysis or processing by other users. In order to authorize data access to a particular group of users, the owner of the data should create a project (see Chapter 4.2) and add other users to it. This grants access to data associated with this project. Finally, project owners still have the opportunity

to decide, whether other project members have only read permissions or both, read and write permissions on data within the project.

4.5 MVI data exploration and analysis

The aspects mentioned so far point out the necessary requirements for the basic framework of the *BioIMAX* platform. As the exploration and analysis is one of the major goals of the *BioIMAX* system, this section addresses general aspects concerning interfaces and functionalities that should be considered in the development of *BioIMAX*.

A prerequisite for and the first step in the exploration and analysis of MVI data is the development of an appropriate interface, i.e., a Web-based *image viewer* for displaying image data of an MVI and for visualizing analysis results regarding visual exploration and evaluation tasks. An image viewer in *BioIMAX* should enable users to visually browse through an MVI stack with the objective to gain an initial overview about the raw image signals. Therefore, a Web-based image viewer should include navigation functionalities such as zooming, panning and flipping through a series of images, which are well known from conventional desktop image viewing tools. However, an MVI describes a special type of image series, where each single image or channel depicts a distinct molecular meaning of the same visual section of a biological sample obtaining a stack of aligned images of the same size (see Section 2.1). Each image position or region has corresponding positions or regions in all other images and the comparative observation of such regions in different images is an essential aspect regarding first steps in MVI analysis. Thus, the demand for displaying a series of aligned images poses non-trivial challenges to the realization of an image viewer, which is usually not owned by conventional image viewers. All image navigation functions have to be performed on the whole image stack in parallel, i.e., all images of the stack have to be displayed at the same scale and at the same position in the image viewer. This should allow sequentially flipping through an MVI stack without losing the orientation regarding an observed region.

A pure visual inspection of an MVI by flipping through the image stack image by image is not sufficient to capture the combinatorial information inherent in the MVI data at once, in particular with increasing data dimensionality. Thus, scientists need further visualization tools, which allow a more efficient and detailed visual analysis or exploration of the multivariate signal domain. The analysis and exploration results should be interactively displayed in the image viewer, which therefore has to be extended with additional visualization functionalities. Thereby, methods from the fields of information visualization (Card et al., 1999; Chen, 2004; Spence, 2007; Ware, 2004), visual datamining (Keim, 2002), and exploratory data analysis (EDA) (Tukey, 1977) are ideally suited to cope with the analysis of multivariate datasets, especially if the exploration goal is vague. Thus, the *BioIMAX* system should include selected methods from these fields tailored to the needs of multivariate image exploration. This way, it is possible to capture and to quantify differences or similarities of MVI channels with different molecular meanings and to generate and verify hypotheses of the respective MVI data. Here, it should be possible to compare a selected set of whole channels as well as to focus the visual exploration on defined image regions following Ben Shneiderman's

mantra *Overview first, zoom in and filter, and details on demand* (Shneiderman, 1996). This is a valuable feature, since scientists often have their own expert knowledge, e.g., regarding special biological signals or spatial molecular organizations, so with *BioIMAX* they should be able to concentrate on particular image details. For this reason, image data should be presented in a visual form allowing for interacting with it, so that humans are directly involved in the data exploration process and combine their general or specific knowledge with computational visualization power (Keim, 2002). To this end, *BioIMAX* should provide several data displays reflecting different image data aspects with varying degrees of detail. This ranges from manual annotation based on initial visual inspection to fully automatic datamining or unsupervised learning techniques.

4.6 Collaboration

Basically, collaboration in *BioIMAX* can be divided into two categories: *data sharing* and *communication*. The option to share data can be considered as the necessary basis for an efficient communication about image data, hypotheses and analysis results and should be covered by the project concept mentioned in Chapter 4.2. *BioIMAX* projects enable a group of collaborating users to collect and share several types of data with respect to a specific analysis question.

Communication can take place at a general level, i.e., exchanging general information about image data and results or arranging organizational issues, which is usually done by email correspondence. Therefore, *BioIMAX* should provide an internal messaging system, so that general information related to *BioIMAX* remains within this platform, thereby reducing the need for additional external tools. In the context of multivariate bioimage analysis, communication is directly related to image content. This refers to discussions about the application of specific analysis methods and their intermediate results, e.g., concerning the accuracy of analysis methods, the comparison to gold standards, or to discussions and verifications of findings and hypotheses based on data exploration. Thus, analysis related decisions are associated to particular image regions of interest, i.e., discussions need to be linked to image coordinates. This leads to less trivial design issues in database and graphical user interface development. *BioIMAX* users need an innovative user interface, which should enable them to easily label image regions with graphical annotations and to link textual semantic information to these graphical annotations. This includes comments, questions or even comprehensive conversations, e.g., in a chat-like manner, facilitating collaborative work on image content with several other users.

4.7 Platform usability

Nowadays, the usability of software solutions plays an important role in software development projects and has led to separate research branches focussed only on usability aspects in software development. Particularly in science, the usability of analysis software is a crucial point, since the amount and complexity of data, that has to be processed is steadily increasing

and the management of data constitutes a considerable part in the analysis process. The actual analysis of data and scientific reasoning should not unnecessarily be impeded by the installation and maintenance of complicated and overloaded software solutions. This also applies to the *BioIMAX* system and the analysis of multivariate image data. In this context, a distinction can be made between the usability for users of *BioIMAX* and for developers of future *BioIMAX* tools and analysis functionalities.

From the user perspective, the *BioIMAX* platform should hold the following properties regarding software usability:

- *Platform independence.*
- *No local installation and maintenance costs. BioIMAX* users should not be bothered with annoying installation or compilation routines of software packages and external libraries on their local machine. This would have the additional advantage that they are not responsible for upgrades to the latest releases of the software.
- *Platform availability.* The platform should be available and accessible at any time independently of the user's whereabouts. This applies to the availability of the platform itself and also to all data stored within the system.
- *No local data storage management.* All *BioIMAX* relevant data should be stored and managed on a centralized data server, in order to avoid unnecessary storage costs and management of data in complex folder structures on the local machine. Data should be integrated into the running system as simple as possible, which requires an easy-to-use interface for importing data. In addition, using a centralized data storage on remote server facilitates effortless data sharing with collaborating users.
- *Data search and retrieval.* The demand for a centralized storage of data on a remote server implies that the *BioIMAX* platform has to provide an appropriate data browser to query the database and display desired subsets of retrieved data. With this data browser interface, users should be enabled to search, browse, filter, extend, modify and manage their own datasets, data associated to specific projects or foreign datasets from other users (provided that they are at least assigned with read privileges).
- *Easy access to data processing tools.* On the basis of the data browser, users should be able to select specific datasets and to immediately invoke data processing tools for exploration or analysis available for these datasets.
- *Software as a service.* Since with *BioIMAX* the data should be hosted on a remote server, time-consuming and CPU-expensive analysis processes such as datamining or unsupervised learning tasks should also be performed on a remote server instead of running on local workstations. The concept of outsourcing software applications on a remote server architecture has been termed "Software as a Service" (SaaS) (Turner et al., 2003). In the context of *BioIMAX*, the SaaS approach provides the following major advantages: First, users could continue working with other aspects of *BioIMAX*

or they can close the application, while waiting for the completion of the extensive processes. Second, the remote processes do not consume any resources of the local workstation. Thus, *BioIMAX* users should be able to simply start new processes on the server and should be informed, e.g., by email, when the process on the server has been finished and the newly generated datasets and results are available within the *BioIMAX* platform.

- *Modular interface design.* In general, *BioIMAX* should provide an easy-to-use graphical user interface (GUI). The design of the interface should be well-structured and clearly organized, in particular regarding the application of different data processing and analysis tools. Therefore, tools (including functionalities with respect to specific aspects of data analysis) should be realized in a modular manner, e.g., by providing separate graphical windows for different tools. The objective is to split the GUI into several parts allowing for a clear arrangement on the user's desktop screen and therefore, avoiding an overloaded GUI.

From the software developers point of view, it should be possible to develop and implement new analysis functions and tools, which can be integrated into the existing and running system with a minimum of effort and system downtime. For this reason, appropriate programming interfaces should be provided, which guaranties that new tools have access to existing datasets and that new types of data can easily be integrated into the database. This implies that the *BioIMAX* system comprises a flexible and transparent data model in combination with an appropriate data management.

Figure 4.1 graphically illustrates the idea and the concept of the Web-based *BioIMAX* system. Based on the defined requirements in this section, the technical realization and implementation of the *BioIMAX* platform generally involves the following steps, which will be described and illustrated in detail in the next sections.

- Design and development of a database in connection with a database management (DBMS), in order to store and manage all the data and information collected while using *BioIMAX*.
- Development and implementation of a graphical user interface as application front-end.
- Realization of the communication between the application front-end and the database server.

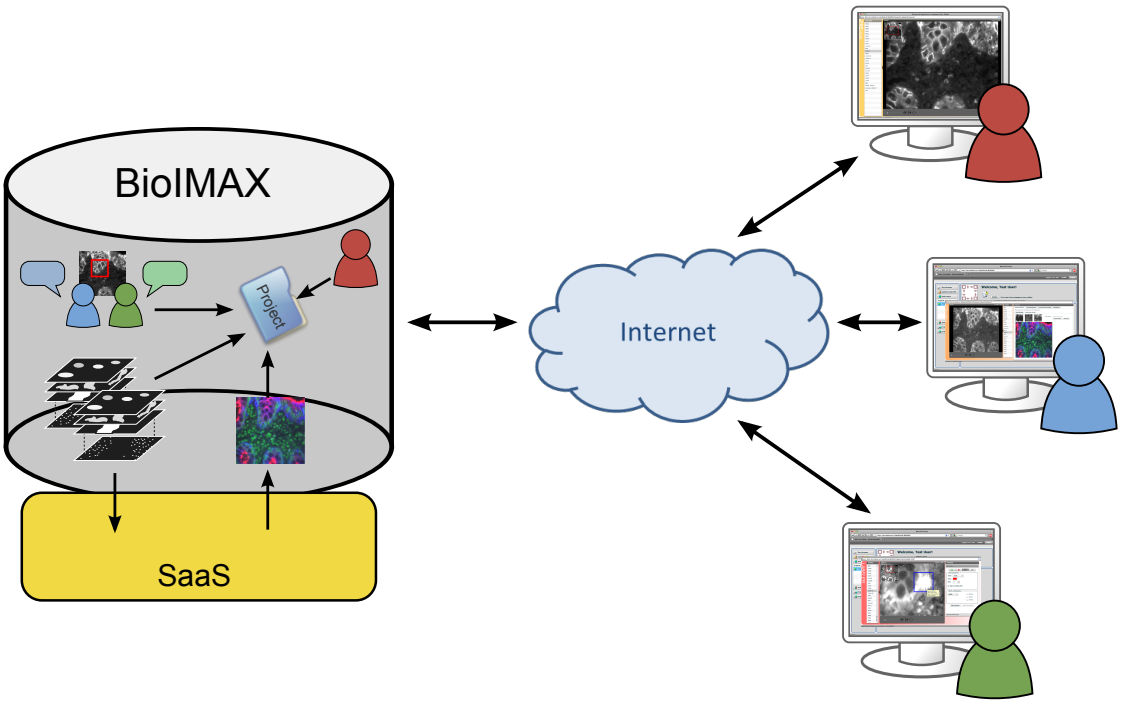


Figure 4.1: The *BioIMAX* idea. This figure represents a simplified schematic illustration of the idea of the *BioIMAX* system. *BioIMAX* should allow scientists of different disciplines to collaboratively work on research questions regarding the analysis of multivariate bioimage data by using a Web-based software that runs in a standard Web browser. The major goal is to cover all basic facets in the analysis of MVI data, i.e., data management, data analysis, and scientific collaboration, by one single platform accessible over the Internet. Via an individual user account *BioIMAX* users should be able to easily upload their image data into a central database, use analysis tools provided within the platform, and share their data, results, knowledge, and scientific findings with other users in a quick and uncomplicated way. The database centrally holds all data such as images, analysis results, and communication between different users and stores the relations and links between these instances. Time-consuming and computational expensive analysis routines should be outsourced to a powerful application server following the modern concept of Software as a Service (SaaS). Access rights should guarantee, that data remains confidential within the *BioIMAX* platform. A project concept should allow users to give access to their data for selected individuals. *BioIMAX* should foster the collaborative work on bioimage data using state-of-the-art Internet technologies without the hurdles of installing and maintaining additional software on the users' desktop machines.

CHAPTER 5

Architecture

In light of the requirements formulated in the last chapter, the architectural and technical issues of the *BioIMAX* system are addressed in this chapter. At first, Section 5.1 presents a comprehensive specification of the *BioIMAX* data model. The careful definition of an appropriate data model that reflects the variety of data and processes within *BioIMAX* is considered to be the necessary basis for the development and implementation of a Web-based platform for collaborative analysis and exploration of multivariate image data. A detailed explanation of the technical realization and implementation design issues follows in Section 5.2.

5.1 Database design

The design of the *BioIMAX* database is built on the *relational database model*. For this reason, all concepts, data and their relationships, which should be covered by the *BioIMAX* system have been mapped to a conceptual data model, reflecting the concepts and data of the *BioIMAX* world in an abstract way and have been finally implemented using the relational database management system (DBMS) MySQL[1]. In this section, the essential entities of the conceptual data model and their roles within *BioIMAX* are described and illustrated. Due to the multitude of demands on the entire *BioIMAX* platform, the *BioIMAX* data model has to reflect several types of data, information and concepts with varying degrees of complexity and

[1] http://www.mysql.com

interconnections. In general, the types of data can be divided into two categories: *analysis data* and *meta data*, which are described in detail in the following subsections.

5.1.1 Analysis data

The analysis data category subsumes those types of data and information, which are directly involved in the exploration and analysis procedure of MVI data. An analysis process in *BioIMAX* already starts with searching for specific original MVI datasets followed, for instance, by an initial visual inspection of single images of a selected MVI stack. Therefore, the analysis data category includes both, raw images from original MVI datasets, forming the central datasets in *BioIMAX*, and all kinds of data *derived* from the original MVIs. In *BioIMAX*, derived data is the result of a transformation or interpretation process, investigating different features of entire MVI stacks or particular sets of images from an MVI. This kind of projection from the original image domain to a specialized representation of MVI features is obtained through one or more data processing steps (see Figure 5.1(a)). Examples of data regarding this category can be characterized as follows.

- First of all, this category comprises the raw and unprocessed images associated to an MVI.
- In the simplest case, derived datasets are special copies of the original MVI, i.e., each channel of the original MVI is transformed in the same way, e.g., through an intensity normalization enabling the display of images in the Web browser or through scaling the image size, e.g., to create thumbnails of images of an MVI. Transforming MVIs in this way results in a new MVI stack with the same number and order of images as with the original MVI.
- Data from manual graphical or textual image annotations, represented by several properties such as position, size, color, shape or plain text.
- Results from semi-automatic analysis strategies such as image segmentation, classification or data mining routines.

The integration of the analysis data category into the conceptual data model is a challenging task, since particular types of MVI transformations produce different result data with varying data structures. Some of the data structures can completely be mapped to the data model described by specific entities, which can be managed by the relational DBMS, e.g., the data structure characterizing graphical or textual image annotations. Other data structures from an MVI transformation are more complex, in particular image files, descriptive models of the data and their results stored as special proprietary file formats or combinations of multiple data structures that can be considered as *distributed data structures*. It would be difficult to integrate these complex data structures into a relational database in an efficient way and therefore, they cannot be reflected by dedicated entities in the data model. To solve this problem technically, such data has to be stored on a separate dedicated file server. The relational database holds the information where the data is stored on the file server

and contains additional meta information and parameters about the external data. Due to this kind of data handling, using a relational database in combination with an external file server, it is possible to associate data with complex data structures to all other data and information managed by the relational DBMS. A detailed description of the results and their data structures obtained by different MVI transformations and on how the different results are stored and reflected by the data model, will be given in Chapter 6, where the different *BioIMAX* interfaces and their respective data handling are illustrated.

5.1.2 Meta data

In contrast to the analysis data category, the meta data category comprises all types of data and information, that are not directly involved in the transformation process for the exploration and analysis of MVI data. Basically, data regarding this category can be considered as administrative data, which is particularly responsible for the management and organization of all data concerning the analysis data category within the *BioIMAX* environment. With this category, analysis data can clearly be assigned to particular *BioIMAX* users or specific projects, and it holds additional meta information about original MVIs imported into the *BioIMAX* system. The meta data category contains three major entities, which can be regarded as essential cornerstones of the *BioIMAX* data model: *User*, *Project* and *ImageStack*. In the following, their roles in the data model and their properties will be specified in detail.

User entity

The *User* entity is the top level entity of the entire data model and the central element of the meta data category, which is involved in all *BioIMAX* operations. Basically, the *User* entity is required to identify *BioIMAX* users by a unique user ID. The major roles of the *User* entity with respect to the entire *BioIMAX* system can be described as follows:

- Representing relevant user account information to control the registered access to the *BioIMAX* platform.
- Associating all user generated content within *BioIMAX* to a particular user, who will be defined as the owner of the respective content.
- Being essential for the implementation of a rights management, in order to control access to user-specific content for other users.
- Constituting the key element for all collaboration and communication issues within *BioIMAX*.

Project entity

The *Project* entity is the basic component for the management of user-created projects. In combination with the related entities *Project_views*, *Project_members*, and *Project_invitations* the major objectives of the project management can be characterized as:

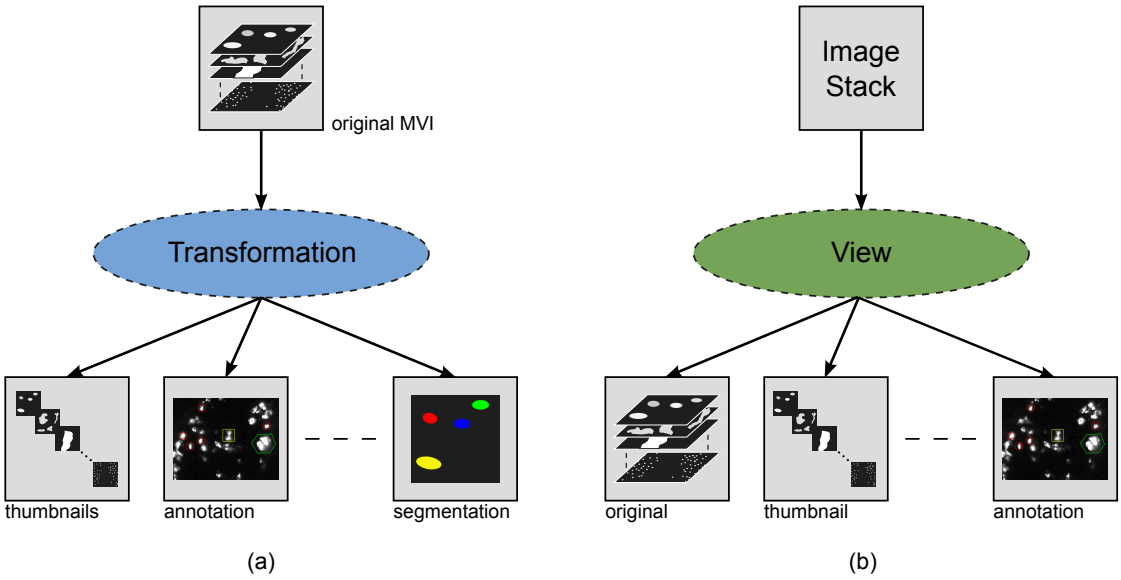


Figure 5.1: Transformation and View concept. The graphic on the left (a) schematically illustrates the concept of the transformation or projection from the original image domain to a specialized representation of MVI features, e.g., thumbnails, graphical annotations or segmentation results. Transformations are obtained through one or more data processing steps. The graphic on the right (b) depicts the idea of the View concept within the *BioIMAX* data model. Transformation results can be considered as a special views or perspectives on the original ImageStack and are subsumed under a generalizing *View* entity. Thus, each result datasets corresponding to one specific MVI transformation is identified with a unique View within the system. A special role in the View concept plays data from an original ImageStack. Images from an original ImageStack are also treated as analysis data, even though they are not a result of a transformation process. For this reason, an additional View type describing original image data has to be defined: the *original view*.

- Collection and organization of user generated content with respect to a defined biological or analysis relevant topic.
- Linking of an arbitrary number of *BioIMAX* users as members to the respective project.
- Control of the projects access rights and the procedure of inviting foreign users to a project.

ImageStack entity

For a structured integration and organization of MVI datasets including its derived data it is necessary to provide an entity, which allows a clear identification of new MVIs as source datasets within *BioIMAX*. For this reason, the data model includes the *ImageStack* entity, where each original MVI is represented as an instance of the *ImageStack* entity with a unique ID being the key to the respective instance. The term *ImageStack* is a generalizing

synonym for an MVI, since it is generally possible to import an arbitrary stack of images of the same size with a defined order into the *BioIMAX* system, even if the single images of the stack have no multivariate characteristics, e.g., images showing signals from different samples or from different visual fields. In the following, the term ImageStack is also used to describe a multivariate image. The *ImageStack* entity is assigned to the meta data category, since it can be considered as an abstract representation of an ImageStack, which holds only general information and properties *about* an ImageStack at hand and *does not* contain the raw image data. As mentioned before, the raw images of an ImageStack are associated to the analysis data category and are managed by the *View* entity, which will be described in detail in the next Section 5.1.3. Instances of the *ImageStack* entity are uniquely assigned to its owner, who has uploaded the ImageStack to the system via the user ID as a foreign key. In addition, the *ImageStack* entity enables the integration and association of additional meta information such as imaging parameters or biological and medical information about the imaged sample to a unique ImageStack ID.

The three basic meta data entities and their related entities can completely be mapped to the data model and will be managed by the relational DBMS. A complete overview of the *BioIMAX* data model and the relationships of the single entities is illustrated in Figure 5.3.

5.1.3 View concept

Finally, the *View* entity is the last, but the most essential cornerstone of the basic *BioIMAX* data model. The idea of the View concept is, that any analysis result can be considered as a special *view* or *perspective* on an original ImageStack imported into the system and is comparable to the ImageStack transformation principle mentioned above (see Figure 5.1(a,b) for a graphical illustration of both concepts). One of the major objectives of the *View* entity is to connect the analysis data category to the meta data category in a distinct way, thereby avoiding unnecessary redundancies and complex links within the database. A *View* instance collects and aggregates all analysis results and their data structures regarding a particular ImageStack transformation. ImageStack transformations or projections in *BioIMAX* are represented by single *View* instances, each of them characterized by a specific *View type*, which match the type of the respective transformation. Each *View* instance is clearly associated to an *ImageStack* instance by using the unique ID of the respective *ImageStack* instance as foreign key in the *View* instance. All results of one transformation regarding a particular ImageStack are associated to a respective *View* instance using a unique *View* ID, indendently of how the results are stored on the server or how they are distributed. Thus, all data of the analysis category is connected to the respective source *ImageStack* entities through the *View* entity.

The handling of raw images from an original ImageStack plays a decisive role in the View concept. As mentioned before, original images are also associated to the analysis data category. Since these images are not a result of a specific transformation step, there exists no direct View type equivalent to a transformation. Thus, an additional View type called *original view* has been defined, which allows the assignment of raw images of an original ImageStack to a particular *View* instance. For each newly imported ImageStack the raw

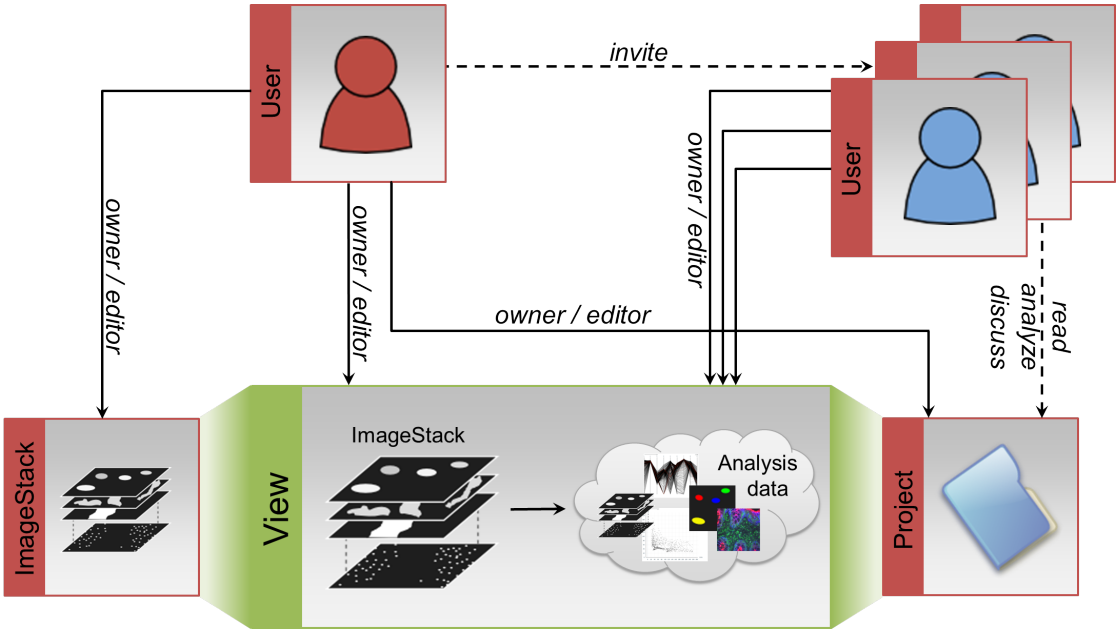


Figure 5.2: The basic *BioIMAX* data model. Illustration of the core entities forming the basic conceptual scheme for the relational database model and their functional roles within the *BioIMAX* system. The *BioIMAX* data model consists of four basic entities: `User`, `Project`, `ImageStack`, and `View`. A registered *BioIMAX* user uploads a multivariate ImageStack to the *BioIMAX* database. Afterwards he/she is uniquely associated to the newly integrated image data as its owner and has the exclusive rights to edit the ImageStack, to control its availability for other users, or to delete the ImageStack. All data generated based on an ImageStack, either automatically or via analysis tools within *BioIMAX*, is represented by respective Views, which describe a specific perspective on the original ImageStack. The major objective of the View concept is to connect the entities of the meta data category (`User`, `Project`, `ImageStack`) to the analysis data category. In order to share data with a selected group of individuals, a user can create and manage a project instance with which data in the form of Views as well as invited users can be brought together. Using projects as communication channel, collaborating users get access to data from other users, can generate new project related content, and discuss relevant datasets, e.g., regarding specific biological analysis questions.

images are stored on the server without any previous processing of the image data and are associated to the respective *View* instance via a unique *View* ID. Together with the type *original view* all data of the analysis data category can completely be connected to the meta data category of the *BioIMAX* data model.

In sum, the *BioIMAX* data model has a simple but powerful structure, since it is based on just a few essential entities, which allows a fast and transparent access to all data stored in the database. In particular, the View concept plays a key role in this data model. The essential benefits of the View concept with respect to the entire data model are pointed out in the following.

- First, the View concept allows an easy and flexible integration of novel algorithms, tools, or interfaces, which potentially generate new types of data. These new data types are usually stored using appropriate data structures, e.g., database tables or even more complex data structures, that are not yet included in the *BioIMAX* data model. However, the *BioIMAX* data model can be extended with these new data structures without changing the rest of the data model, which applies in particular to the meta data category. The developer only has to define a new View type, which should describe and identify the newly embedded datasets within the data model. In this way, each dataset is connected to the basic entities of the meta data category. The same procedure also refers to modifications and updates of existing tools or functionalities, which can be performed without adapting the entire data model.

- Second, the way the system deals with complex and even distributed analysis data structures is simplified considerably. This refers to data handling aspects such as searching for specific data, displaying the search results, and assigning analysis results to particular projects. In particular, the View concept reduces the amount of data transferred from the *BioIMAX* main application to dedicated analysis tools (see Chapter 5.2.4 for detailed explanation of the *BioIMAX* interface concept). Thus, working with abstract Views within *BioIMAX* rather results in a faster and a more performant system than by using and providing the total number of corresponding analysis datasets at any time the system requests the data.

- Finally, the association of additional meta information to specific analysis data is easily possible via *View* instances in a central manner independently of the type of meta information.

In Figure 5.2 the conceptional data model based on the essential entities *User*, *Project*, *ImageStack*, and *View* and their functional roles within the *BioIMAX* system are illustrated, whereas Figure 5.3 depicts the detailed overview of the entire data model showing all entities and their interconnectivity.

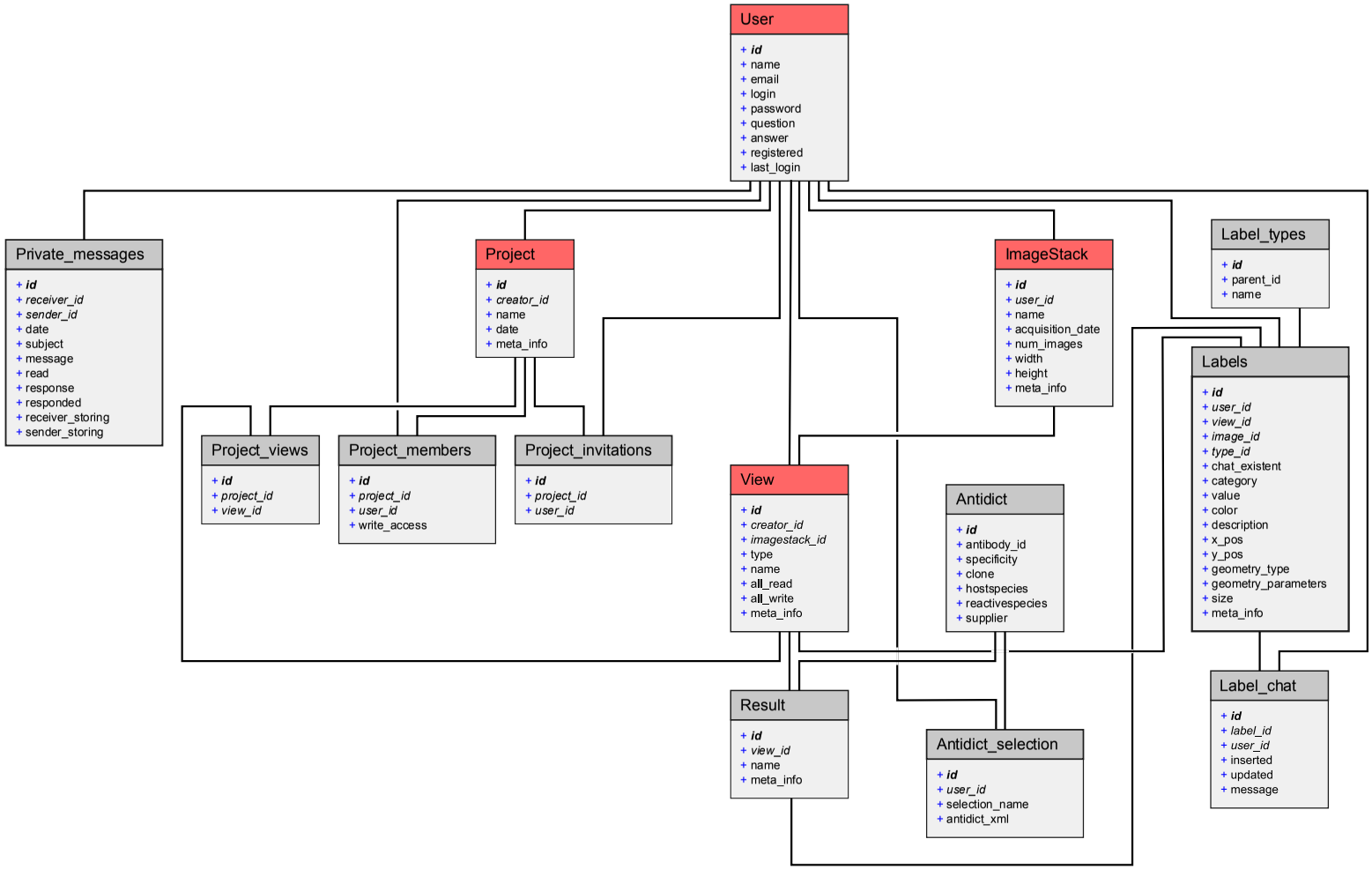


Figure 5.3: Complete overview of the *BioIMAX* data model. This schematic diagram depicts a detailed overview about of the complete *BioIMAX* data model including all entities and their interconnections. Each entity contains a number of attributes describing their properties. Key attributes are italicized, whereby the primary keys are additionally written in bold. The entities `User`, `Project`, `ImageStack`, and `View` are the essential cornerstones of the data model and are highlighted with a red colored header. Relationships between entities are represented by lines connecting the entities.

5.2 System design

In this section, the components of the *BioIMAX* system architecture and its technical realization are described. Since *BioIMAX* is developed as a Rich Internet Application, this section at first gives a brief overview about the current trend of the World Wide Web towards Web2.0 and the associated development of recent Web technologies and addresses its relevance regarding modern scientific work. Finally, the technical details of the *BioIMAX* architecture are illustrated and described.

5.2.1 A Short history of the Web

Since Tim Berners-Lee invented the World Wide Web (WWW) in 1989[2] at CERN (European Organization for Nuclear Research)[3], it has undergone a considerable development over the last two decades. While at the beginning, the WWW was a system for exchanging statically linked information available only for a limited number of users, it is nowadays an indispensable and extremely dynamic communication medium in our modern society. The WWW exploits the possibilities and advantages offered by the Internet, being at the same time a world-wide broadcasting capability, a medium for the dissemination of information and a system for collaboration and interaction between individuals and their computers independent of their geographical whereabouts. During the development of the WWW, Tim Berners-Lee has introduced three essential standards, on which the WWW is still based and which defines its core functionalities. WWW resources are formatted using a markup language called HTML (HyperText Markup Language) and are identified by unique addresses, the URLs (Uniform Resource Location). For requesting HTML documents and transferring them from the server to the user's computer, Berners-Lee developed the network protocol HTTP (HyperText Transfer Protocol). The release of the graphical Web browser named Mosaic (later Netscape Navigator) in 1993 revolutionized the way how the Web was used and led to its popularization. Through an easy-to-use Web browser, users were able to request and view extremely interlinked HTML documents. Based on the WWW concept, documents no longer are considered as coherent files, but rather as heterogeneous objects, which are connected to other resources such as image, audio or video files through a network structure.

While in the early days of the WWW, Web content has been centrally generated and published by a small group of Web designers, the usage behavior of the Web has changed dramatically in the last few years, not least due to the increased availability and bandwidth of Internet connections world-wide. Web users are no longer passive consumers of Web content, they themselves more and more become an active part in the generation and organization process of their information and relationship management in the WWW. This change in using the Web constitutes the core characteristic of a modern Web type, referred to as *Web2.0*. The term Web2.0 has been coined by Tim O'Reilly (OReilly, 2007) and is discussed controversially, since it mistakenly suggests, that it describes an update of the WWW to a new technical version. However, it rather characterizes new interactive ways of using the Web and the

[2] http://www.w3.org/History/1989/proposal.html
[3] http://public.web.cern.ch/public/

changed public awareness regarding the Web. Thus, recently Web2.0 is increasingly replaced by the term *Social Media*. Due to advanced browser interface capabilities such as those provided by content management systems, data- and media-sharing platforms, Wikipedia[4], or social networking platforms, Web end-users without experience in software development and Web design are fostered to qualitatively and quantitatively generate, manipulate and disseminate own Web content, referred to as *user generated content*, so that the WWW no longer represents a network for only requesting and retrieving interlinked information.

As a consequence, the term Web2.0 is closely related to technical innovations regarding the design of user-interfaces and improved software and storage facilities, which have already been developed in the mid-ninetees, but they generally became available with the advent of broadband Internet connections world-wide, opening up new vistas for the public Internet community. In the beginning of the WWW, information were purely encapsulated into HTML pages statically interlinked with other resources and provided only a limited degree of rudimentary interaction possibilities such as buttons, forms or hyperlinks. Each user interaction required a full reload of the Web page in the Web browser by generating new HTML files on the remote Web server. This process is rather inefficient, since for each reload of the Web page the creation and transfer of the complete HTML page have to be performed, which leads to longer waiting times and cumbersome information presentation for the reader of the Web page.

With the introduction of *JavaScript* in 1995, for the first time it was possible for designers of HTML Web pages to use a programming tool that enables them to design and generate advanced interactive and more dynamic HTML pages. JavaScript is a scripting language implemented as part of the Web browser. Since JavaScript code will not be executed on the remote Web server, but runs locally in the Web browser, user actions are performed more quickly making the Web page more responsive. Using JavaScript it is possible to access and change existing HTML elements or objects directly, without reloading the complete HTML page from the Web server. Here, JavaScript interacts with the *document object model* (DOM)[5] of the current HTML page. The DOM has been defined by the World Wide Web Consortium (W3C) as a standard convention that describes how HTML documents has to be formally represented and how to interact with their elements and objects using a specific programming language such as JavaScript.

However, while using HTML pages with embedded or included JavaScript code, the problem of refreshing the HTML page still persists, when new information is requested from the Web server triggered by user action, e.g., loading data from a database. The Web browsers, and therefore the users, still have to wait until the Web server has generated a new HTML document, which entails a complete reload of the current HTML page. Modern Web technologies solve this problem, by applying special client-side engines, which allows them to asynchronously exchange data between the Web browser and the server, avoiding limitations in usability and behavior of the HTML page during the data loading process. One of the most well-known and frequently used Web technologies in this context is the *AJAX* (Asynchronous

[4] http://www.wikipedia.de
[5] http://www.w3.org/DOM/

JavaScript And XML) framework (Garrett, 2007). AJAX is not a newly developed technology, but it rather characterizes a combination of traditional techniques such as HTML, CSS, DOM, XML, XMLHTTPRequest, and JavaScript that are combined properly, in order to facilitate the dynamical exchange of data from the client to the server and vice versa. Here, the XMLHTTPRequest object (based on a W3C specification) plays an important role in this Web concept, since it is responsible for wrapping a client request structured in XML, which is sent via the AJAX engine using HTTP to the Web server. The Web server processes the request and sends back the response to the client-side AJAX engine, which in turn triggers JavaScript to update only relevant parts of the HTML document instead of the entire page. This process is performed in the background, letting the user continue to interact with the Web page, which speeds up Web pages' performance, responsiveness, and interactivity.

Many of modern websites have been realized with frameworks like AJAX including content management systems, weblogs, social network platforms, various wiki Websites, media sharing platforms, or Web applications. Using AJAX in combination with CSS, graphically appealing Web applications can be created, whose functionalities and behavior resemble rather desktop applications than conventional Web pages. The concept of Web2.0 is built around such user-centric Web applications, providing a high degree of usability and powerful interactions, which has great effects on the user Web experience. The development and delivering of such modern and rich Web applications, usually called *Rich Internet Applications* (RIA), has increased dramatically in recent years and they are more and more shaping the Web. RIAs can be considered as full software running in a Web browser and its architecture corresponds to the so-called *fat client*, in contrast to traditional Web applications referred to as *thin clients*. The major goal of RIAs is, to extend the conventional hypertext-based Web with new capabilities by combining its lightweight distribution architecture with the interface interactivity and computation power of desktop applications (Fraternali et al., 2010). There is a lack of a precise definition or specification of the RIA term in the literature, however, some general characteristics can be emphasized, which distinguish RIAs from conventional Web applications.

- RIAs offers enhanced graphical user interface capabilities. In addition to visually more appealing and flexible design possibilities, RIAs enable sophisticated user interaction functionalities such as live validation, auto complete, or drag&drop. Furthermore, RIAs support improved interactive multimedia presentations like embedding animations, images, audio or video files. RIA interfaces are usually designed as a single application, which manages all user interactivity on the client side. This permits loading, displaying, and updating individual page elements without refreshing the full application at each user interaction. In addition, with RIAs it is possible to change interfaces dynamically by loading parts of the presentation logic at runtime, such as extended interaction events or customizing of existing widgets or elements.

- RIAs enable client-side processing, by moving part of the computation from the server to the client. Offloading business logic to the client leads to quicker responsiveness and to an optimized communication behavior. This allows users to navigate, filter, and

manipulate data using the computation power of clients' workstation before sending it to the server.

- RIAs permit asynchronous server communication. Both, the server and the client can trigger communication, while program components in the client stand ready to receive and execute asynchronous server commands. This kind of bidirectionality avoids unnecessary server roundtrips as it is usually the case in thin-client applications.

5.2.2 Science2.0

The very nature of the World Wide Web is sharing and exchanging arbitrary types of information in a community-driven environment. This paradigm recently has been pushed to a next level towards Web2.0. Information and knowledge sharing has always been a crucial part in scientific work, resulting in a large number of cross-disciplinary research collaborations. Thus, the question is raised, how science could substantially benefit from the recent Web developments allowing for new and more efficient ways to solve scientific problems. This question led to a new phrase: *Science2.0* (Shneiderman, 2008; Waldrop, 2008). The idea of Science2.0 is to use Web2.0 technologies to share several aspects of scientific work with colleagues and interested third parties. Here, *sharing* is at the heart of Science2.0. Web2.0 software tools and principles enable sharing of ideas, hypotheses, questions, workflows, data, experimental results, and even applications efficiently accessible via the Internet from anywhere. The social network characteristic of the Web2.0 fosters the communication and cooperation among collaborating researchers, taking science to a new level of connectivity and interactivity. Science2.0 becomes an important complement to the existing system of peer-reviewed publications in journals and conference proceedings, which is traditionally considered as "the" medium of scientific communication. Using new collaborative technologies could enhance science communication, both before publication, when generating ideas and hypotheses, and after publication, when discussing results.

The development of *BioIMAX* is focussed on this concept of Science2.0. *BioIMAX* is the attempt to create an collaborative scientific environment for bioimage analysis exploiting the recent developments and social characteristics of modern Web2.0 technologies, i.e., sharing and discussing data and results and collaborative application of RIA based analysis.

5.2.3 RIA frameworks

RIAs can be developed and implemented with several different technologies, which can be generally divided into two categories. The first category comprises RIA implementation frameworks that are based on using standardized and established Web technologies such as the AJAX framework natively supported by most Web browsers. The second type of RIA development technologies produce solutions requiring additional third-party plug-ins and runtime environments, e.g., Adobe Flash/Flex or Microsoft Silverlight, allowing for running their own native code within the Web browser. In the following, an overview about the most recent well-known RIA frameworks is given including the Adobe Flex framework, which is the RIA technology used to develop and implement the *BioIMAX* client-side.

Google Web Toolkit

The Google Web Toolkit (GWT)[6] was introduced by Google in 2006. It is an open source framework for the development of Web applications based on the AJAX framework and has been built under the Apache license[7]. GWT allows the creation of complex AJAX (JavaScript and HTML) client front-end applications in Java. With GWT, AJAX applications are completely written in Java and the resulting sources are compiled to highly optimized JavaScript by a Java-to-JavaScript compiler, which will be deployed in the Web browser. This allows to produce Web applications that run on all major browsers throughout any operating system without requiring any proprietary plug-ins. The GWT SDK (Software Development Kit) ships with a large variety of core Java APIs and Widgets. Graphical components can be implemented using known Java graphics libraries such as Swing or SWT, but will rendered in dynamically created HTML, which enables one to design and manipulate Widget appearances employing CSS[8]. For the generation of GWT applications, developers are writing and running classic Java applications in an hosted mode on their personal machine employing the Java Runtime Environment. Using their favorite integrated development environment (IDE) including typical features like compiling, debugging, refactoring, and testing, developers are able to create both client- and server-side applications in Java. Server-side code is packaged in services facilitating the communication between the client and the server by using asynchronous remote procedure calls accessed with a remote servlet on the server.

Microsoft Silverlight

Microsoft Silverlight[9] is an application framework for developing and delivering rich and interactive Web applications. Silverlight (SL) is a relatively young technology, its first version has been released by Microsoft in 2007. Just recently, the latest version, Silverlight 5, has been released. Since the first version, Silverlight provides a simple presentation framework using XAML (eXtensible Application Markup Language), which is a declarative XML-based markup language. XAML forms the essential part in the Windows Presentation Framework (WPF)[10], Microsoft has developed for graphical display and animation of Windows desktop applications. With XAML the layout of the SL user interfaces (UI) is defined and declared. Allowing for creation of simple animations with geometrical primitives or text and integrating video-based content in the early SL stages, the recent version enables the integration of complex graphical UI elements, controls and widgets such as DataGrids, TreeViews or various layout panels and their interactivity. One of the major features since SL version 2 is the integration of the .NET framework into SL. From then on, SL is based on a complete implementation of the common language runtime (CLR)[11], which includes a wide variety of functions of .NET 3.5, insofar they are reasonable for browser applications. In contrast to

[6] http://code.google.com/webtoolkit/overview.html
[7] http://code.google.com/webtoolkit/terms.html
[8] http://www.w3.org/Style/CSS/
[9] http://www.microsoft.com/silverlight/
[10] http://msdn.microsoft.com/en-us/library/ms754130.aspx
[11] http://msdn.microsoft.com/en-us/library/8bs2ecf4%28v=VS.110%29.aspx

SL 1, where interactivity of the UI components was achieved exclusively by using JavaScript running in the browser, SL 2 and further versions allow developers to basically apply arbitrary .NET languages for application development with SL. This fact can be considered as an advantage over other RIA development frameworks, since the development of SL applications requires no learning of new programming languages, so that developers using any .NET language can generate SL Web applications. Another feature in this context is, that both the client applications and the server side programs can be implemented using .NET languages. .NET code is compiled into a Microsoft Dynamic-Link Library (DLL) and the application together with XAML is deployed on a Web server such as from Apache or from Microsoft (IIS)[12]. The communication with the Web server is performed by using the standard HTTP_GET method. Running SL applications in the Web browser requires the installation of a proprietary browser plug-in being the run-time environment for SL, which is available for multiple browsers including Internet Explorer, Firefox, and Safari supported for Microsoft Windows and Mac OS X operating systems. In order to bring SL to Linux, FreeBSD or other open source platforms, a free software implementation called Moonlight[13] has been developed by Novell in cooperation with Microsoft, which basically includes functionalities of SL versions 1 and 2. For the development and implementation of SL applications, Microsoft provides a complex ecosystem for SL designers and programmers such as Visual Studio 2008 or Expression Blend.

JavaFX

As in the case of Microsofts Silverlight, JavaFX[14] is another framework for creating modern rich Web applications. Still a young technology, first release was at the end of 2008, JavaFX aims at providing a development environment for producing cross-platform applications, which can be run either on mobile phones, on desktop computers or in Web browsers. JavaFX is a branch of the Java specification and family provided by Sun Microsystems. For the implementation of JavaFX applications it provides a statically typed, declarative scripting language named JavaFX Script. It allows to integrate a large amount of functionalities and interfaces of the Java class library, since the compiler translates JavaFX Script code into normal Java byte code, which will be run on a Java virtual machine. Due to this fact, JavaFX developers can make use of virtually the entire Java world. For the graphical presentation of JavaFX applications JavaFX Script includes separate class libraries that allows developers to integrate existing Java graphics libraries such as Swing or Java 2D in an easy and declarative way. Single UI components can be designed or modified with CSS. In addition, JavaFX incorporates a graphics engine, which is capable of taking advantage of hardware graphics accelerators and it comes with a set of plug-ins that enables the direct integration of Adobe Photoshop and Illustrator objects. JavaFX applications are deployed as JavaApplets in the Web browser downloaded from a Web server using the Java Runtime Environment (JRE). The connection to the Web server is realized with HTTP_GET, REST or Webser-

[12] http://www.iis.net/overview
[13] http://www.mono-project.com/Moonlight
[14] http://javafx.com/

vices. JavaFX applications are able to run on any operating system and Web browser that has the latest JRE installed. JavaFX provides plug-ins for integrated development environments (IDE) such as Eclipse or Netbeans, which allows a more comfortable implementation of JavaFX applications. With JavaFX it will be possible to develop rich Web applications, either for experienced Java developers or developers without special knowledge in the Java programming language.

HTML 5

HTML5 is going to become the latest revision of the HTML standard and is still a work in progress and under heavy development. Since 2007, HTML5 is developed in a cooperation between the World Wide Web Consortium (W3C) and the Web Hypertext Application Technology Working Group (WHATWG)[15]. The major goal of HTML5 is to establish a new standard for the generation of sophisticated Web applications, which meets the increased demands of the new Web2.0. The developers of HTML5 argue, that using a new Web standard in the future will reduce the needs for proprietary Web technologies such as Adobe Flash/Flex or Microsoft Silverlight, whose plug-in based concepts are often considered as a drawback. At present, an intensive work is carried out, in order to develop a comprehensive specification of the HTML5 standard[16], which introduces various new capabilities and features, touching not only HTML itself. HTML5 is considered as an umbrella term that covers a collection of different standard Web technologies. HTML5 still remains a markup language exclusively used to format data and information. Important aspects like interaction and graphical design still has to be covered by traditional technologies such as JavaScript, CSS and DOM. Thus, even these Web standard has to be adapted to be compatible to the new HTML5 standard. HTML5 is not intended to fully replace the current version HTML4.01, but it extends it with new features, so that Web applications using HTML4.01 are still compatible with HTML5.

Besides new markup elements for improved semantical document structure and new graphical form elements with more functionalities, in particular the new media elements and the canvas element are drawing attention. These new elements make it easy to directly include and handle multimedia content such as audio or video and graphical content into HTML, without using third-party proprietary plug-ins or APIs. As mentioned before the, the HTML5 family also comprises the latest versions of JavaScript, CSS, and DOM. The HTML5 specification defines how the new elements interact with JavaScript and DOM, resulting in a new JavaScript API allowing for enhanced interaction capabilities, e.g., drag&drop, dynamic drawing using the canvas element or multithreading. The CSS3 specification includes new and adapted styles for customizing the presentation of HTML5 interface capabilities.

In sum, since it is not yet an official standard and not all features of HTML5 are supported by all major Web browsers, there is a great disagreement, whether HTML5 should be considered as a part of the RIA family or whether it will in fact supersedes the established plug-in based RIA technologies. It will take several years of development and testing time until it is possible to verify to what extend HTML5 is comparable to existing RIA technologies.

[15] http://www.whatwg.org/
[16] http://dev.w3.org/html5/spec/Overview.html

Adobe Flex

In 1996, Macromedia introduced a new multimedia platform named *Flash*. Flash enables Web designers and developers to enrich Web pages with both multimedia content and enhanced Web page interactivity. In its beginning stages, Flash was mainly used as animation tool, e.g., for intro animations on Web pages, animated banner ads or for embedding audio or video files. Since Flash has been extended with the *ActionScript* scripting language in version 4, it was possible to program interactive browser based Web applications completely based on Flash. Early Web pages realized with Flash using ActionScript can be considered as the first RIAs. Flash is both the origin of the RIA era (Macromedia first coined the term RIA (Allaire, 2002)) and the most frequently applied RIA technology in these days. Flash applications are running in the Web browser using the proprietary, freely available *Flash Player* (Flash runtime environment), which is available as a plug-in for all major browsers and operating systems. However, developing RIAs purely with Flash is challenging for "traditional" programmers, since it is rather intended to be a tool for creating media applications than a framework for developing and implementing sophisticated browser based applications.

For this reason, Macromedia and later Adobe have released an extended development framework called *Flex*[17], especially designed to develop and deploy RIAs based on the Adobe Flash platform. The goal of the Adobe Flex framework is to provide a workflow and programming model, which allows developers who are not familiar with designing Flash applications can easily create powerful and robust RIAs. The Flex framework basically consists of two important components, the declarative, XML-based markup language MXML and ActionScript (AS) known from the Flash platform. With MXML, the layout and the behavior of the display elements of the user interfaces is defined, whereas AS is responsible for client logic, procedural control and event handling, in order to achieve dynamic interactivity. In recent years, AS (current version 3) has evolved to an object-oriented programming language based on the ECMAScript standard[18] and has a similar syntax and semantic as Java or C# and constitutes the core language of the Flash Player. AS code can directly be added to an MXML file identified as separate script block or imported as external files. Technically, MXML describes AS on a higher level of abstraction, which provides functionalities and components of AS in a more simple way. During the compilation process, MXML code is parsed and converted to AS code in an intermediate step. The resulting AS classes along with the user-defined AS classes are compiled into Flash bytecode stored in an executable SWF (Shockwave Flash) file. This SWF file can be deployed in the Web browser using the Flash Player plug-in.

In addition to MXML and AS, the Flex framework provides a comprehensive collection of standard components to design and lay out the graphical user interface. It includes a wide variety of containers such as layout or navigator containers, controls (buttons, lists, trees and grids) and predefined charting and graph components. These components can be individually customized and their graphical appearance modified through style properties or inline or external CSS, which enables developers to freely design their own user interfaces

[17] http://flex.org/
[18] http://www.ecma-international.org/publications/standards/Ecma-262.htm

depending on their personal needs. Other features such as drag&drop, animation effects, model dialogs, or form validations round out the Flex application framework.

Flex applications can work with several server technologies. Therefore, Flex allows applications to access data from the persistence layer or trigger server-side business logic at runtime in three different ways: using HTTP calls usually by exchanging XML-based data, through requesting SOAP (Simple Object Access Protocol) (Curbera et al., 2002) based Web services and through AMF (Action Message Format) based services via remote objects. Each of these methods are supported by separate Flex classes or MXML tags, respectively that allows an easy implementation of communication issues.

The Flex framework serves either with a free open source SDK containing all predefined class libraries, application services and a standalone command line compiler or with a proprietary commercial IDE, the Flash Builder, built on Eclipse. The IDE also contains the Flex SDK and other useful development tools such as a debugger or a visual designer and is available free of charge for non-commercial use, e.g., by students or unemployed developers.

In sum, Adobe Flex is a powerful and the most established framework for the development of RIAs, not least through the wide variety of predefined and flexible user interface components, which allows developers to rapidly and effectively create RIAs tailored to their needs. Since the Flash Player has nearly 95% penetration rate, i.e., 95% of Internet connected computers having Flash Player installed, it creates a massive audience for Flash/Flex Web applications. Thus, using Flash/Flex ensures to reach the widest possible audience without bothering users to install or maintain additional plug-ins or software. This important fact led to the decision to employ the Adobe Flex framework for the development of the Web-based user interface front-end of the *BioIMAX* platform. The complete overview about the architecture of the *BioIMAX* system is given in the next section.

5.2.4 *BioIMAX* architecture

The diagram in Figure 5.4 schematically illustrates the entire architecture of the *BioIMAX* system. The architecture can be regarded as a four-tier architecture, with the *RIA client* representing one tier, a *Web server* representing the second tier, an *application server* as third tier, and a *database server* representing the fourth and final tier. Below, the technical realization and the role of each of the four tiers and their interconnections are described in detail.

Rich Internet Application (first tier)

As mentioned before, the *BioIMAX* user interface (UI) has been realized with Adobe Flex and the resulting Flash application is run in a Web browser using the Flash Player runtime environment. In this architecture, the Web browser represents the *client* that is running locally on the user's workstation. Through the Flash based RIA, the Web browser allows *BioIMAX* users

- to access data such as multivariate images or analysis results stored on a remote server

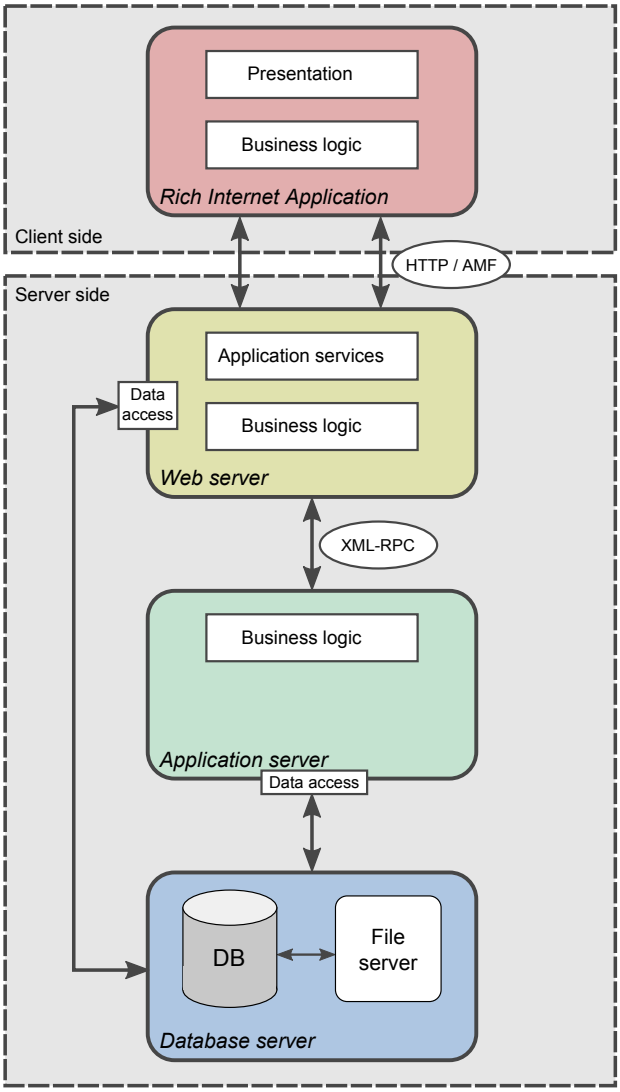


Figure 5.4: Architecture of the *BioIMAX* system. The system is based on a client-server architecture. The client is represented by a Web browser that is running a *Rich Internet Application*, which is responsible for data presentation and provides a graphical interface incorporating substantial business logic that enables users to dynamically and interactively manipulate and process parts of the data directly on the client side. Through the RIA, the client communicates with a remote server, in order to access and process data via a *Web server* and to call server-side services triggering computational expensive data processing routines performed on a powerful *application server*. Via data access interfaces the Web server and the application server retrieve data from and store data to a *database server*. The Database server consists of the relational database management system *MySQL* implementing the *BioIMAX* data model and a separate *file server* that holds physical data such as images or complex analysis data.

- to interactively display, navigate, filter, and manipulate the data directly on the client-side

- to call server-side services in order to trigger computational expensive data processing routines

in an environment resembling desktop applications via the Internet. Conventional client-side Web front-ends are only responsible for the presentation of the data, whereas the manipulation of the data is performed on the server side initiated by simple and limited client-side user interface options. In contrast, using RIA technologies, considerable parts of the computation can be shifted to the client side. This allows more dynamic interaction capabilities while at the same time reducing the server and network traffic costs. Thus, the functionalities and logic of the *BioIMAX* RIA front-end basically can be divided into two parts, *presentation* and *business logic*.

Presentation

The presentation part refers to the graphical design of the *BioIMAX* front-end as well as to the display of the different data structures that are called from the server and transferred to client. The graphical front-end is basically designed and structured using the standard graphical UI elements predefined by the Flex framework, e.g., layout containers, input controls or advanced components such as datagrids or charting facilities. The style of the components is customized with CSS. In addition, the *BioIMAX* front-end includes also custom UI components for advanced interface capabilities that could not be realized with only one standard Flex component, instead they consist of a combination of standard UI elements creating new individual UI components.

BioIMAX comprises a large number of different functionalities that can be divided into distinct categories. In order to create a well-structured and clearly organized graphical user interface, which should allow *BioIMAX* users to clearly navigate through the system and its data, the *BioIMAX* front-end is split into several parts depending on the category of functionalities. Each category is implemented as a stand-alone Flash application considered as a toolbox that includes all functionalities and interface facilities of the respective category. These toolboxes are stored as individual SWF files that are triggered through the *main* application (see Figure 5.5(a) for a schematic illustration). Technically, the running main Flash application, which is embedded in HTML code, calls and initiates the external SWF file of a specific toolbox and simultaneously sends a set of connection parameters to the toolbox application. Using a JavaScript routine the SWF of a toolbox runs in a separate Web browser window. This toolbox concept has several advantages. First, it avoids a cluttered and overloaded graphical user interface. Second, it enables *BioIMAX* users to clearly arrange parts of application on their desktop screen, e.g., on a multi-screen system (see Figure 5.5(b)). Finally, it considerably reduces loading time of the application, since initially only the main application is started and other toolboxes can be loaded when required at runtime.

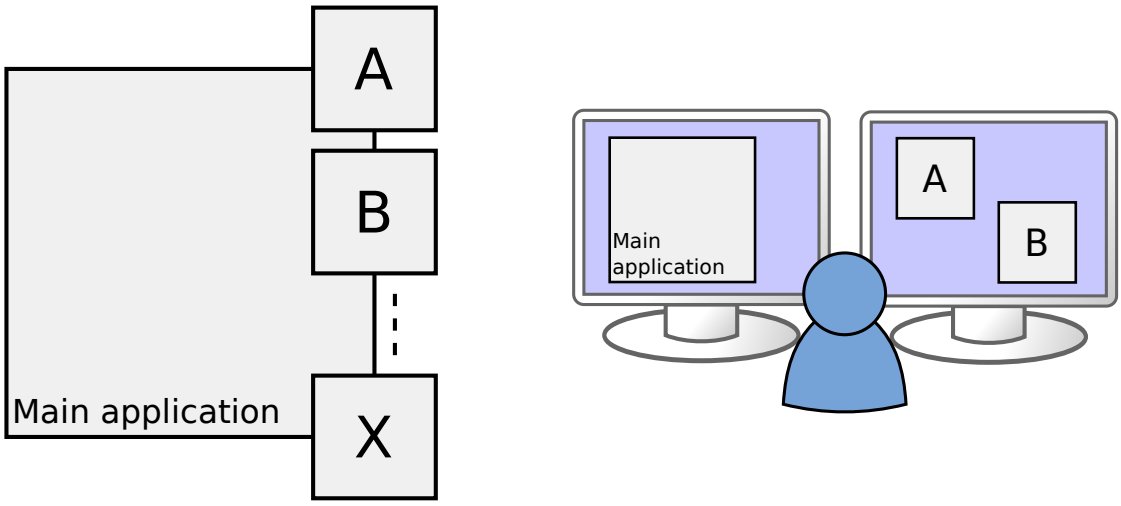


Figure 5.5: The toolbox concept. (a) shows a sketch of partitioning different categories of functionalities into different toolbox interfaces that are running in separate browser windows independently from the *BioIMAX* main application. This interface architecture has two main advantages. First, it reduces loading time of the entire application considerably, since the toolboxes A, B, ..., X are initiated at runtime. Second, it enables users to clearly arrange parts of the application on their desktop screen, e.g., on a multi-screen system (b), whereby cluttered and overloaded user interfaces are avoided.

Business logic

The business logic part of the *BioIMAX* RIA is responsible for dynamically and interactively manipulating and processing data directly on the client side. Data manipulation can be achieved by using the interactivity functions and capabilities of the predefined Flex UI components, e.g., navigating, filtering, or sorting entries in the Flex datagrid. Furthermore, the *BioIMAX* client includes customized interactions between different UI components, in particular if one UI component requires data as input from another UI component and vice versa, i.e., the changed state of presented data in one UI component has a direct impact on one or more other UI components. In addition to using the interaction functionalities inherent in the Flex UI components, the *BioIMAX* RIA incorporates extended data processing routines, which are computed in the clients background, e.g., extracting user-selected image regions of interest and transforming the pixel values of the regions into an appropriate data structure, which is capable to be read in by other UI components of the presentation layer. All data interactivity and manipulation is achieved using ActionScript classes and functions.

The *BioIMAX* RIA front-end is implemented based on the *Model-View-Controller* (MVC) software pattern as far as possible. The idea of the MVC architecture is to manage the presentation logic and the business logic separately without affecting the other. This should

enable a flexible and an easy to maintain and modify design of the application. The *model* represents the data domain, which is independent from the presentation and business domain. The *view* represents the user interface (presentation) and manages the rendering of the *model* in a suitable way allowing for user interaction. The *controller* holds the business logic and is responsible for receiving user input and triggering requests to the model domain that performs specific actions based on the user input.

The design and all functionalities of the *BioIMAX* Web front-end are described and illustrated in detail in the next Chapter 6.

Web server (second tier)

The second building block of the *BioIMAX* system architecture is the *Web server*. A Web server basically serves as an connection entity located between the client application and the data stored on the server. A client application, commonly a Web application, formulates a request for a specific resource on the server and sends this request via a transfer protocol such as HTTP or via remote procedure calls (RPC) to the Web server. The Web server processes the request and delivers the content of the resource to the client. The client, in turn, uses the response to update the current state of the Web application. In addition to simply delivering static content, e.g., HTML documents including any additional content (images, CSS, JavaScript), many generic Web servers support dynamic creation of Web content by server-side scripting. Information from different sources such as different databases can be dynamically collected and combined before returning the content to the client. In this way, the Web server and its behavior can be individually designed and configured by using scripting languages. Depending on the computation power of the host server machine, higher scripting languages even allows to implement sophisticated business logic applications. Thus, the Web server is not only responsible to collect data from different sources, but also to manipulate the data using specific processing routines. Here, the Web server forms the interface between the client application and server-side databases. The Web server receives a request from the client and executes the requested routines or scripts. These scripts process data either from databases and delivers the requested results to the client or data sent from the client, in order to store the processed results in a database.

For the *BioIMAX* architecture the *Apache HTTP Server*[19] was used. The Apache Web server is an open source and cross-platform implementation of an HTTP server, maintained by the Apache Software Foundation under the Apache license v2.0[20], and is the most frequently used Web server in the Internet[21].

In combination with the Apache Web server, *PHP*[22] was used to realize server-side programming. PHP, which stands for the recursive acronym *PHP: Hypertext Preprocessor*, is a popular open source and cross-platform scripting language, which has originally been devel-

[19] http://httpd.apache.org/
[20] http://www.apache.org/licenses/LICENSE-2.0
[21] http://news.netcraft.com/archives/2011/05/02/may-2011-web-server-survey.html
[22] http://www.php.net/

oped to generate dynamic HTML output to be sent to the client. However, PHP is not only limited to generate HTML output, it includes a wide range of libraries, classes, and functionalities, which allows programmers to implement sophisticated server-side processing routines producing various other types of output such as images, animations, or XML. Finally, one of the most essential features of PHP is its support for a large variety of databases.

Regarding the *BioIMAX* architecture, the function of the Web server with PHP can be characterized as follows:

- Loading requested data from a database and sending the retrieved data back to the client or storing data that has been sent from the client to the Web server in a database.
- Performing parts of the server-side business logic. The Web server includes PHP routines that process data either for storage or for presentation to the client application, e.g., manipulating image data or verifying user login data.
- Initiating computational expensive routines on an application server (see next paragraph *Application server*).

For the connection and the data transfer between the client (Flex application) and the Web server the *BioIMAX* architecture uses two different communication techniques: via *HTTPService* or via *Action Message Format* (AMF) based services. Adobe Flex provides for both techniques specific classes, which are instantiated within the application and are able to send and retrieve data to and from the Web server. These techniques are based on the *remote procedure call* (RPC) principle. In the RPC concept, the communication starts with the client sending a request to the server (the Web server in this case) that executes a specified procedure with the supplied parameters. After processing the request, the Web server sends back a response to a callback function on the client continuing its application process.

HTTPService This technique allows the Flex application to execute conventional HTTP calls such as *GET* or *POST*, in order to exchange simple XML based information. In *BioIMAX*, HTTPServices are mainly used to initiate computational expensive routines running on a remote compute server. Here, the communication is unidirectional, i.e., the required parameters have to be send only from the client to the Web server. The Web server provides specific PHP services that process the request and calls programs on the compute server, which store the results directly in the database, so that the client does not need any immediate response from the Web server (a detailed description of the *BioIMAX* application server and its connectivity is given in the next paragraph). Although this communication technique is limited to transferring only XML-based information and the transfer process is slower than with other techniques, it requires considerably lesser programming effort and is sufficient for the aforementioned demands.

AMFPHP For a bidirectional transfer of larger and more complex data objects, Adobe Flex supports an advanced client-server communication protocol AMFPHP[23], which

[23] http://www.silexlabs.org/amfphp/

is free open source PHP implementation of the Action Message Format (AMF) and is based on remote object classes. Except for the initiation of services on a compute server, AMFPHP is used for any client-server communication issues in the *BioIMAX* architecture. AMFPHP allows Flex applications to communicate directly with PHP class objects on the server by binary serialization of ActionScript native data types and complex objects into AMF requests to be sent to server-side services. The Web server deserializes the request, finds the corresponding remote class, calls the remote method using the specified parameters, and returns the data as serialized objects. The serialization and deserialization into a binary format is natively supported by the Flash Player and is generally more compact than other representations, e.g., XML, which makes AMFPHP one of the fastest client server communication protocol available for Flash applications. Another advantage of AMFPHP using the latest AMF3 is, that it supports resources from database connections, i.e., directly returning database queries from the remote method, which will be understood and handled by AMFPHP. This simplifies the implementation and development process considerably.

Application server (third tier)

As already mentioned in the last paragraph, the *BioIMAX* architecture includes a remote *application server* constituting the third component of the this architecture. The application server is exclusively used to perform high-level business logic on datasets stored in the *BioIMAX* database. The principle of the application server is similar to those of the Web server. The application server provides sophisticated services that are initiated and executed with specific parameters via a remote request. However, the requests are not directly sent from the client application to the application server, but via the Web server as an intermediate layer. The client sends an initial HTTP request to services located on the Web server (see last paragraph for details), which are designed to formulate a new RPC to be sent to the application server.

The application server implements all services that are too computational expensive and time-consuming to be made available on the Web server, e.g., supervised or unsupervised machine learning algorithms. In the *BioIMAX* architecture these routines are implemented in C++ using several additional libraries such as machine learning or image processing libraries. The routines are executed on a powerful compute cluster allowing for parallel programming resulting in far better performance of the programs. Once the application server receives a request from the Web server, it calls the dedicated routine with the supplied parameters, which loads and processes data from the database, e.g., images, and stores the results in the database. Finally, the user, who has initiated the request, is informed via email, when the procedure has finished and the results are available to be visualized and processed by the client application.

The Web server and the application server are connected via XML-RPC[24] (Laurent et al., 2001), which is another transfer protocol to make procedure calls between software running

[24] http://xmlrpc.scripting.com

on disparate operating systems or servers. XML-RPC uses XML to encode its calls and HTTP requests as transport mechanism. XML-RPC supports implementation in several programming languages, making it a powerful and generic application.

The communication between the client application, the Web server and the application server is unidirectional (see Figure 5.4). Results produced on the application server are directly stored in the database and can be loaded with the respective visualization or analysis tools of the *BioIMAX* client application.

This principle of calling and using sophisticated software programs hosted on a powerful remote compute server accessible via the Internet is considered as *Software as a Service* (SaaS), which is a common software delivery model usually used for most business applications and refers to the evolving technology named *Cloud Computing*. Using this software model in *BioIMAX* allows users to apply sophisticated algorithms to their data called from their Web browsers, without installing and maintaining the software on their local client workstation. This saves a lot of computation costs and time, which is of particular benefit for *BioIMAX* users.

Database server (fourth tier)

The last and final tier of the *BioIMAX* architecture is the database server. Here, all types of data imported into the system or produced by the system are organized and managed. The relational database management system (RDBMS) *MySQL* was used to implement the conceptional data model described in Section 5.1. Here, all entities of the *BioIMAX* data model, which is illustrated in Figure 5.3, are translated into relations or tables, respectively. The *BioIMAX* database can be accessed by the *BioIMAX* Web server as well as the application server. Those types of data that cannot be directly mapped to the relational database model, e.g., images, are stored on a separate file server. The MySQL database holds the information, where the external data is located on the file server. Due to the combination of a RDBMS with a dedicated file server, it is possible to incorporate complex data structures into the *BioIMAX* system.

5.3 Summary

In light of the requirements formulated in Chapter 4, this chapter has pointed out the definition of the *BioIMAX* data model that has been developed to model the variety of data and processes collected within the system. Based on an excursion into the history of the World Wide Web and an overview about current trends in Web technologies, important aspects of the system architecture for the realization of a Web-based RIA front-end and its corresponding back end architecture have been addressed and illustrated. The next chapter is focussed on the realization and implementation of the *BioIMAX* Web front-end, emphasizing on the different user interface capabilities.

CHAPTER 6

Implementation and Methods

In the previous chapter the formal design and the architecture of the *BioIMAX* system was presented and described. On this basis, the following chapter is focussed on the realization and implementation of the system. It emphasize both, the graphical design of the user interface of the *BioIMAX* Web front-end and the functionalities of single components and tools either as client-side interactivity logic or as user interface in order to request server-side services or logic. Additionally, this chapter regularly addresses aspects regarding data handling, in particular while describing the analysis and exploration tools, provided they are relevant for the respective *BioIMAX* components:

Loading Which data has to be loaded while starting a tool? How is the data loaded from the server?

Processing Which data processing routines are available? What are the goals and the purposes of these functionalities?

Storing How is the output of a specific tool stored on the database server? How are the data structures realized in the *BioIMAX* data model?

The structure of this chapter generally follows a possible workflow from a potential user perspective: beginning with the login followed by the import, retrieval and management of image data and concluding with several data analysis and collaboration aspects.

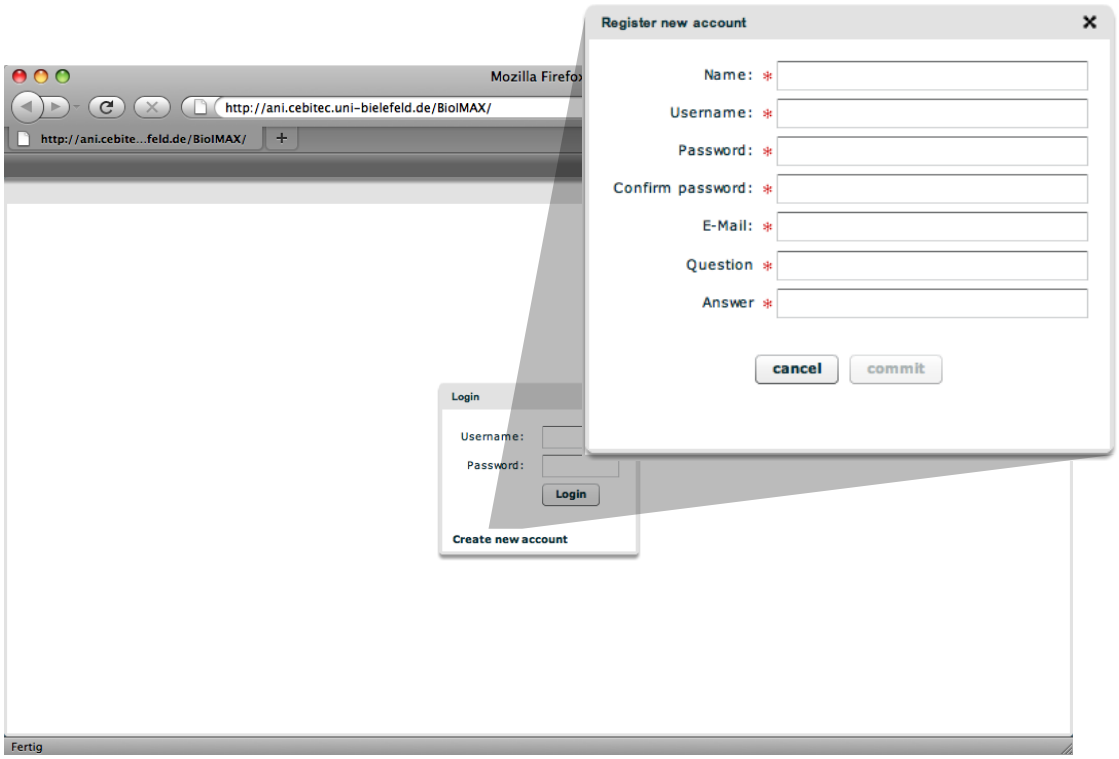


Figure 6.1: Registration and login procedure. Prior to the first login to the *BioIMAX* system, new users have to create a unique *BioIMAX* account via the registration form accessible on the startup screen of the *BioIMAX* application. After a confirmation process users can login to the system using their defined username and password.

6.1 Start working with *BioIMAX*

The *BioIMAX* system can be accessed at http://ani.cebitec.uni-bielefeld.de/BioIMAX with a standard Web browser provided that the latest version of the Adobe Flash Player is installed. Similar to other Web portals such as social network or messaging platforms, users need to register a new *BioIMAX* account prior to the first usage (see Figure 6.1). After filling out the registration form a request is sent to the *BioIMAX* administration staff, which has to confirm the registration. With the confirmation process a new user instance with a unique ID is created and inserted into the database (see *User* relation in Figure 5.3), which holds all user account details. Finally, the user is informed via the given email address, that the registration was successful and that a *BioIMAX* account has been generated.

Once a user has been authenticated by a username and password login procedure, she or he is presented with the *BioIMAX* main page (see Figure 6.2). The main page resembles design aspects of well-known social media platforms, in order to create a personalized environment. It is the starting page providing access to all *BioIMAX* facilities, i.e., users are able to update

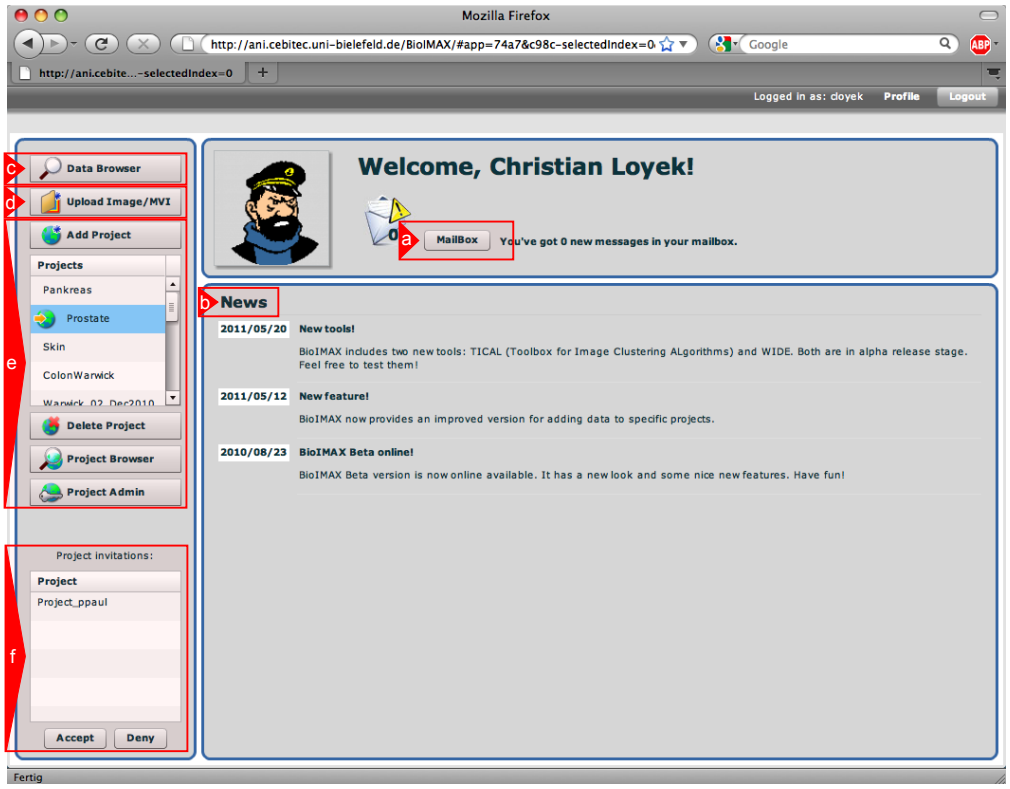


Figure 6.2: The *BioIMAX* main page. This is the starting page providing access to all *BioIMAX* facilities. It is divided into three parts: First, a personalized header with access to the system internal messaging system (a). Second, a news box serving as information area displaying the latest news on *BioIMAX*, e.g., service messages, system upgrade announcements or new tools and functionalities (b). Third, a navigation toolbar including the following data handling capabilities: Starting a data browser to search and retrieve data stored in the database (c), uploading new image data into the system (d) and organizing and sharing data with other *BioIMAX* users via user-defined projects (e,f).

their account details, to get in contact with other registered users, to upload, organize, and retrieve image data and results, and to start several analysis and exploration tasks. The *BioIMAX* main page is divided into three areas:

Header The main page header creates a personalized Web environment enriched by an individual welcome text and a user-defined avatar and provides access to the system internal messaging system (see Figure 6.2(a)).

News box The news box keeps all users informed about the latest news on *BioIMAX* such as service messages, system upgrade announcements, bug reports, or about new or improved tools and functionalities (see Figure 6.2(b)).

Navigation toolbar This panel is the starting point for all kinds of data handling capabilities

available to the users in *BioIMAX*. It provides access to a data browser (see Figure 6.2(c)), in order to search and retrieve the variety of different data (see Section 6.2 for detailed explanation of the data browser). Additionally, the navigation toolbar serves as an interface to integrate new image data into the system (see Figure 6.2(d)) and to organize and share data with other users via user-defined projects (see Figure 6.2(e,f)). The latter two functionalities are described in the following paragraphs.

6.1.1 Importing MVI data

For the import of new image data *BioIMAX* includes an upload routine that allows users to quickly and simply integrate new multivariate ImageStacks into the *BioIMAX* database. The upload procedure is controlled through an easy-to-use form offered by the graphical interface of the *BioIMAX* Web front-end and comprises three basic steps illustrated in Figure 6.3.

Step 1 Select an appropriate name for the ImageStack.

Step 2 Select all single images that are associated to the ImageStack via a standard file chooser dialog provided by the Adobe Flex framework. In the current status of *BioIMAX*, only image files with the following file formats are supported: JPEG, PNG and GIF. Subsequently, the selected images are displayed as a list of file names encouraging the users to assign each image with a descriptive textual tag via respective text fields (see Figure 6.3(a)). Here, the user has the option either to enter free text or to select predefined tags, which are directly filtered and suggested from an extensive tag list by analyzing user keystrokes while typing (see Figure 6.3(b)). Since *BioIMAX* is frequently used for MVI data obtained with multifluorescence imaging techniques (see Chapter 2.1.2), the list of predefined tags mainly contains entries of an antibody dictionary. This dictionary is based on details and information about antibodies and biological reagents freely available on *Linscott's Directory of Immunological & Biological Reagents* [1]. In the data model, the antibody dictionary is represented by the *Antidict* relation (see Figure 5.3) and the database comprises approximately 300.000 instances. A single tag is characterized by specific attributes, i.e., specificity (tag name), clone, host and reactive species and a supplier. This should allow the users to specify tags for their images more precisely. In case a tag is not available in the list or the attributes does not match with the antibody that has been used for staining, it is possible to extend the current list with new entries, in order to reuse them for further images of the same type. If the text field is left blank, the file name of the respective image is selected and associated as descriptive tag.

Step 3 In the last step, users have the opportunity to save the current selection of image tags in a separate tag list that will be associated to the current user. This allows them to reuse their personal tag list for other ImageStacks containing the same number of images with the same antibody profile. Using a saved list of tags (see Figure 6.3(c)),

[1] http://www.linscottsdirectory.com/search/antibodies

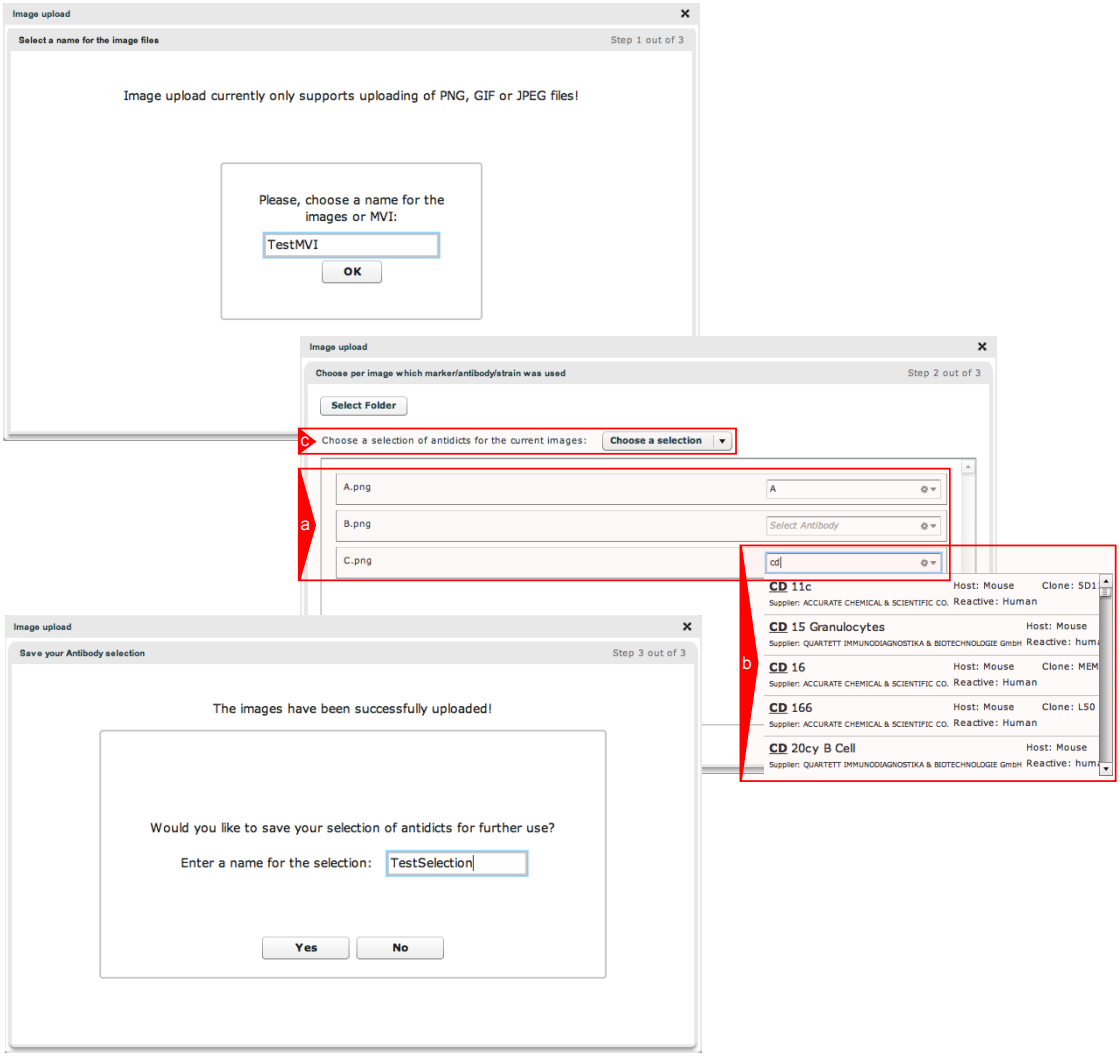


Figure 6.3: Importing ImageStacks to *BioIMAX*. The import of new multivariate image data comprises a three-step upload procedure. In the first step the user defines a characteristic name for the ImageStack. In a second step single image files belonging to one ImageStack have to be selected via a file chooser dialog that will be displayed in a list of image files (a). Afterwards users are encouraged to assign each image with a descriptive textual tag either by free text or by selecting predefined tags (b). Finally, users are asked to store the current list of defined image tags, in order to reuse them for other ImageStacks showing the same "tag profile". This prevents cumbersome and time-consuming retyping the same list of tags for each new ImageStack again (c).

cumbersome and time-consuming retyping single tags for each image again is avoided, which applies in particular to ImageStacks with an increased number of images. For this reason, if users decide to save the current selection to the database, they have to enter a title for the selection in this step.

Data processing and storage

After selecting the images and specifying any required information, the image files are loaded by the client application and are combined with the other information such as the image tags to a value object (VO). This VO is send from the client to a remote object located on the Web server. The Web server receives the VO and executes the requested service responsible for preprocessing and storing new ImageStacks to the database using the delivered parameters of the VO.

Before any data is stored in the database or on the file server during the upload process, some automatic preprocessing steps are carried out on the Web server. As mentioned in the requirements in Chapter 4.3, *BioIMAX* requires special copies of the original ImageStack, i.e., one containing intensity normalized images with an appropriate file format enabling the Web browser to display the images and one containing thumbnails of the original images for preview issues. These copies are generated on the Web server via routines implemented in PHP. As a consequence, three ImageStacks playing different roles within *BioIMAX*, but originate from only one original ImageStack have to be stored in the database in a suitable way. Additionally, some further meta information are generated on the Web server such as number of images in the ImageStack, image width and height or the acquisition date. Once the preprocessing routines on the Web server are completed, all accumulated data, i.e., the three ImageStacks and their additional information, are stored on the database server according to the defined data model introduced in Chapter 5.1. A detailed description of this process is given in the following, which continually refers to entities and relations of the *BioIMAX* data model illustrated in Figure 5.3.

- First, a new instance of the *ImageStack* relation is created that can be considered as an abstract instance, in order to uniquely identify image data of the three newly generated ImageStacks and to link their associated or derived data to this unique instance. This *ImageStack* instance is characterized by different attributes. The attribute *user_id* is a foreign key and refers to the *User* instance representing the user, who is currently importing the original ImageStack. This user can be considered as the owner of the original ImageStack and its copies. The name of the ImageStack has been defined in the first step of the upload and image selection procedure in the client application. All other attributes refer to the meta information that are previously created by the Web server.

- In a next step, three *View* instances are created for each of the ImageStacks, the original one and its two copies. The View concept has been described in detail in Chapter 5.1.3. Each *View* is clearly linked to the above-mentioned *ImageStack* instance via the foreign key attribute *imagestack_id*. The three Views are characterized by special

View types according to their role within *BioIMAX*. The *View* corresponding to the raw original ImageStack is represented through the type *orig*, whereas the two newly generated ImageStacks are described by the types *web* and *thumb*. The values of the *View* attributes *creator_id* and *name* are identical with the values of the attributes *user_id* and *name* of the corresponding *ImageStack* instance. Additionally, this step includes the definition of initial access rights for each of the three *Views*. The access rights are represented by the attributes *all_read* and *all_write* and their values are initially set to 0. Thus, only the owner of the ImageStacks has access rights in the first instance. Later, the owner has the option to modify these rights.

- The *Views* with the types *orig*, *web* and *thumb* represents the three ImageStacks and their single images, which have to be stored in the database by default. The actual raw image files of the respective ImageStacks are stored on a separate dedicated file server. In order to assign these image files to the three *Views* and therefore, to all other data stored in the relational database, an entry in the *Result* relation is created for each single image. These entries contain the path to the image files stored on the file server and are uniquely linked to the respective *View* instance. Each image in the *Result* relation is specified with the previously defined textual tag via the *meta_info* attribute. If the user has selected the tags from the predefined list of antibody tags, the *meta_info* stores the ID of the respective entry in the *Antidict* relation as foreign key. Otherwise the user-defined free textual tags or the file name of the image are stored.

- Finally, the list of selected antibody tags is stored with an *Antidict_selection* instance for potentially reusing them for further ImageStacks. Here, in addition to the unique *user_id* and the name for the selection list that has been defined by the user in the third step of the client upload interface, all antibody tags are combined in one XML string and stored as value of the *antidict_xml* attribute. The single tags are either represented by IDs being foreign keys to respective entries in the *Antidict* relation or by free text, which is similar to the storage of the tags in the *meta_info* of the *Result* relation.

In this way, each single image in *BioIMAX* is associated to one of the three ImageStacks via the respective *View* instances and thereby clearly linked to the abstract *ImageStack* relation with its meta information and to an individual user as its owner. The role of the different multivariate ImageStack versions in *BioIMAX* is pointed out in the following sections of this chapter.

6.1.2 Sharing data via projects

Sharing of data in *BioIMAX*, i.e., either original ImageStacks or result data obtained through analysis or transformation of specific ImageStacks, is supported through the *project* concept. Projects allow *BioIMAX* users to collect and organize an arbitrary subset of datasets regarding a defined biological or analytical question and to share this project specific data with other

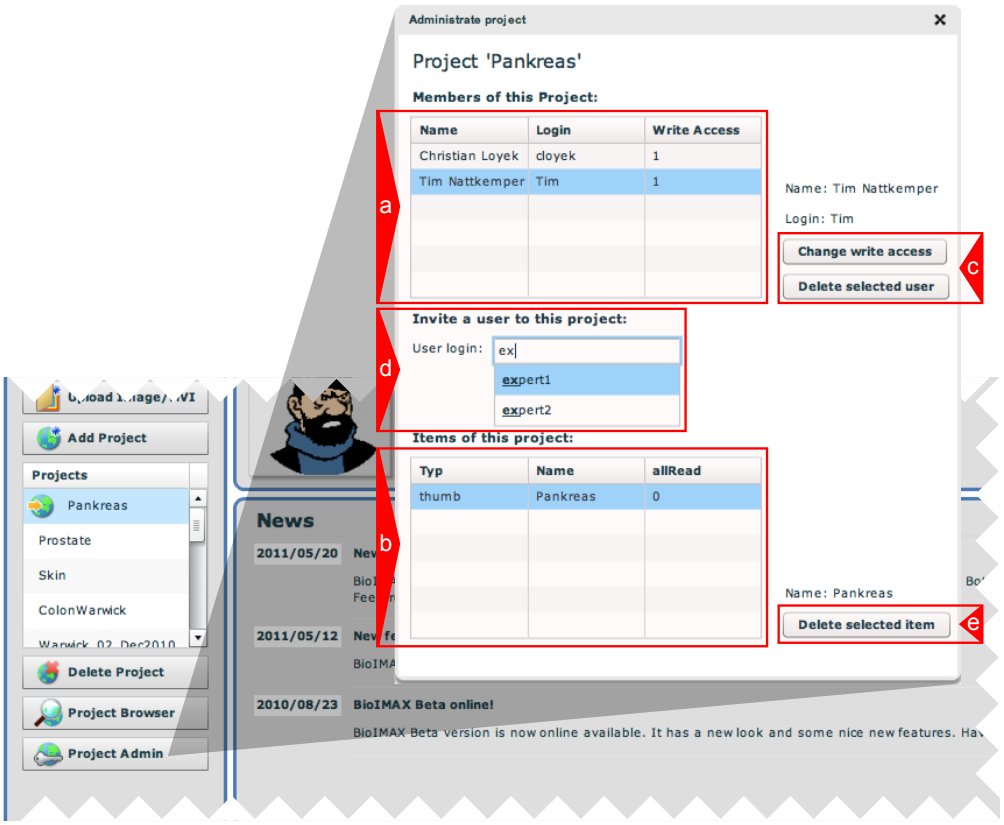


Figure 6.4: Administration of *BioIMAX* projects. The project administration interface generally allows each project member to inspect the current status of the project, i.e., which users are participating in the project (a) and which datasets are assigned to the project (b). Furthermore, project owners especially are authorized to manipulate the current status of the project: changing write access of existing project members and deleting selected members (c), inviting new users (d) and deleting selected datasets from the project (e).

collaborating *BioIMAX* members. The creation, deletion and management of user-defined projects is controlled via the navigation toolbar on the main page (see Figure 6.2(e,f)). Here, users can easily create own projects only by specifying a name and short description about the purpose of the project. Once a project has been created, it is added to the list of the user's personal projects (see Figure 6.2(e)), whereby identifying the user as the owner of the project.

Each project in the list can be managed through a separate administration window illustrated in Figure 6.4. The project administration window allows each project member to inspect the current status of the project, i.e., which *BioIMAX* users are associated to the project (see Figure 6.4(a)) and which data items are assigned to the project (see Figure 6.4(b)). In addition, project owners are authorized to manipulate the current status of the project:

- Owners have the option to grant write access to existing project members, so that they can add new data to the project themselves. Additionally, owners can remove members from their projects (see Figure 6.4(c)).
- Owners can invite further *BioIMAX* users to their projects (see Figure 6.4(d)). Project invitations are displayed on the *BioIMAX* main page of the invited users (see Figure 6.2(f)) and have to be accepted before the invited users have access to the respective projects.
- Finally, owners can delete single data items from their projects (see Figure 6.4(e)).

Adding new datasets to projects is carried out with the *Data Browser* and is described in detail in the next Section 6.2.

Data storage

User-defined projects and information about associated members and data items are stored in the *BioIMAX* database as follows:

- For each newly generated project a unique *Project* instance is created that is linked to a specific *User* instance via the foreign key *creator_id* and is characterized by a *name*, a short description of its purpose through the *meta_info* attribute and the creation *date*.
- The relation *Project_members* links *BioIMAX* users to projects by combining the user and project IDs as foreign keys of the respective *User* and *Project* instances. Additionally, the *Project_members* relation specifies the *write_access* status of project members.
- As mentioned before, ImageStacks and derived result data are represented in *BioIMAX* via a certain *View* instance and its respective *View* types. When a user adds data to a project (see Section 6.2), a new *Project_views* instance is created that links the *view_id* as foreign key of the *View* instance representing the respective data to a specific project. In this way, one *View* with its associated data can be assigned to multiple projects.
- Finally, project invitations are stored in the *Project_invitations* relation, until the user has accepted the invitation.

The project concept allows *BioIMAX* users to easily build up small communities regarding a dedicated biological or analytical problem and to share their data and results with other users. This kind of data sharing is the first and essential step towards collaborative work in *BioIMAX*. Another crucial aspect covered by the project concept is data protection and security. In addition to general privilege mechanisms, i.e., access rights for each uploaded or produced dataset, data access can be limited to certain individuals and groups of researchers using *BioIMAX* projects, whereas project-related data remains hidden for non-members.

6.2 Querying the database (Data Browser)

Since *BioIMAX* is designed to store a large amount of different data and data types, it provides a separate interface called *Data Browser*, in order to manage and search for desired image data and results obtained from analysis and exploration processes (illustrated in Figure 6.5). This interface allows users to search, retrieve, browse and organize their own data, project-related data, and any other data being accessible to the public. The *Data Browser* can be called in two different ways via the navigation toolbar on the main page. First, it can be started initially showing datasets that have been uploaded to the database by the current user (see Figure 6.2(c)). Second, it can be invoked as a *Project Browser*, being the regular *Data Browser*, but displaying exclusively data associated to a selected project (see Figure 6.2(e)).

While starting the *Data Browser*, user- or project-specific data is queried and retrieved automatically from the server and the retrieval results are presented as *data items* in a sortable datagrid. Each row in the datagrid represents one search result or data item, respectively, and depicts relevant data properties, i.e., the item type, name, its particular owner, and information about possible project references. In general, the search results are distinguished into three different types: (1) ImageStacks, (2) single images, and (3) result data. Each data item representing result data (3) is visually enriched by a descriptive graphical icon (as an example see Figure 6.5(a)). ImageStacks and single images are illustrated in the datagrid with thumbnails of their respective images (see Figure 6.5(b)), which have been automatically generated during the upload procedure (see Section 6.1.1 for details). Datagrid items representing ImageStacks allow the user to rapid eye-balling through the ImageStack by moving the mouse cursor over the thumbnail View from left to right and vice versa. Using such interactive thumbnail Views allows users to get a quick overview of the data stored in the database and fosters a fast visual identification of interesting image sets or single images.

Once the *Data Browser* interface has been initialized, the user has several options to submit another query to the database, either by selecting defined search categories (see Figure 6.5(c)) or by submitting a keyword query (see Figure 6.5(d)). Using the keyword search users can search for specific images indexed with descriptive tags, e.g., antibody tags, that has been defined during the upload process. Here, the same autocomplete mechanism as with the upload routine (see Figure 6.3(b)) is used for suggesting predefined tags. This kind of image retrieval follows the text-based image retrieval (TBIR) strategy mentioned in the requirements formulated in Chapter 4.3. Finally, users are able to filter the existing set of search results (see Figure 6.5(e)) in the datagrid according to the aforementioned types, i.e., ImageStack, single images, and result data, which subsequently refreshes the list of displayed data items in the datagrid accordingly.

In addition to searching and retrieving data from the *BioIMAX* database, the *Data Browser* provides options to manipulate and manage data items (see Figure 6.5(f)). Here, a user can remove selected data items from the database, control read and write permissions or assign an arbitrary number of data items to specific projects, provided the user is the owner of the selected datagrid item.

Finally, the *Data Browser* serves as starting point for all data analysis and exploration

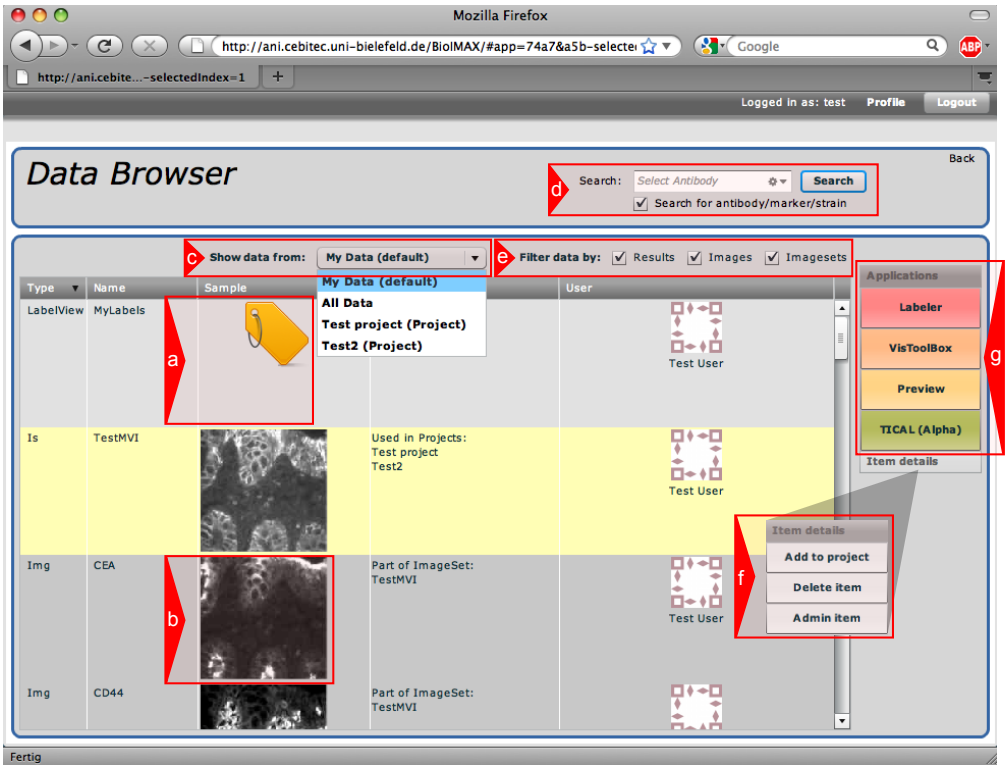


Figure 6.5: The *Data Browser*. This tool serves as graphical user interface that allows *BioIMAX* users to search, retrieve, filter, browse and organize their own data, project-related data, and any other data stored in the database being accessible to the public. Retrieval results are displayed as items in a datagrid depicting interesting information that are visually enriched by descriptive icons and thumbnails (a,b). The *Data Browser* provides different search and filter mechanisms (c,d,e) in order to find desired datasets and includes important administrative functionalities (f). Finally, it is the starting point for any data analysis and exploration tools available via a dynamic application toolbar (g).

tasks in *BioIMAX* . Depending on the semantic type of a selected datagrid item, the *Data Browser* suggests several tools (see Figure 6.5(g)) that can be started using the selected item as initial parameter. The list of possible tools is adapted to the type of a selected datagrid item, since not all tools are suitable to be started with each type of data. Each datagrid item corresponds to a certain *View* instance in the database. Thus, when a user invokes a tool, all data associated to the *View* instance corresponding to the selected data item is requested from the server and passed on to the respective tool as a value object containing the raw image data and any additional datasets and parameters, e.g., the user ID or antibody tags. Triggering a *BioIMAX* analysis or exploration tool is opens a new external browser window that contains the respective tool working independently from the *BioIMAX* main page (see Chapter 5.2.4 for details).

The following sections describe and illustrate the different analysis and exploration tools

available in *BioIMAX* and give a detailed explanation what kind of data is required to initialize the respective tools.

6.3 Image Viewer (Preview)

As already mentioned in Chapter 4.5, one of the most central and essential parts of all *BioIMAX* tools for analysis and exploration of multivariate image data is the *Image Viewer*. The Image Viewer is a graphical user interface component, especially designed

- to display multivariate images in its full extent, in order to initially gain visual insights into the raw multivariate signal space,
- to provide an interactive starting point for analysis and exploration purposes, e.g., to trigger exploration events based on selected regions of interest,
- to visualize and interact with analysis and exploration results directly on the corresponding image content.

According to Chapter 2.1, multivariate images represent a special type of image series containing a set of mutually aligned images of the same size, where each image pixel corresponds to distinct pixels in all other images. This pose non-trivial challenges to the development and realization of an appropriate image viewer as described in Chapter 4.5. In contrast to conventional desktop-based image viewers, the ImageStack has to be considered as a coherent unit, i.e., all navigation and manipulation functions provided by the Image Viewer has to be performed on the entire stack in parallel. This should allow comparative ImageStack navigation, without losing the orientation regarding an observed image region.

The *BioIMAX* Image Viewer provides easy-to-use navigation functionalities that are well known from conventional desktop-based image viewers. It allows users to scroll either sequentially through the ImageStack image by image or to select specific image channels directly using the image tag list that has been defined by the user during the upload of an MVI (see Figure 6.6(a)). The transition from one image to another is performed immediately without any delay, which avoids a blank screen between two consecutive images. Avoiding such blank screen is an extremely valuable feature and is not owned by many conventional image viewing tools. In case of a blank screen, the human visual system does not recognize any significant differences between two images, referred to as *change blindness* (Simons and Rensink, 2005), so a direct comparison of image content of consecutive images is hardly possible. Furthermore, the Image Viewer includes image zooming and panning functionalities (see Figure 6.6(b,c)). Here, the crucial feature is, that zooming and panning are performed not only on the image currently displayed in the viewer, but on each image of the stack in parallel, i.e., at any time all images are displayed on the same scale and at the same position in the viewer. This allows users to sequentially flip through the ImageStack, focussing on interesting image regions and comparing such regions without losing the orientation.

As mentioned before, the Image Viewer with its basic functionalities is a central component of all visual analysis and exploration tools and it is extended by advanced interactivity and

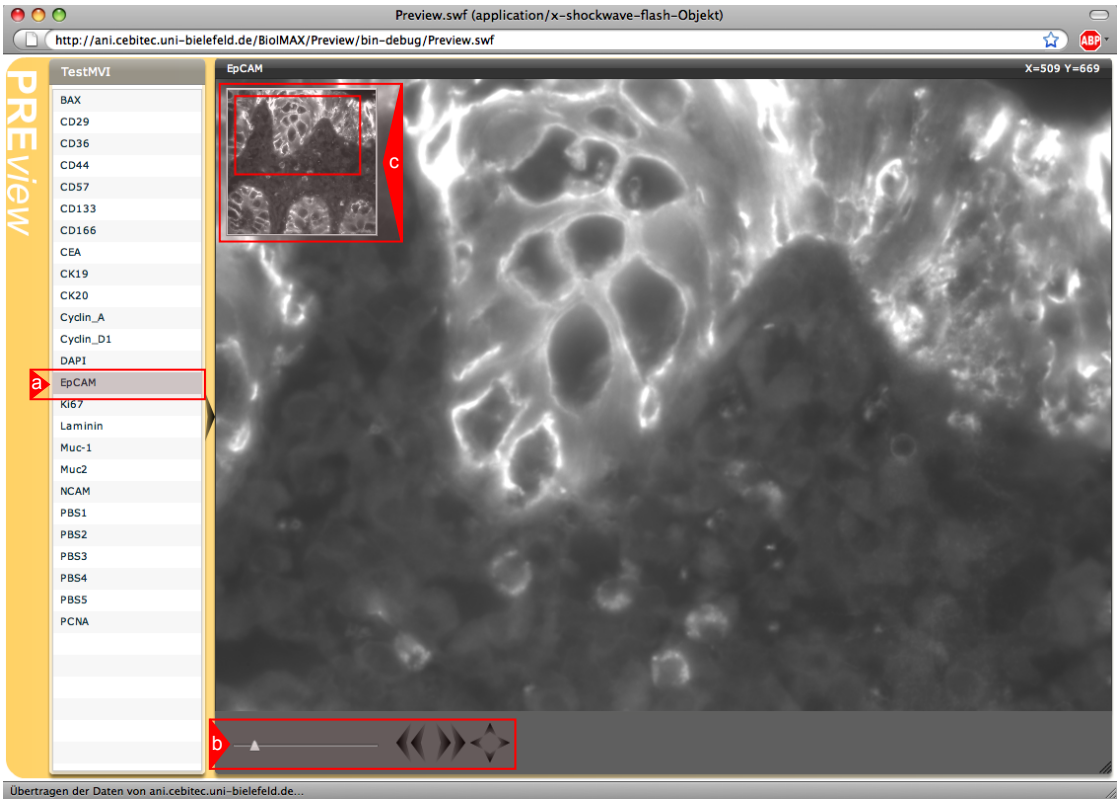


Figure 6.6: The *BioIMAX Preview* tool. The major objective of this tool is to provide initial insight into raw multivariate image signals on a large scale. It includes the *BioIMAX* Image Viewer especially designed to display and navigate through MVI stacks. Images of an MVI stack are arranged in the order the MVI has been uploaded to the system, displayed as a list of textual image tags on the left (a). The selection of specific list items subsequently triggers the display of the respective image in the Image Viewer window on the right. For navigation purposes, the viewer offers standard navigation functionalities that are well-known from conventional desktop-based image viewers: zooming, panning, flipping through the ImageStack (b) and a small window serving as an "outline map" that gives an overview about the current zoom status and orientation (c). The special feature of this viewer is, that by zooming and panning performed on the image currently visible on the screen, the scale and position of all other images in the stack is adapted in parallel. This allows users to navigate through an MVI stack without losing the orientation, e.g., regarding an observed image region.

visualization capabilities tailored to the purpose of the respective tools. In its simplest case, the Image Viewer itself is considered to be an initial exploration tool giving first insights into the raw image signal domain. For this reason, the *BioIMAX* system provides an interface called *Preview* that solely embeds the basic Image Viewer (illustrated in Figure 6.6). The main objective of the *Preview* tool is to offer the basic image navigation functionalities that should allow users to quickly and easily view, browse and compare raw images of a selected ImageStack.

The following section briefly describe how the *Preview* tool is invoked and how the image data is transferred from the server, in order to be displayed in the Image Viewer. The *Preview* tool is invoked via the *BioIMAX Data Browser* (see Figure 6.5(g)) by previously selecting an ImageStack from the datagrid. The corresponding image data and any further information is passed on to a new Web browser window initializing the *Preview* application as a stand-alone application using the delivered data as its parameters (see Section 5.2.4 for details). Data that is send to the *Preview* tool is collected on the Web server as follows:

- The selected ImageStack datagrid item is associated to the three *View* instances in the database with the types *orig*, *web* and *thumb* created during the upload process of the respective ImageStack (see Section 6.1.1 for details).

- For displaying images in the Image Viewer, those images that corresponds to the *web* View are required.

- According to Section 6.1.1, the Web server queries all entries of the *Result* relation that are associated to the *web* View and requests the respective image files from the file server and sends these images in combination with their user-defined image tags to the newly instantiated *Preview* application.

The aforementioned data loading procedure of an ImageStack also applies to all other *BioIMAX* analysis tools described in the next sections. For this reason, it has been only explained once in this section and is omitted in the following sections.

6.4 Semantic Image Annotation (Labeler)

The idea of the *BioIMAX Labeler* tool is to provide the users with an interface that allows them to graphically and semantically annotate image regions in single images/channels of an ImageStack. In general, the *Labeler* tool aims at two goals. First, users shall be enabled to label interesting image regions associated with *low level semantics* and second, chat-like discussions should be linked to image regions to describe morphological features with *high-level semantics*. As with the *Preview* tool the *Labeler* tool is invoked via the *Data Browser* with a selected datagrid item corresponding to an ImageStack. In the following, both types of semantic image annotation with the *Labeler* tool are described and illustrated in detail.

6.4.1 Low-level semantic image annotation

With the *Labeler* users can place sets of graphical objects on arbitrary images of an selected ImageStack, in order to label specific image regions. For this reason, the *Labeler* tool embeds the Image Viewer with its basic functionalities that have been described in Section 6.3 to navigate through an ImageStack (see Figure 6.7). A graphical label is characterized by visual properties, i.e., shape, color, size and position that can be defined and adjusted by the user at any time (see Figure 6.7(a)). In case of the *Labeler*, the Image Viewer has been extended with interactive capabilities that enable users to set selected graphical objects on transparent layers, which are associated to the respective images. Using transparent layers superimposing each single image allows a dynamic modification of labels, i.e., changing their visual properties such as color, position, form or size (see Figure 6.7(b)), showing labels from different channels at the same time and easily removing existent labels. This simplifies an appropriate storage of labels and its characteristics to the database. The layers with its graphical objects behave in the same way as the images in the Image Viewer regarding zooming or panning, so that the size and the offset of the labels are dynamically adapted to the scale and the position of the ImageStack in the viewer.

Graphical annotations can be associated with special types that semantically describe the image content at different scales delineated by the graphical label, e.g., to distinguish labels describing different cell types or even different cell compartments such as Nuclei or Ribosomes. Users have the option either to select a predefined label type from a list of semantic categories (see Figure 6.7(c)) prior to positioning the label on the respective image or to extend the existing 'ontology' of semantic categories and types, in order to define and use own label types (see Figure 6.7(d)). Labels are stored to the *BioIMAX* database with their associated selected semantic types, represented by short *low-level* semantic tags, and are available for further analysis purposes, e.g., used as training samples for automatic image processing or machine learning methods.

Storing image annotations

The *Labeler* tool allows users to store sets of image annotations to the *BioIMAX* database, in order to reuse them for further analysis purposes or to share them with other *BioIMAX* users. For each set of labels a new *View* instance is created identified by a *label* type. This View is assigned to a specific *ImageStack* instance, so that all labels can be associated to an ImageStack via a unique View. For the storage of single labels, the database includes a *Label* relation (see Figure 5.3). Here, each single label is represented by a unique *Label* instance, whose attributes characterize the visual properties of the labels such as *position*, *color*, *size* or *geometry*. The *Label* instances are linked to a respective *View* instance using the *view_id* as foreign key. In order to associate a label to a certain channel of the ImageStack, each *Label* instance includes an *image_id* as foreign key that is linked to the respective entry of the *Result* relation. The *user_id* specifies which user has created the label, since it is possible that several users have edited or manipulated the same set of labels. Through the *type_id* the selected semantic type of the label is defined. The list of possible

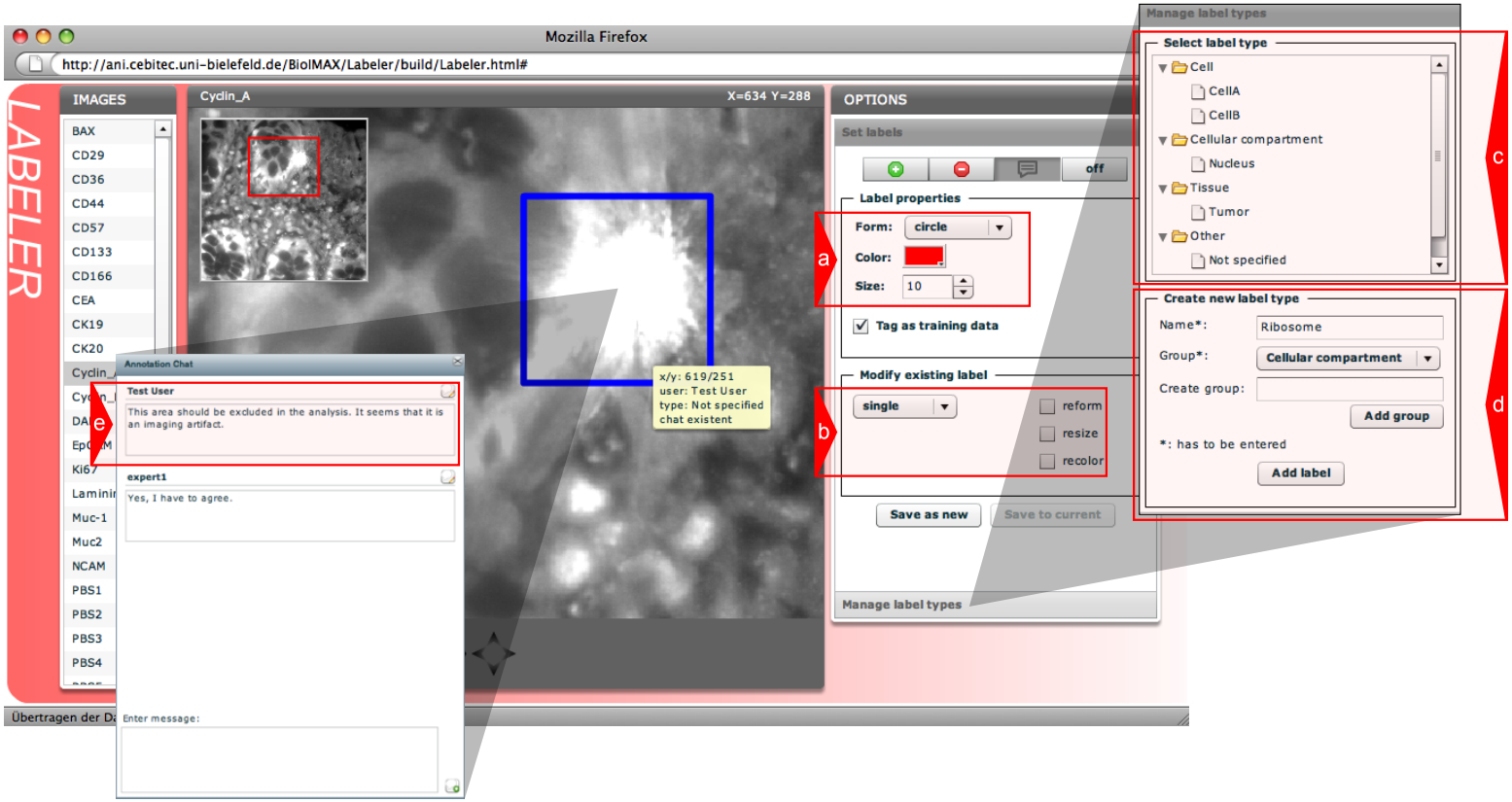


Figure 6.7: The *Labeler* tool. This user interface aims at labeling interesting image regions with semantics at different scales for arbitrary analysis purposes. The *Labeler* embeds the standard Image Viewer described in Chapter 6.3 and extends its basic features with functions allowing users to graphically annotate images or channels of a selected MVI. A graphical label (blue rectangle) can be directly placed on the currently visible image. It is characterized by visual properties, i.e., shape, color, size and position that can be defined and adjusted by the user at any time (see Figure 6.7(a,b)). Each single label can be associated to a low-level semantic category or type that can be selected from a predefined list (c) or by defining new label categories or types (d). Finally, the *Labeler* tool provides a special feature to link chat-like discussions as high-level semantics to image regions. Therefore, it includes a chat window (e) that can be invoked for each single label and the conversation will additionally be stored to the database linked to the respective annotation.

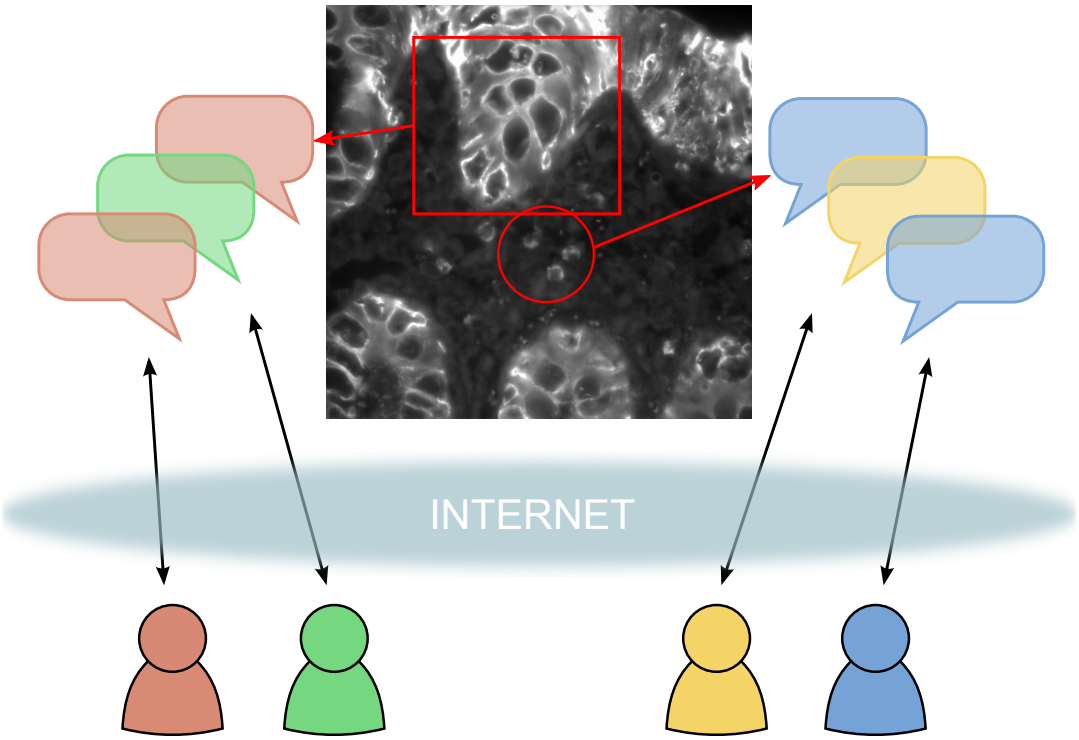


Figure 6.8: Illustration of chat-like discussion about image regions with the *BioIMAX Labeler*. The Figure demonstrates two fictitious conversation scenarios about different image regions from the same image. In this way, scientists from different locations are able to exchange specific knowledge about previously labeled image regions in a fast and uncomplicated way via the Internet, while the stored states of communication content are directly linked to image coordinates.

semantic label categories and types, which can be extended dynamically by each user, is stored in a separate `Label_types` relation.

Once a set of labels is stored to the database, it is accessible again via the *Data Browser* as datagrid item (see Figure 6.5(a)). Existing label results can be easily viewed and modified by selecting the respective datagrid item and invoking the *Labeler* again. Here, the *Labeler* is initialized with the selected label set and with the ImageStack linked to the label set. Additionally, the *Data Browser* allows users to share specific sets of labels, e.g., by adding a label result to a dedicated *BioIMAX* project.

6.4.2 High-level semantic image annotation

In addition to low-level semantic tags that can be associated to graphical image annotations, the *Labeler* tool provides the feature to link chat-like discussions to image regions. In this way, *high-level semantics* such as questions, comments or entire discussions can be linked to morphological image features. For this reason, the *Labeler* tool provides a chat window (see

Figure 6.7(e)) that can be invoked for each single graphical annotation and allows a group of users to communicate about a selected label and the conversation will additionally by stored in the database together with the label. This facilitates Web2.0 style collaborative work on one image via the Internet independently from the users' whereabouts, while the stored states of communication content are directly linked to image coordinates/ROIs. Figure 6.8 illustrates possible communication or discussion scenarios on the same image.

Storing label discussions

Chat-like discussions with the *Labeler* are asynchron, i.e., users have to separately store their contributions to a current label discussion in the database, before other users can reply. Other users involved in the dialog have to invoke the *Labeler* with the respective label and can compose and store new chat messages. The messages are stored in the database using the `Label_chat` relation. Each instance of this relation represents one message from one user regarding a specific label conversation and is associated to the respective `Labels` instance via the *label_id* as foreign key.key.

6.5 Exploratory Data Analysis (VIStoolBox)

The *BioIMAX VIStoolBox* (VISualization toolBox) provides several visualization and exploration methods that allow users to visually and interactively analyze and "browse through" the raw signal domain of multivariate ImageStacks. Methods of the *VIStoolBox* aim at mapping the high-dimensional information inherent in multivariate bioimages onto a physical screen space and integrating the users into the data exploration process, e.g., for co-location studies aiming at detecting and identifying correlations or anti-correlations of image signals in different channels of the multivariate ImageStacks. For this reason, this tool includes techniques from the fields of *Information Visualization*, *Visual Datamining*, *Exploratory Data Analysis* and *Co-location Analysis* that have proven to be powerful, in order to gain fast and initial insights into the structure of multivariate data domains. These methods and their functionalities are not only limited to the analysis of single images, but in particular are capable of simultaneously exploring and comparing signals of multiple images of an ImageStack interactively, following Ben Shneiderman's Information Seeking Mantra: *Overview first, zoom in and filter, and details on demand*.

Once the *VIStoolBox* has been initialized with an ImageStack selected from the *Data Browser*, different exploration methods are available, categorized in a tab-separated user interface (see Figure 6.9(a)). Additionally, the interface includes the standard Image Viewer component, in order to select sets of images from the image list for specific exploration tasks, to limit the focus of analysis to a subset of image pixels, and to display and visualize exploration results directly on the image content. The selection of images for specific exploration methods is obtained by dragging the respective image list items to the application-specific drop targets. In the following the different categories of exploration methods provided by the *VIStoolBox* are described and illustrated.

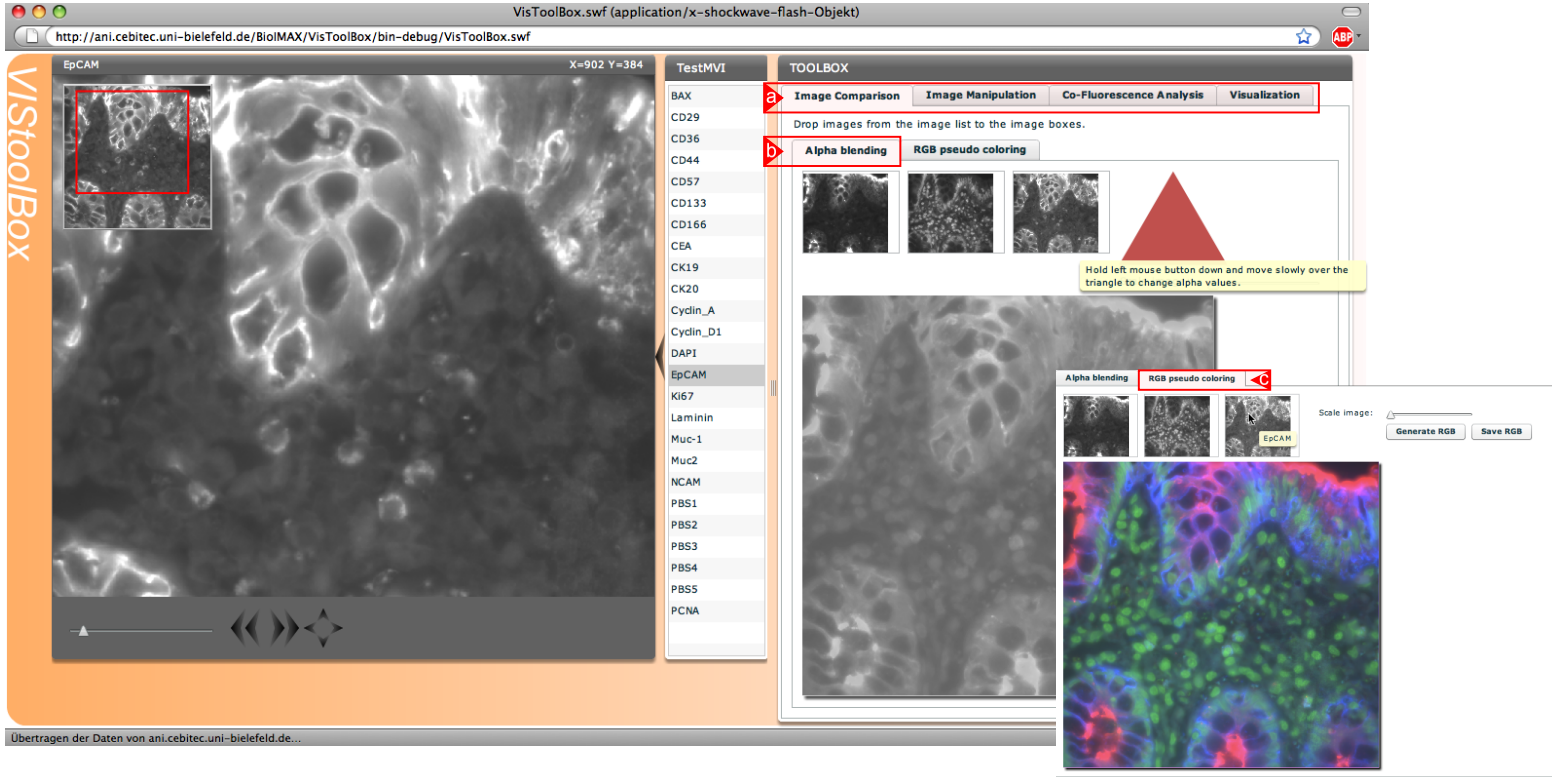


Figure 6.9: The *VIStoolBox*. This tool is a visualization toolbox especially designed to visually and interactively explore the multivariate image signal domain. Via the embedded Image Viewer users select specific image channels for the analysis with exploratory visualization techniques that are categorized in a tab-separated interface (a). The methods of the category *Image Comparison* are illustrated in this screenshot. Here, three different channels can be compared simultaneously on a structural level. The *Alpha Blending* interface (b) presents three selected images superimposed in one image display. In order to compare these images, the user can adjust the opacity values of the three images simultaneously by moving the mouse cursor over the red opacity triangle. The opacity value of the respective images adapts in real-time according to current position of the mouse cursor in the triangle. The *RGB pseudo coloring* method (c) combines three selected channels in one pseudo color fusion image highlighting differences of corresponding pixels encoded by color values.

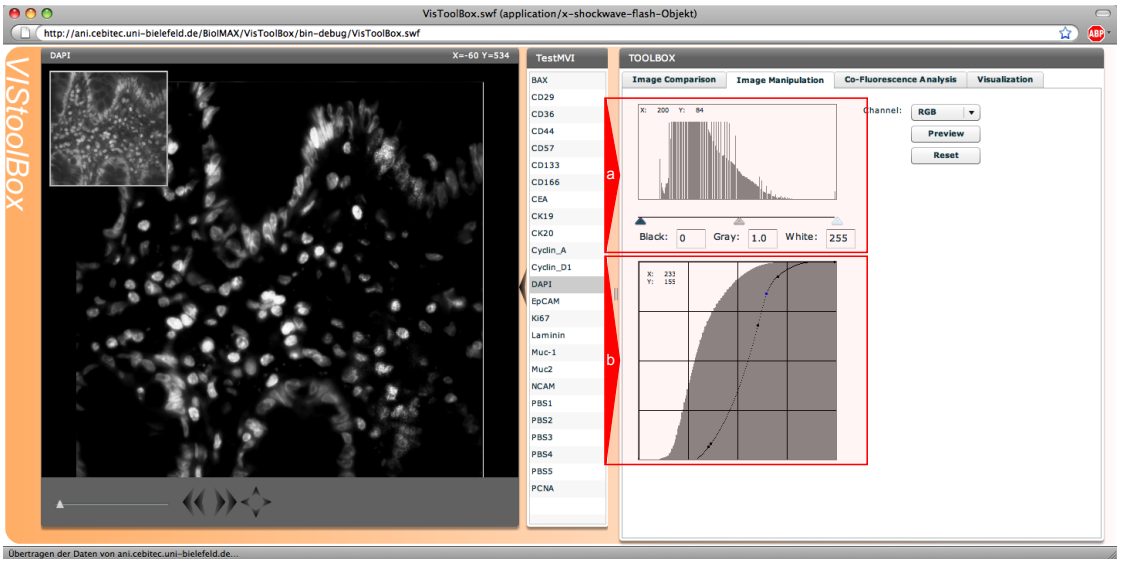


Figure 6.10: Image Manipulation. This part of the *VIStoolBox* allows users to manipulate the gray value distribution of images via two interactive histogram dialogs. The first histogram represents the relative frequency (a) and the second histogram shows the cumulative distribution (b) of gray values of the currently selected image. By manipulating the distributions in the histograms, i.e., setting upper and lower thresholds in (a) and adjusting a gamma curve in (b), users are able to correct or enhance image quality for a better visual interpretation or discrimination of image content. The effect of the manipulation process is adapted and displayed in the Image Viewer on left.

6.5.1 Image comparison

The *VIStoolBox* provides two different methods to compare up to three single images of the ImageStack simultaneously on a structural and morphological level: the *Alpha blending* and the *RGB pseudo coloring* methods.

Alpha Blending

The *Alpha blending* method (see Figure 6.9(b)) aims at comparing three selected images of the ImageStack by superimposing them as layers in one display and manually adjusting the opacity value of the respective layers by moving the mouse cursor over an opacity triangle. Thus, the user can interactively detect structural differences or similarities between the selected images.

RGB pseudo coloring

With the *RGB pseudo coloring* method (see Figure 6.9(c)) users are able to generate a pseudo color fusion image from three selected images. Here, each image is interpreted as one color channel in an RGB image. The color of a pixel or region of the resulting RGB

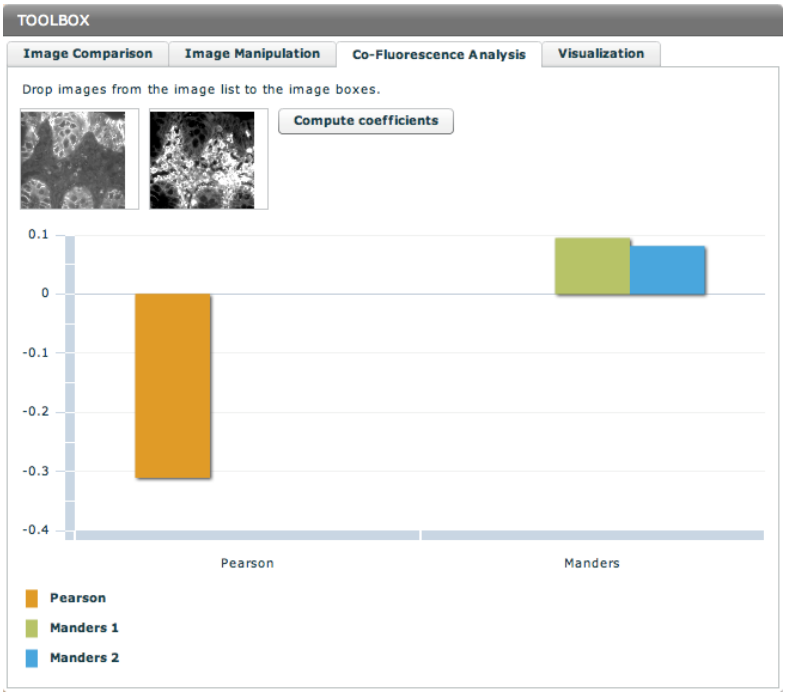


Figure 6.11: Co-Fluorescence analysis. This tool calculates statistical values (Pearson correlation coefficient and Manders' score) describing the co-location of corresponding pixel values of two selected channels. The results are statically displayed in a bar chart.

image indicates the accumulated amount of signals in the three images. As a result, three images are combined in one single display simultaneously, allowing users to rapidly identify structural differences or similarities encoded by color values.

6.5.2 Image Manipulation

The *VIStoolBox* includes two interactive histogram dialogs to manipulate the gray value distribution of selected images, e.g., to filter out irrelevant signals such as signals belonging to the background of the image or outliers. One histogram represents the relative frequency (see Figure 6.10(a)) whereas the other shows the cumulative distribution of image gray values of the currently selected image (see Figure 6.10(b)). The user can manipulate the distributions, e.g., setting thresholds in the first histogram or modifying gray values via a gamma curve. A visualization of the result is adapted in real-time in the Image Viewer.

6.5.3 Co-Fluorescence analysis

In addition to visual exploration and analysis facilities, the *VIStoolBox* provides methods to 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

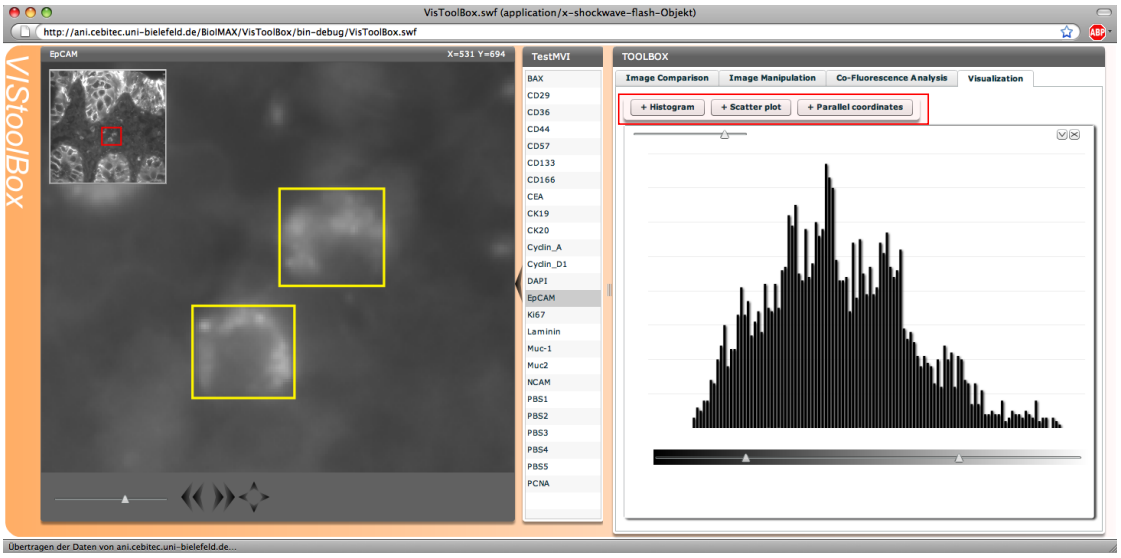


Figure 6.12: Univariate data visualization. The *VIStoolBox* contains three interactive visualization plot interfaces (see red rectangle) that display pixel values of an arbitrary number of outlined image regions. This screenshot illustrates the application of the histogram plot. Here, the relative distribution of univariate pixel values of the two selected regions (yellow rectangles) are visualized as a bar chart. Users can manipulate the histogram display, e.g., scaling the width and height or filtering out specific values by setting upper an lower thresholds. This allows a more detailed exploration of the data space.

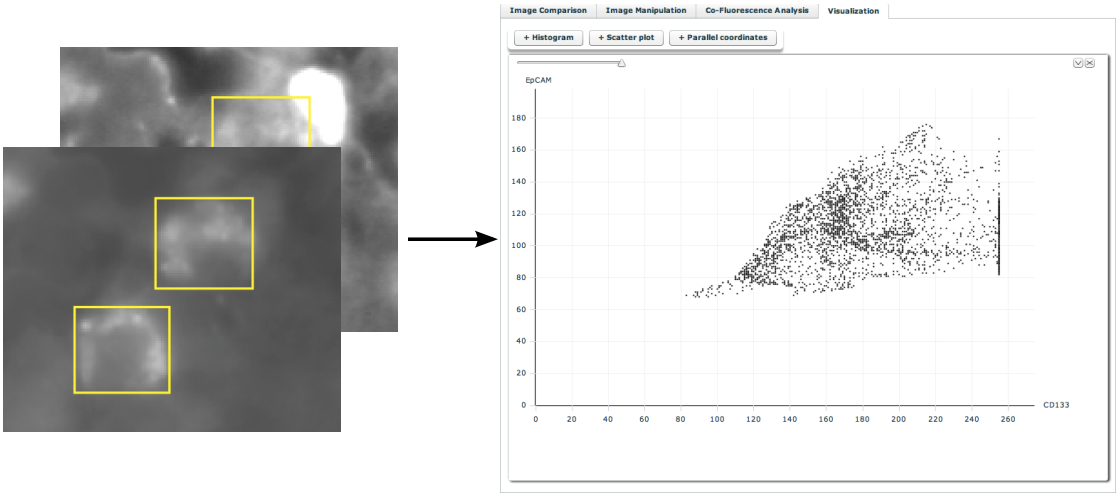
in fluorescence microscopy (Manders et al., 1992). The results are statically displayed as a bar chart (see Figure 6.11)

6.5.4 Visualization

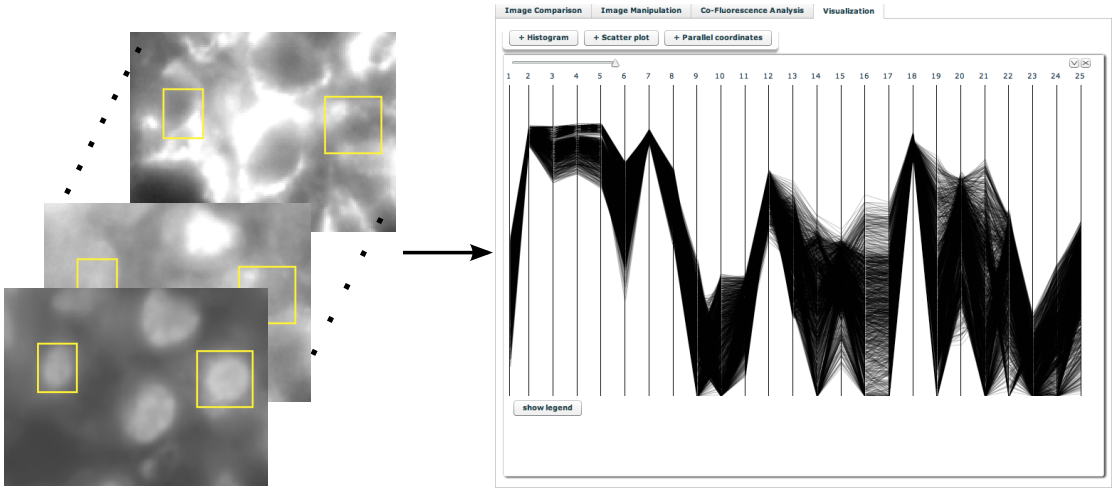
Finally, the *VIStoolBox* provides several interactive data visualization displays that allows a detailed exploration and analysis of signals of selected image regions. As starting point, the users have to outline one or more regions of interest (ROIs) by drawing a rectangle on the currently displayed image in the Image Viewer (see Figure 6.12). In a next step, they have the option to invoke one of the visualization techniques (see Figure 6.12(red rectangle)), which initialize a new plot dialog for uni-, bi-, or multivariate visualization of the image features. Dependent on the chosen plot dialog, users are asked to select images from the image list via Drag&Drop. Finally, all pixel values corresponding to the ROIs are directly extracted from the selected images in the client application and are immediately displayed in the respective plot. In the following the different visualization techniques are presented and illustrated.

Univariate data visualization

For the visual exploration of signals within single images or channels, the visualization component of the *VIStoolBox* provides a histogram display illustrated in Figure 6.12. This



(a)



(b)

Figure 6.13: Bi- and multivariate data visualization. In the above graphic the selection of image regions in two different images and the visualization of their pixel values with the two-dimensional scatter plot is illustrated (a). Each point $\mathrm{f}(s_i, s_j)$ in the plot corresponds to one or more pixels $\mathbf{p} = (x, y)$ with s_i and s_j being the signal values of $\mathbf{p}$ from the images I_i and I_j. An example visualization of selected pixel values from 25 different images with the parallel coordinates plot is shown in the graphic at the bottom (b). In general, each line in the plot corresponds to a k-dimensional signal vector $s(\mathbf{p}) = (s_i, s_j, ..., s_k)$ of pixel $\mathbf{p} = (x, y)$ with $i, j, k \in [1, n]$ (n is the dimensionality of the multivariate ImageStack).

histogram allows the quantitative visualization of *univariate* data and displays the relative distribution of signals from the respective image ROIs. The user has the option to manipulate the display of the distribution by scaling its width and height or filtering out irrelevant values by setting upper and lower thresholds. This allows the user to focus on interesting ranges of the histogram and thus, a more detailed exploration of the data at hand is possible.

Bivariate data visualization

The second visualization component of the *VIStoolBox* is the *scatter plot*. With the scatter plot selected signals from two different images can be displayed in one single graphical display. Here, the pixel values of two selected images that corresponds to the same (x,y) location within the ROI are represented as points in a scatter plot (see Figure 6.13 (a)) spanning a two-dimensional data space usually referred to as bivariate data distribution. As with the histogram display, the width and the height of the scatter plot can be manually adjusted, e.g., for a better separation of points that are located close to each other. Figure 6.14 exemplary illustrate a study of the bivariate image characteristics of two outlined cells.

Multivariate data visualization

The scatter plot display described in the last section allows the display of bivariate datasets. In order to explore ROIs from more than two images, the *VIStoolBox* includes a third visualization component, the *parallel coordinates plot* (Inselberg and Dimsdale, 1990). A parallel coordinates plot is a popular tool that allows the simultaneous visualization of multidimensional data in a compact two-dimensional display, in order to visually reveal the multidimensional characteristics of specific ROIs, e.g., to identify groups or clusters of pixels. For each location (pixel) within the ROI the respective values of k selected images are extracted and combined to a k-dimensional signal vector $s(\mathbf{p}) = (s_i, s_j, ..., s_k)$ with $i, j, k \in [1, n]$ (n is the dimensionality of the multivariate ImageStack). $s(\mathbf{p})$ is interpreted as a point in an k-dimensional feature space. In the parallel coordinates plot variables of the k-dimensional data point are represented by vertical and typically equidistant parallel lines (parallel coordinate axes). A point in the k-dimensional space corresponding to a specific pixel $\mathbf{p} = (x, y)$ of the ROI is represented by a polygonal line with vertices on the parallel axes. The height of the vertices is proportional to the value of the respective variable. The concept of visualizing n-dimensional image signals in a parallel coordinates plot is illustrated in Figure 6.13 (b).

Interactive Brushing and Gating

The proposed visualization components of the *VIStoolBox* facilitate different views and perspectives on the data. However, the different displays presents the data only in a static form and they provide no information about origin of single data items. Therefore, the data displays are enriched with additional interaction capabilities that enable an advanced visual exploration of selected image signals. Each data display provides the option to select a subset of data points or objects in the respective plot, which triggers highlighting the referring pixels in the image, displayed in the Image Viewer. This enables visual linking of components

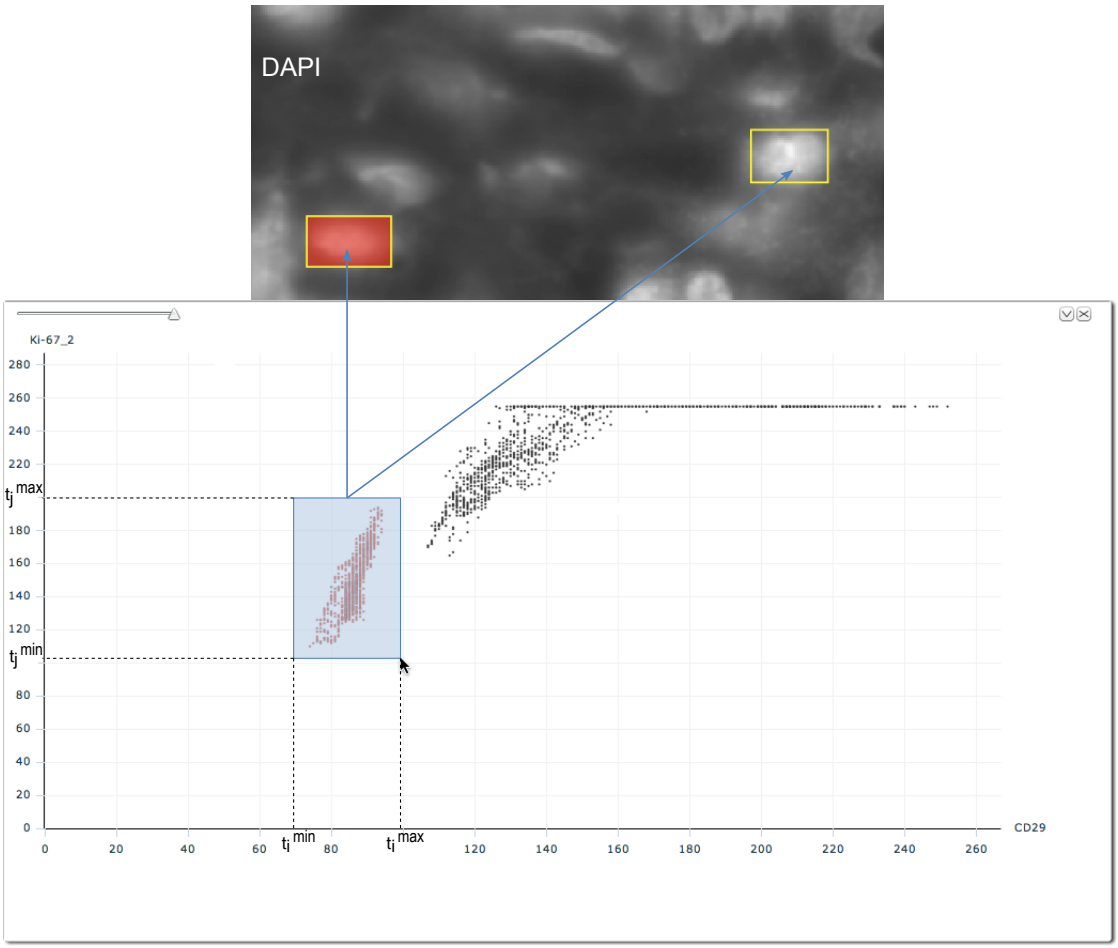


Figure 6.14: Interactive exploration of bivariate data from selected regions of interest (ROIs) in the image. Here, a study of the bivariate co-location characteristic of two cells investigating the protein signals of the channels CD29 and Ki-67_2 in a scatter plot is illustrated. The cells have been identified and outlined via the DAPI channel. The pixel values corresponding to the same (x,y)-location within the ROIs are displayed as points in the scatter plot. Selection of points $f(s_i, s_j)$ in the plot triggers highlighting the referring pixels in the image (displayed as red regions superimposing the original image), with respect to the following criterion: $\Gamma = \left\{ f(s_i, s_j) | \mathrm{t_i}^{min} \leq f'(s_i) \leq \mathrm{t_i}^{max} \wedge \mathrm{t_j}^{min} \leq f'(s_j) \leq \mathrm{t_j}^{max} \right\}$, with Γ being the selection of points in the scatter plot, $\mathrm{t_i}^{min}$ and $\mathrm{t_i}^{max}$ defines the minimum and maximum of the selection range regarding CD29 values and $f'(s_i)$ is the CD29 value of point $f(s_i, s_j)$. The same applies to the Ki-67_2 values, accordingly. This process is often referred to as "Gating" or "Link-and-Brush". This example shows, that the scatter plot is able to reveal co-location characteristics of selected image regions that cannot be quantified by visual comparison of the channels, e.g., clusters of points or groups of outliers. Via "Link-and-Brush" it is possible to close the "semantic" gap between the features visualized in the scatter plot and the spatial location of selected features in the image domain.

in complex data representations to the original signal space. This interactive technique is usually referred to as *Interactive Link-and-Brush* or *Gating* (Becker and Cleveland, 1987) and is exemplary illustrated in Figure 6.14.

All computational tasks in the *VIStoolBox*, e.g., the generation of an RGB pseudo color image, the calculation of statistical values (Pearson coefficient or Manders' score), or the extraction of image regions and their interactive visualization in a plot display, are performed on the client side.

6.6 Datamining tools

Since with the growing number of variables, a pure visual inspection of MVI data using the techniques of the *VIStoolBox*, proposed and described in the last section, only give a limited view on the image data and the complex high-dimensional manifold given by its n-variate features, i.e., image signals. As a consequence, scientists are interested in applying techniques from unsupervised learning such as clustering or other datamining methods such as association rule mining to reduce the complexity of the data and to detect interesting patterns that can be visualized and explored. These techniques are different from the methods and techniques mentioned before in that they previously require the generation of descriptive models based on a subset of image data. Using specific descriptive models of the data allows users to explore and analyze image data with interactive visualization techniques from new perspectives. In the design of *BioIMAX*, several interfaces that allow the integration of such algorithms are currently under development within several student projects.

As these algorithms usually are computationally expensive, they will run on the *BioIMAX* remote compute server, which has been introduced and described in Chapter 5.2.4 (*Application server*) and should be initialized and started via respective graphical Web interfaces from the *BioIMAX* client application. Once a model is computed and stored in the database, its results are immediately available to be visualized and explored via the Internet inside *BioIMAX*. In the following, two classes of datamining techniques and first versions of their Web interfaces are outlined, i.e., *frequent itemset mining* and *image clustering algorithms*. These tools are still prototypes in alpha release state and are either embedded in the *BioIMAX* system for testing purposes or still stand-alone versions not yet connected to *BioIMAX*.

6.6.1 Frequent Itemset Mining (FIST)

Frequent itemset mining is a popular and frequently used datamining method introduced by Agrawal et al. (Agrawal et al., 1993). It aims at identifying interesting patterns such as association rules or correlations and relationships between variables in large databases. Originally, it was applied to investigate customer behavior in supermarkets in terms of the purchased products by calculating association rules for discovering regularities between items (products) in a large set of customer transactions (purchase) and this has coined the term *market basket analysis*. Transactions are considered as vectors containing a set of n binary

attributes $I = \{i_1, i_2, ..., i_n\}$, which represents the presence or absence of the respective item i_k in the market basket (transaction). Following Agrawal et al., association rules are implications of the form $X \Rightarrow Y$, where $X, Y \subseteq I$ and $X \cap Y = \varnothing$. In order to concentrate on interesting rules from the set of all possible rules, different significance constraints can be applied, in order to filter out seldom rules. Prominent constrains are minimum thresholds on *support* and *confidence*. The support of an association rule is defined as the proportion of transactions that contain $X \cup Y$ to the total number of transactions in the database and is considered as the statistical significance of the rule. An itemset is called *frequent* if its support is greater than a minimal support threshold. The confidence value is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that contain X. Thus an association rule is a pattern that indicates when itemset X occurs, then Y occurs with a certain probability.

In addition to the market basket analysis, frequent itemset mining can also be employed in other fields of application, in order to seek for correlations or association rules in high-dimensional data spaces. Thus, it is well suited for studying correlation and co-location of variables, i.e., image signals, in multivariate images, e.g., protein signals obtained with multifluorescence imaging techniques. In case of multivariate images, an n-dimensional signal vector associated to a pixel in the image is considered as a transaction consisting of n variables or items, respectively. All signal vectors of an ImageStack form a large set of transactions and this can be used to mine frequent itemsets and association rules similar to the market basket analysis. For this reason, a new tool, called *FIST* (Frequent ItemSet mining Tool), is currently under development that should allow users to apply an association rule mining algorithm to a selected multivariate ImageStack and to explore and visualize the resulting rules. In the following a brief description of the tool and its prospective functionalities is given.

Since association rule mining is based on binary item attributes, images of an ImageStack have to be binarized by setting an intensity threshold in a first step, distinguishing relevant objects or regions. Therefore, *FIST* includes an interface allowing the user to easily set thresholds manually for each image of the stack, whereas the results of the thresholding process are directly displayed in the Image Viewer. After thresholding all images of an ImageStack, the resulting binary ImageStack can be used to start an association rule mining algorithm that is run on the *BioIMAX* remote compute server. Here, all frequent itemsets satisfying a predefined confidence value are extracted to calculate all possible association rules that are stored in the database. In the literature, there exist several different association rule mining algorithms provided by software libraries for different programming languages, which use different strategies and data structures. The best-known algorithm is the Apriori algorithm (Agrawal et al., 1994), which is used in *FIST*. For the visualization and exploration of generated rules, the *FIST* tool will provide a separate graphical interface. The idea is to display a set of rules as a list of graphical representations (icons) that can be that can be filtered interactively by manually setting the constraint values confidence and support, displaying only those rules satisfying the selected constraints. This allows the users a more detailed exploration of rules with a specific probability. Selection of single rules triggers highlighting those image pixels in the Image Viewer where the signal vectors (transactions)

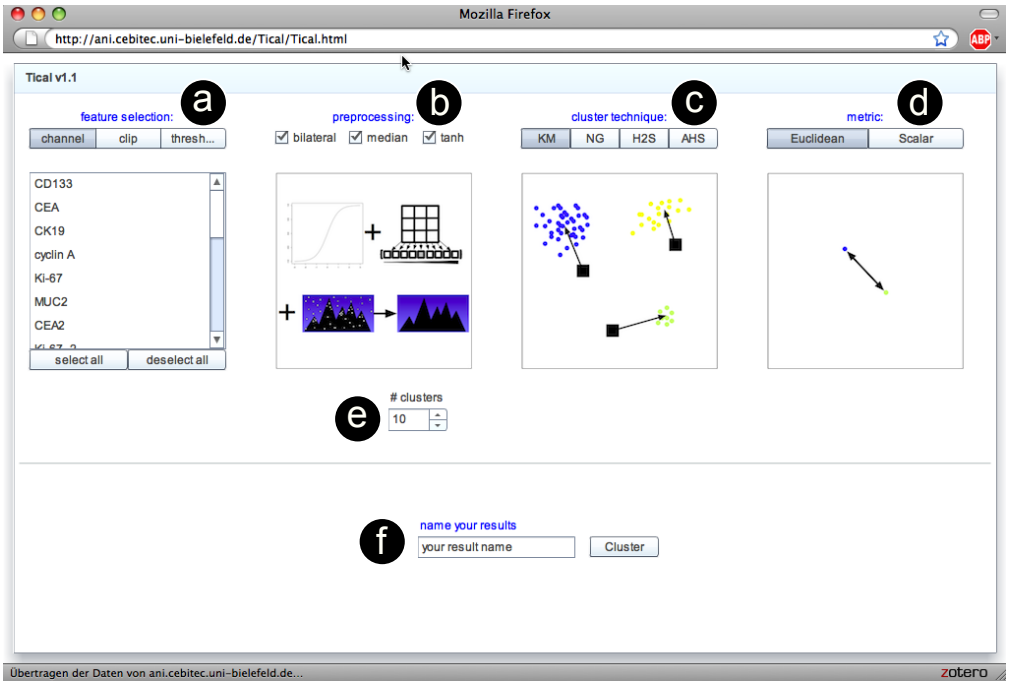


Figure 6.15: The *TICAL* user interface. This interface guides the users through several parameter selections before a clustering job is submitted to and performed on the remote compute server. First, users have to decide which features should be involved in the clustering process obtained by channel selection, image clipping, and thresholding each channel individually (a). The next steps comprise the selection of preprocessing methods (b), clustering techniques (c) and a desired metric (d) that are presented in a visual manner in order to aid the user in selection. Different techniques and parameter combinations are possible. In addition, experienced users have advanced options to adjust the parameters more precisely (e). In contrast, non-experts can always apply clustering using the defaults. Finally, the clustering job can be submitted and the results are stored in the *BioIMAX* database under the specified name (f) as soon as the process has finished.

corresponds to the selected rule.

The *FIST* interface is still a stand-alone prototype tool not yet connected to *BioIMAX* system. Once it has been completed and embedded into the existing *BioIMAX* system, users should be able to detect and localize binary co-location patterns using the association rule mining algorithm, e.g., to find interesting protein combinations in MVIs obtained by multifluorescence imaging.

6.6.2 Image Clustering (TICAL/WHIDE)

With the association rule mining technique binary images are analyzed instead of gray value images. Although this is a reasonable strategy, in order to gain insights into the multivariate signal domain and to identify possible correlations and co-locations between different variables, since it reduces the complexity of the data considerably, it could have some draw-

backs. First, binarization of images requires a high level of expertise and manual interaction for each image of the MVI, which can be quite time consuming. Second, slight changes of the threshold could lead to different transactions, which potentially affect the outcome of the association rule algorithm and its interpretation.

In contrast, other strategies and techniques in the field of datamining are capable of analyzing raw signals, i.e., gray values, of the images. Clustering methods are prominent examples and ideally suited to cope with the original n-variate signal domain with the objective to reveal hidden regularities and structures inherent in high-content images (Herold, 2010). Clustering is an unsupervised learning method and typically applied to determine the intrinsic grouping of data points in a high-dimensional dataset that satisfies a certain similarity criterion. Clustering reduces the complexity of the data by mapping the number of patterns to representative data points (prototypes) to be assigned to graphical parameters such as color allowing for visualization and exploration.

For this purpose, *BioIMAX* interfaces have been developed in several student projects allowing for the integration of a variety of clustering algorithms and visualizations. As with the association rule mining algorithm, the clustering methods are running on the remote compute server as they are computationally expensive and can be selected and started with a tool named *TICAL* (Toolbox for Image Clustering And anaLysis) (Langenkämper et al., 2011), illustrated in Figure 6.15. With *TICAL* several clustering techniques, i.e., *k*-means, neural gas (Martinetz et al., 1993) or self-organizing maps (Ontrup and Ritter, 2006) can be applied to selected images of one or more multivariate ImageStacks. Additionally, *TICAL* provides preprocessing methods to enhance and correct the quality of image signals to achieve better clustering results. Once a clustering result is computed on the application server it is stored into the *BioIMAX* database and can be visualized with another tool called *WHIDE* (Web-based Hyperbolic Image Data Explorer) illustrated in Figure 6.16. *WHIDE* allows the mapping of cluster prototypes to colors which are used to colorize each pixel applying the best matching criterion to the pixel and all the cluster prototypes. This results in a pseudo color image allowing for interactively exploring the image signal domain linked to the spatial domain.

The tools *TICAL* and *WHIDE* are already embedded into the existing *BioIMAX* system and can be accessed via the *BioIMAX Data Browser* (see Figure 6.5(g)), but are still in alpha release status used for testing purposes. With these tools, machine learning strategies and multivariate data interpretation is applicable for non-expert users, since processes are streamlined and integrated in client server data analysis frameworks.

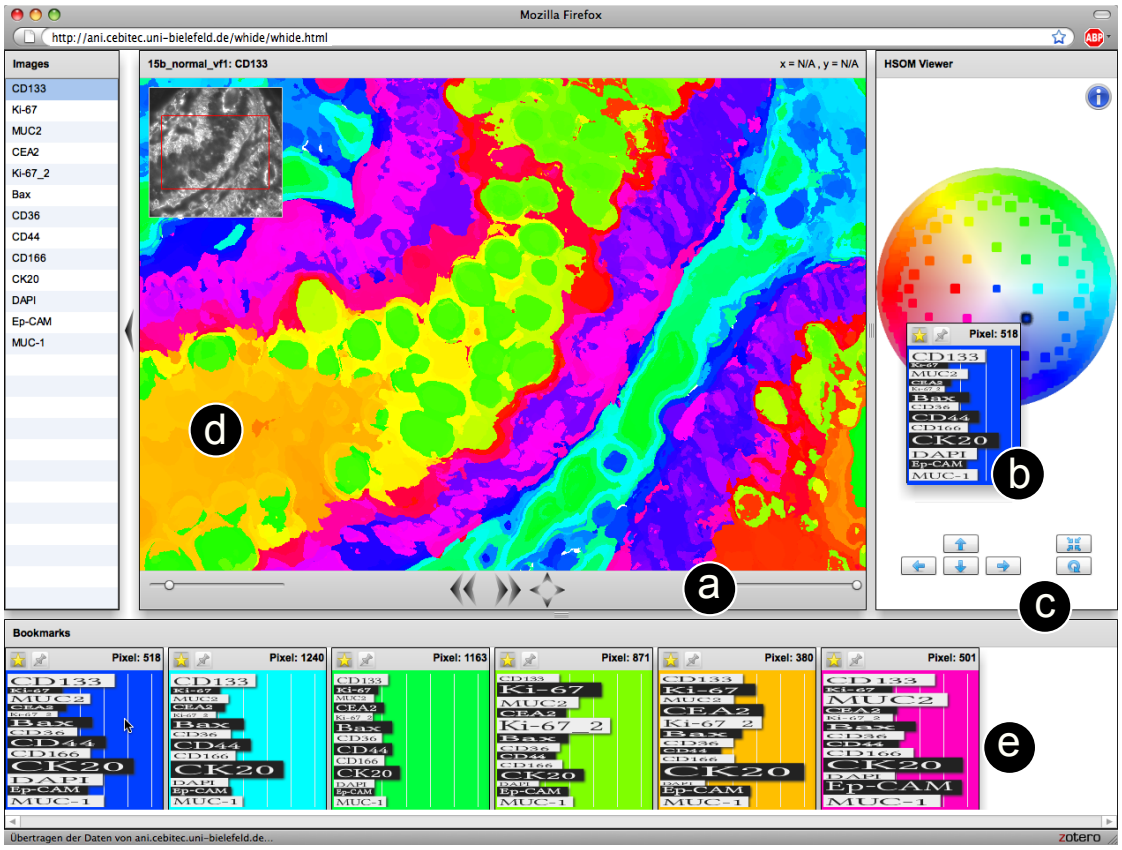


Figure 6.16: The *WHIDE* user interface. A pseudo color visualization obtained with an example clustering algorithm is shown in this screenshot. Prototypes has been mapped to color scale coordinates in a circular color scale varying the hue and the saturation of a color. Similar colors represent similar co-location patterns which point at similar biological functions. *TICAL* supports the process of visual data mining with several functions by modifying the display interactively. Using the slider (a) the opacity of the cluster map is controlled to create a fusion display of the cluster pseudocolor map and one original image of the ImageStack. In the color scale mapping window (b), cluster icons representing cluster prototypes can be shifted using the arrows (c) and the color scale can be rotated to change the coloration of the cluster 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 modifying the color mapping, the colors of the selected cluster icons in the bottom row (e) are adapted accordingly.

CHAPTER 7

Application Examples

In the previous chapters the motivation, the architecture and the implementation of the Web-based *BioIMAX* system for the explorative analysis of multivariate image data was presented and illustrated. Since *BioIMAX* is already an active running system, i.e., it is available via the Web (http://ani.cebitec.uni-bielefeld.de/BioIMAX), several users have uploaded different types of multivariate bioimages, e.g., multifluorescence images, MALDI images, images from spectrometry or multimodal images, and have applied specific functionalities and features of the *BioIMAX* system for their individual purposes. In this chapter, three potential application cases are described emphasizing the usefulness and applicability of certain aspects of the *BioIMAX* system in recent bioimage analysis problems. These application cases focus on bioimages obtained with different imaging modalities showing substantially different image content and signals.

7.1 Studying Bacterial Invasion in High-Content Screening Images

In this study the quantification of cell infection caused by *Listeria monocytogenes* invasion is investigated (Arif et al., 2011). *L. 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 women or immunocompromised individuals (Ramaswamy et al., 2007). The bacteria is an important model organism for infection, intracellular proliferation and host pathogen

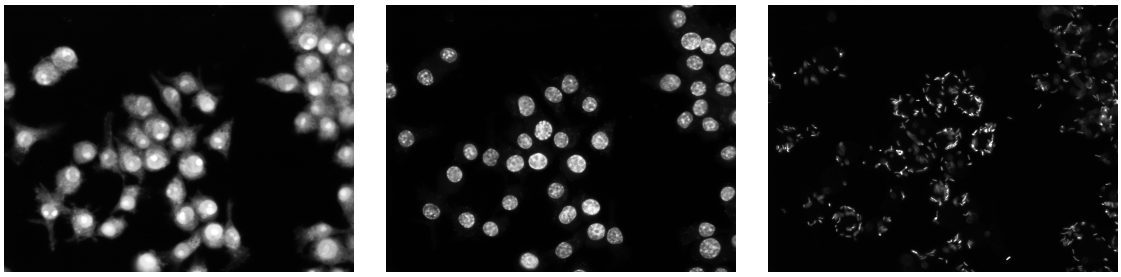


Figure 7.1: Example high-content fluorescence image showing infected cells: (left) Cell channel: cytoplasm, (center) Nuclei channel, (right) Listeria channel: GFP stained Listeria

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 (Ireton, 2007). In order to study the grade of host cell invasion with *L. monocytogenes*, a high-content screen (HCS) has been set up using automated microscopy and *L. monocytogenes* expressing the green fluorescent protein (GFP). Figure 7.1 shows an example high-content fluorescence image, obtained with a $\mathrm{Scan}^{\mathrm{R}}$ screening station (Olympus).

The major objective of this study is to develop an algorithm, which automatically (i) segments cells in each image and (ii) classify each cell regarding its degree of *L. monocytogenes* infection using three-channel high-content screening images. Several experts from different disciplines and institutes at different locations are involved in this study (see authors' affiliation in (Arif et al., 2011)) forming an interdisciplinary research project. In the following, different aspects of the *BioIMAX* system are highlighted and illustrated, which can support the collaborative analysis of *L. monocytogenes* infection, in order to develop new analysis strategies.

Sharing image data

Once all experts involved in this study have generated a *BioIMAX* user account, they can easily upload image data to the *BioIMAX* database and collect these images in a specific project. By inviting the other users of this study to this project they gain access to all associated data from this project from any location in the world provided with an Internet connection

Morphological overview of the data

In a first step, each member of the research project can easily get a qualitative overview of the severity of *L. monocytogenes* infection and of the location of bacteria, i.e., intra- or extracellular using the *Image comparison* methods provided by the *VIStoolBox*. Here, they can compare the three channels of the HCS simultaneously as illustrated in Figure 7.2.

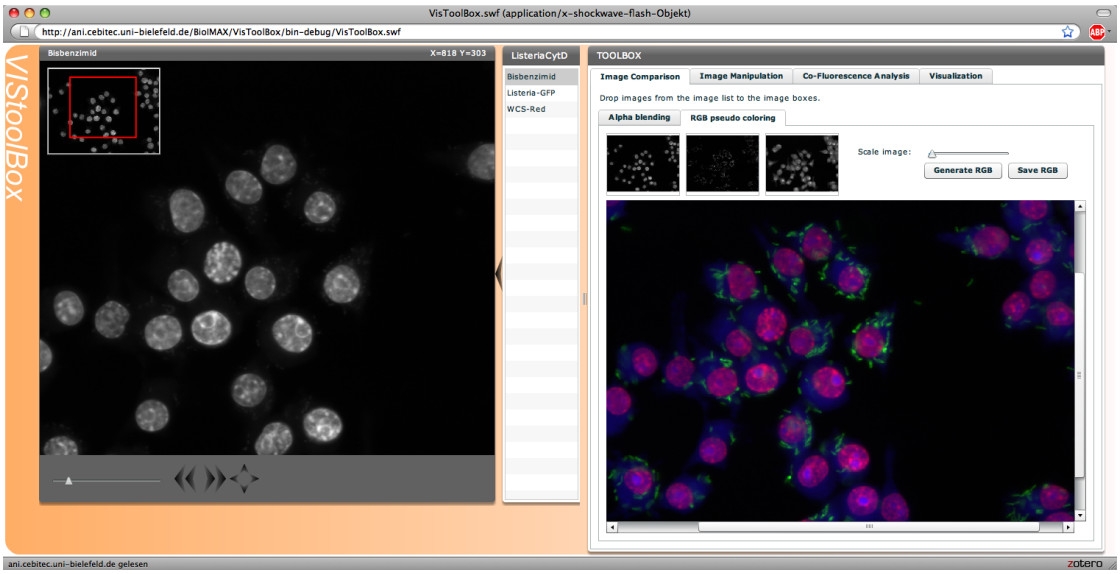


Figure 7.2: Comparing all three channels of an HCS image simultaneously with the *RGB pseudo coloring method*. The *VIStoolBox* provides the possibility to get a first qualitative overview about the location and severity of *Listeria monocytogenes* infection. (Red: nucleus; Green: *L. monocytogenes*; Blue: cytoplasm)

Analysis on single cell level

In order to investigate the cell invasion of *L. monocytogenes* in more detail, users can focus the study at a single cell level, e.g., to examine the severity of cell invasion in the nucleus shown in Figure 7.3.

Evaluation of analysis strategies

With the *Labeler* tool users can label and discuss interesting image regions, which can be important in quantification and evaluation tasks. While developing new analysis strategies and algorithms for high-content image data, researchers have to discuss several aspects about the *original data*, e.g. the trustworthiness of image signals, and about *analysis methods*, e.g. the quality of intermediate results such as registration or segmentation results. Figure 7.4 illustrates a chat-like discussion with the *Labeler* tool about a result image obtained by an automatic segmentation method.

In addition to segmenting images, scientists are aiming at developing an algorithm, which automatically classifies cells regarding their degree of *L. monocytogenes* infection. For this reason, it is necessary to compare its performance to a gold standard derived by a number of independent human experts. To this end, 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 and can start labeling cells, e.g. using circles with different colors, each color representing a specific infection grade (see

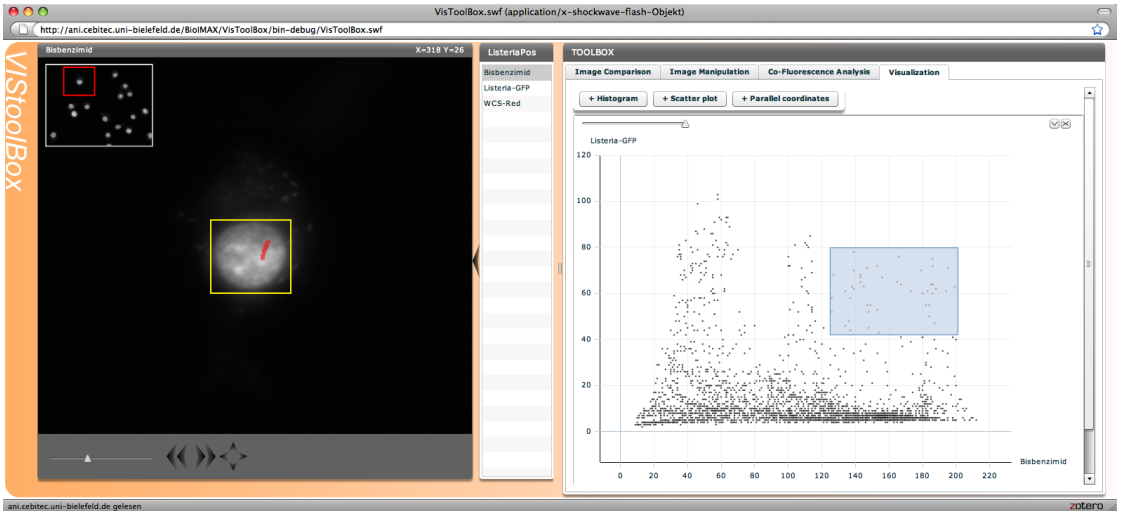


Figure 7.3: Detailed investigation of cell invasion at a single cell level. Here, the pixel values of a selected ROI of the nucleus and the *Listeria monocytogenes* channels are visualized in a scatter plot. The selection of points in the plot showing high values (high image signals) in both channels (see blue rectangle) reveals that *L. monocytogenes* are located within the nucleus (see red region in the selected cell).

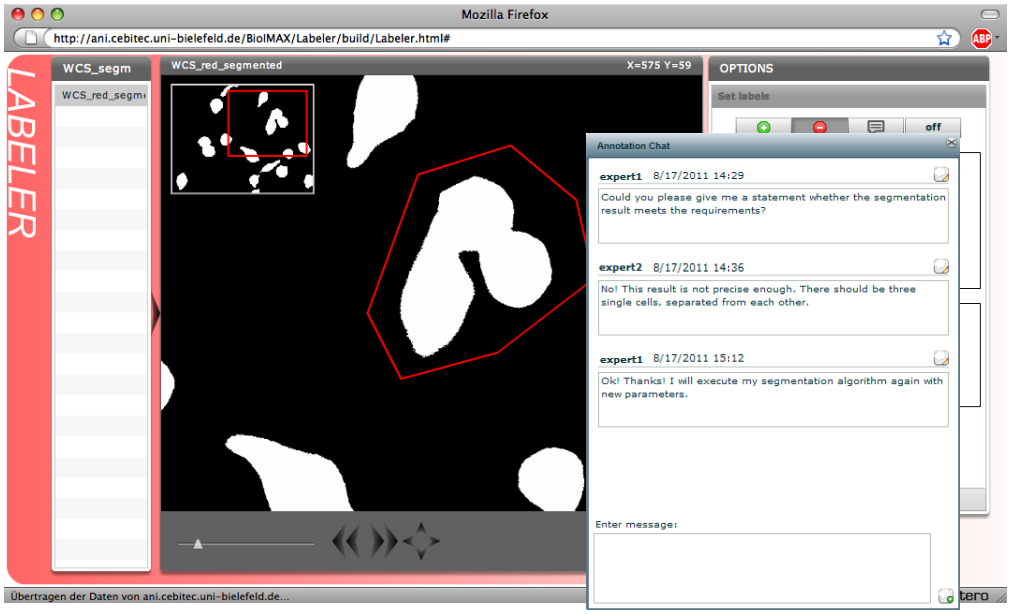


Figure 7.4: Communication and discussion of a segmentation result. This screenshot illustrates a discussion about a segmentation result via the *Labeler* tool. One expert outlines a region in the result image obtained with a segmentation algorithm and formulates a question about the quality of the segmentation result. Other experts involved in the same project can directly respond to the question via the Web without transferring the result data from one expert to another, which is usually a time-consuming task.

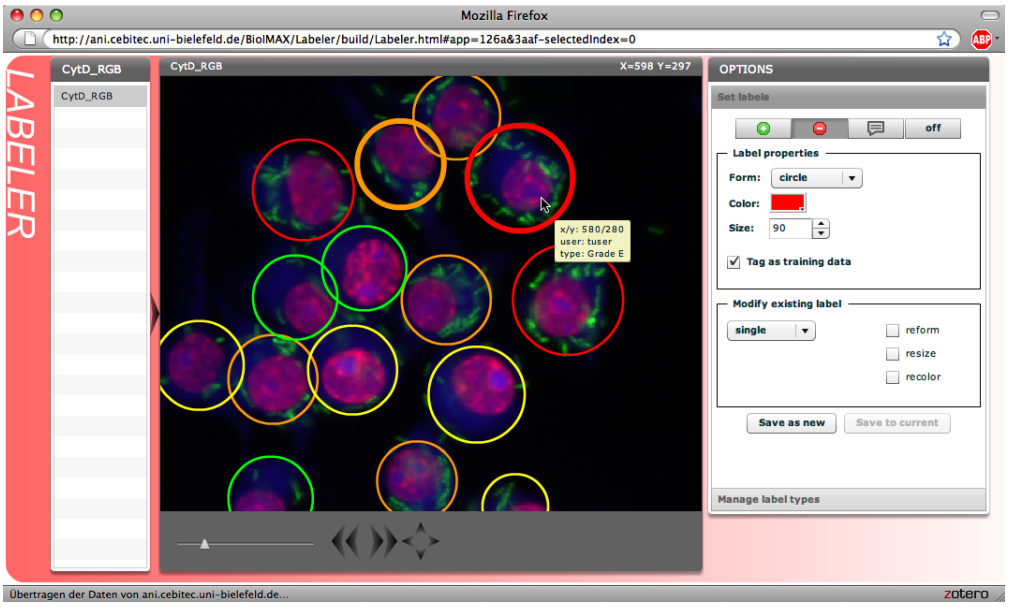


Figure 7.5: Cell image annotation. This screenshot shows annotations of cells in a color fusion image with the *Labeler* tool. Each cell is assigned to a specific semantic category (here: the grade of *L. monocytogenes* infection by an expert. Different grades of infection are represented by different colored circles. In this way, it is easily possible to establish a gold standard with several experts, in order to compare the performance of automatic cell classification algorithms to the gold standard.

Figure 7.5). Using the *Labeler*, the process of establishing a gold standard with several experts is accelerated and simplified, e.g. there is no need to transfer multiple copies of images from one expert to the others. The users can easily login to the *BioIMAX* system and can immediately start labeling from any location. All label results will be stored centrally in the database, are available for any further analysis tasks and can be accessed by the collaborating researchers at any time.

The results of this application example have been published in (Loyek et al., 2011b).

7.2 Collaborative Analysis of Ion Mobility Spectrometry Data

Although the first application example has illustrated the applicability of *BioIMAX* to typical multivariate image data, the general BioIMAX design allows an application to various kinds of two-dimensional scientific images. This way, one can quickly explore, compare, and share images and discuss certain image-related aspects with collaborating experts at different locations, even if the image data is originally not acquired for multivariate analyses. In the following, we demonstrate the usefulness of *BioIMAX* for this kind of image data using data from Ion Mobility Spectrometry (IMS) as an example.

IMS is a method to characterize chemical compounds on the basis of gas-phase ions in an electrical field (Baumbach and Eiceman, 1999). Together with the usage of a multi-capillary column for pre-separation, the resulting data is typically visualized as a heat-map image. Recent applications of the IMS technique show great potential to screen complex mixtures like samples from the headspace of cell cultures and even more complex mixtures like human breath (Baumbach, 2009). After data acquisition and several pre-processing and alignment operations, chemical compounds can be detected, quantified, and compared. Since IMS is still a relatively young and emerging technology, it opens up new vistas and analysis approaches for the field of spectrometry. In addition to the application of existing and established analysis methods, IMS research is an ongoing knowledge discovery process with the objective to gain new insights into the data domain. For this reason, scientists in IMS research projects in first instance need advanced analysis methods, which allow them to explore and visualize the data at hand, in order to generate new hypotheses or to develop improved and specialized analysis strategies. Various facets in IMS research leads to challenges at different levels in data analysis. Therefore, scientists from different disciplines are usually involved in the entire knowledge discovery process, focussing on specific analysis aspects depending on their expertise. This implies, that scientific collaboration plays an important role in IMS research, in order to share and discuss data and results with experts from other scientific fields. Two potential scenarios using *BioIMAX* capabilities with respect to IMS data analysis are described and illustrated in the following.

Collaborative work on IMS data

During the IMS data analysis process, some of the regions in the IMS image cannot clearly be assigned to known compounds, due to unexpected influences that prevent automatic evaluation of the sample. These image regions need to be examined and discussed with experts from different disciplines like medicine, biology, chemistry, or computer science, e.g., to quantify these regions or to avoid misinterpretations or to exclude irrelevant regions in future analysis. For this reason, researchers can upload single images to the *BioIMAX* database, in order to share them with collaborating experts using *BioIMAX* projects. With the *Labeler* tool involved users with different expert knowledge can focus a discussion to conspicuous image regions using the chat facility. In Figure 7.6 a hypothetical communication about an image region with the *Labeler* is illustrated. This example discussion highlights the usefulness of the *Labeler* tool regarding a typical IMS workflow scenario, where researcher from different disciplines are working with the same IMS data.

Comparative IMS analysis

In addition to identification and quantification of compounds in single IMS images, a frequent challenge is the comparative analysis of sets of IMS samples, in order to get first insights into structural differences or similarities between different samples. A typical scenario in the analysis of IMS images is the comparison of the actual sample with two previously taken reference measurements. Before a sample is taken, an instrumental blank and a medium

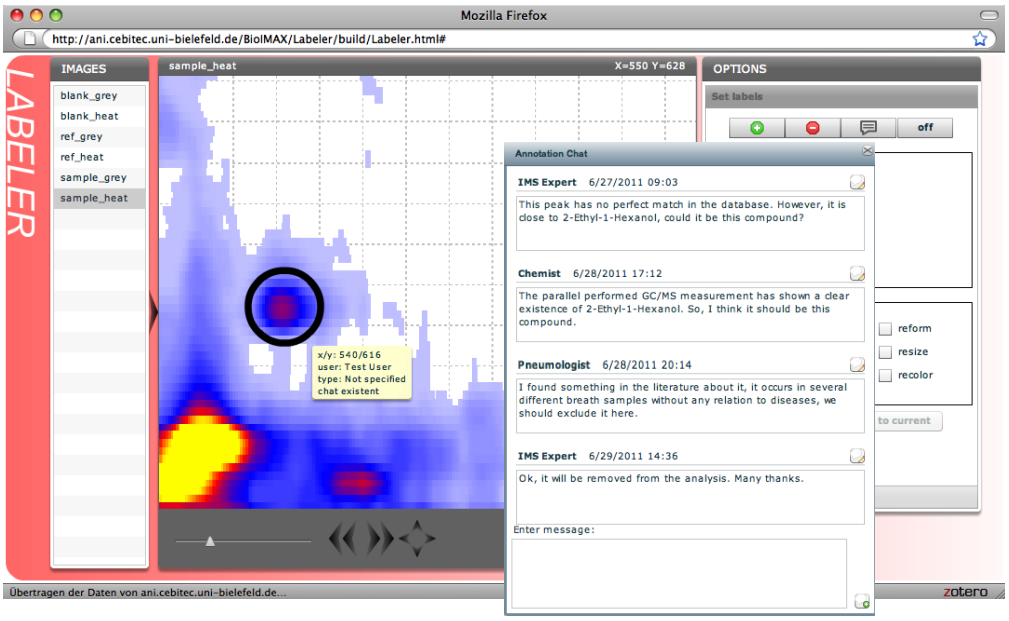


Figure 7.6: Potential discussion of IMS image data. Existing IMS analysis tools allow the application of preprocessings operations like noise reduction, normalisation and alignment (Bödeker et al., 2008; Bunkowski, 2010). They also contain methods for automatic peak detection, quantification and functionality to export the data as images. In case of the analysis of exhaled air, experts from the fields of pneumology, chemistry and computer science need to communicate in order to discuss new or unexpected data. One example subject which is frequently discussed is the origination of so far unknown peaks. With the chat function of the *Labeler*, the treating pneumologist can give information about recently changed medications, which can cause a peak and the computer scientist can check if the peak is caused by computational artifacts. Additionally the chemist can search existing databases if substances with matching characteristics exist and try to identify the peak.

reference is recorded, in order to determine if a compound is caused by the device, the surrounding medium or by the actual sample. For this reason, one could use the *Image comparison* methods of the *VIStoolBox*. In Figure 7.7 the simultaneous comparison of the actual sample with both reference samples using the *RGB pseudo coloring* method is illustrated.

In addition to the comparison of IMS image with reference data, the analysis of IMS data often requires the comparative analysis of sets of registered samples from different experiments or patients, e.g., to reveal differences or similarities between pathological and healthy human breath samples. Here, the analysis is focussed on specific compounds, i.e., regions in the IMS images. The *VIStoolBox* is thereby well suited for a fast and uncomplicated comparative analysis of specific image regions given by its interactive visualization facilities (see Figure 7.8).

The results of this application example have been published in (Loyek et al., 2011a).

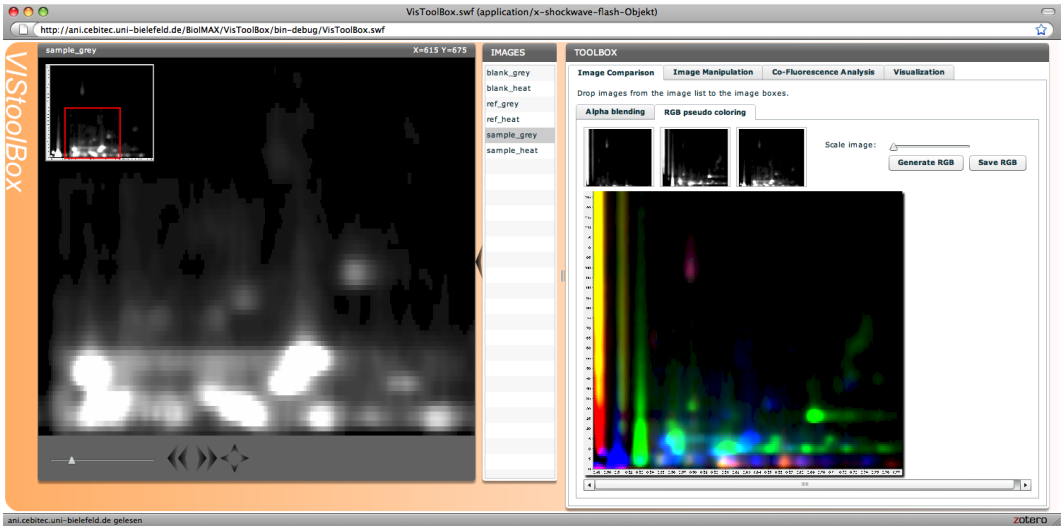


Figure 7.7: Comparison with IMS reference data. In this example the RGB pseudo coloring method is applied to compare the sample with reference images, which are usually produced for each newly generated IMS dataset. With this pseudo color image, experts can get a fast qualitative overview about the structural difference of the sample and the reference data, e.g., if a compound is caused by the device, the surrounding medium or by the actual sample itself.

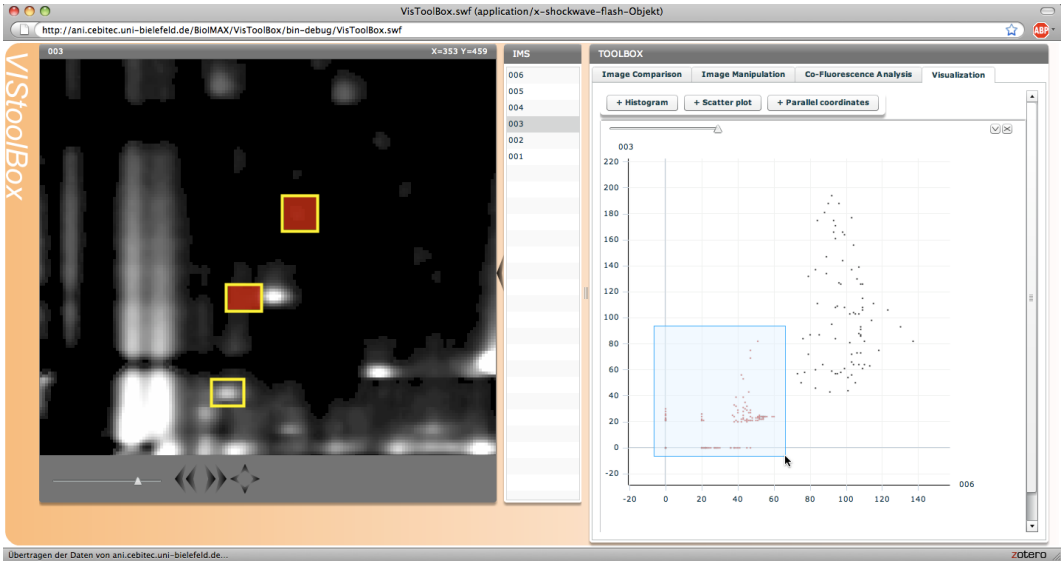


Figure 7.8: Comparative analysis of compounds. Here, the analysis is focussed on specific compounds, i.e., image regions using the bivariate scatter plot visualization tool. In this way, the same regions from different (registered) IMS images can be compared, in order to identify differences between the occurrence of compounds in different samples, e.g., between pathological and healthy human breath samples.

7.3 Collaborative evaluation of epilepsy-causing brain lesions using MRI

In addition to the previously described applications cases in Sections 7.1 and 7.2, systems like *BioIMAX* are promising and important assistance tools for other fields of application. One potential field is clinical diagnostic based on digital image data, which has been acquired with medical imaging techniques such as magnetic resonance imaging (MRI), computer tomography (CT), or ultrasound imaging. In the following, a potential application of *BioIMAX* as part of the clinical diagnostic routines and research in the field of epilepsy is described and illustrated. Here, one prominent pathological phenomena is Focal Cortical Dysplasia (FCD) (Taylor et al., 1971), which is usually diagnosed and identified using 3D-MRI as neuroimaging modality.

FCD, a disorganization of cortical development, is an important cause of medically intractable partial epilepsy. In medically resistant epilepsy patients, only surgical removal of the dysplastic lesions leads to significant reduction or cessation of seizures (Woermann et al., 2004). High-resolution MRI has proven to be the most successful technique to detect FCD lesions and its improvements in the past years has allowed more patients to undergo resective surgery. However, visual analysis and identification of FCD is a challenging task and strongly depends on the experience and expertise of the observer.

In a large number of cases, FCD lesions can not clearly be distinguished from healthy cortex and often remain unrecognized in standard radiological analysis (Tassi et al., 2002). Even for experienced radiologists it is difficult to descry the subtle characteristics of FCD lesions in MR images. This motivates the question to what extend *BioIMAX* could be helpful to lower characteristic hurdles in neuroradiology and to aid the process of detecting FCD lesion in MRI. Typical aspects of this diagnostic problem are pointed out in the following.

Sharing FCD cases for diagnostic reasons

In general, the *BioIMAX* system could serve as a medium to rapidly and easily exchange MRI data between radiologists at different places. In this way, radiologists would be able to discuss MRI data regarding diagnostic aspects of complicated FCD cases and to communicate possible treatment outcomes, e.g., sharing post-surgical images and resulting knowledge. Via the central project concept in *BioIMAX* , expert radiologists could build up small communities and could exchange anonymized FCD data that needs to be jointly investigated.

Furthermore, neuroradiologists without expert knowledge in FCD diagnosis would be able to discuss cases of medically resistant epilepsy patients with experts in the field of FCD diagnostic. They could easily invite experts to their projects, upload relevant image data (e.g., selected slices of the 3D-MRI), label possible suspicious regions with the *Labeler* tool, and ask the experts for their opinion using a chat conversation. The experts on the other hand could examine the images of the non-experts and could give a hint about the suspicion of FCD being the cause of the epileptic seizures.

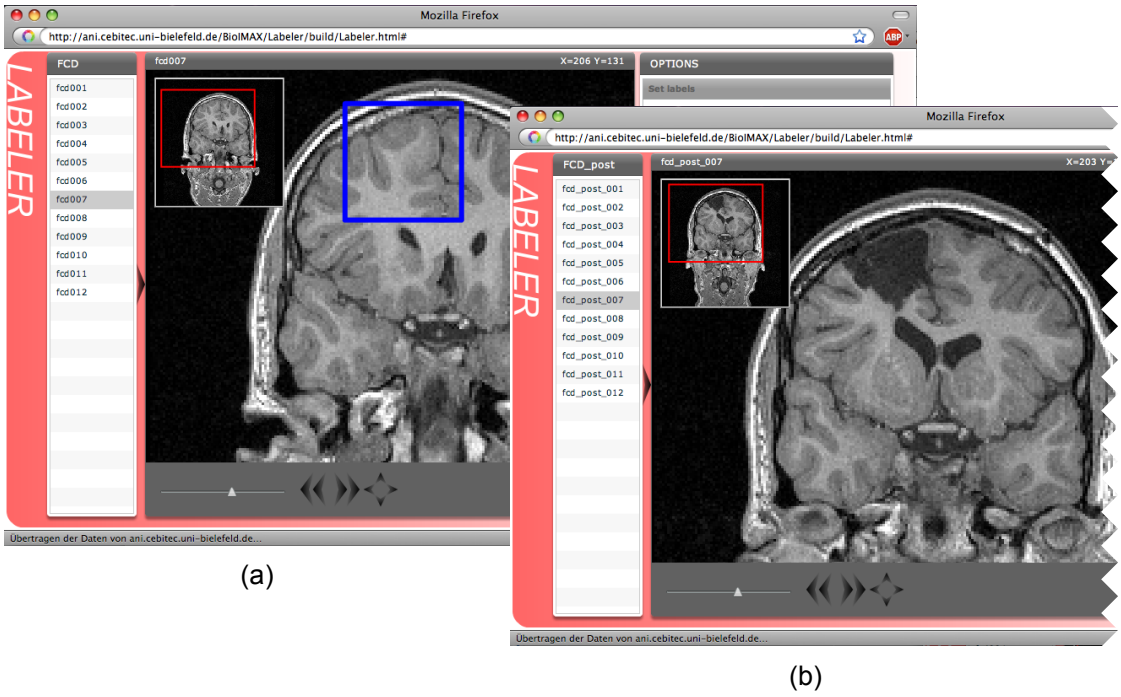


Figure 7.9: Sharing and discussing FCD cases. This figure presents two screenshots of the *Labeler* displaying one slice of a 3D-MRI of the human brain. Radiologists could upload sets of selected MRI slices to the *BioIMAX* platform, share them with collaborating experts via *BioIMAX* projects, and communicate specific diagnostic issues regarding clinical image data. The *Labeler* in (a) shows an MRI of an epilepsy patient with an FCD lesion (see blue rectangular label). In (b), slices of an MRI of the same patient are shown after resective surgery. A large part of the brain tissue producing epileptic seizures has been removed. With the *Labeler* experts could communicate about pre- and post-surgical MRI data regarding FCD diagnostics and surgical outcomes.

Development and evaluation of automatic lesion detection

In addition to the support for clinical practice, *BioIMAX* could play an important role for several research aspects regarding FCD. One point of research focusses on the development and evaluation of automatic image analysis techniques for computerized detection of FCD lesions, in order to augment the radiologic findings based on MRIs (Huppertz et al., 2005; Colliot et al., 2006). In (Loyek et al., 2008) the potential of machine learning based classification of textural image features is investigated. This study aims at detecting local structures in MR images, which are unidentifiable by the experts' naked eye. Therefore, first and second order image texture features of the cortex have been extracted and used for training a support vector machine (SVM) (Cortes and Vapnik, 1995) that should allow a discrimination of healthy and lesional cortical brain tissue. A crucial aspect for the training of the SVM and its evaluation is the availability of a data basis of a sufficient number of healthy and histologically

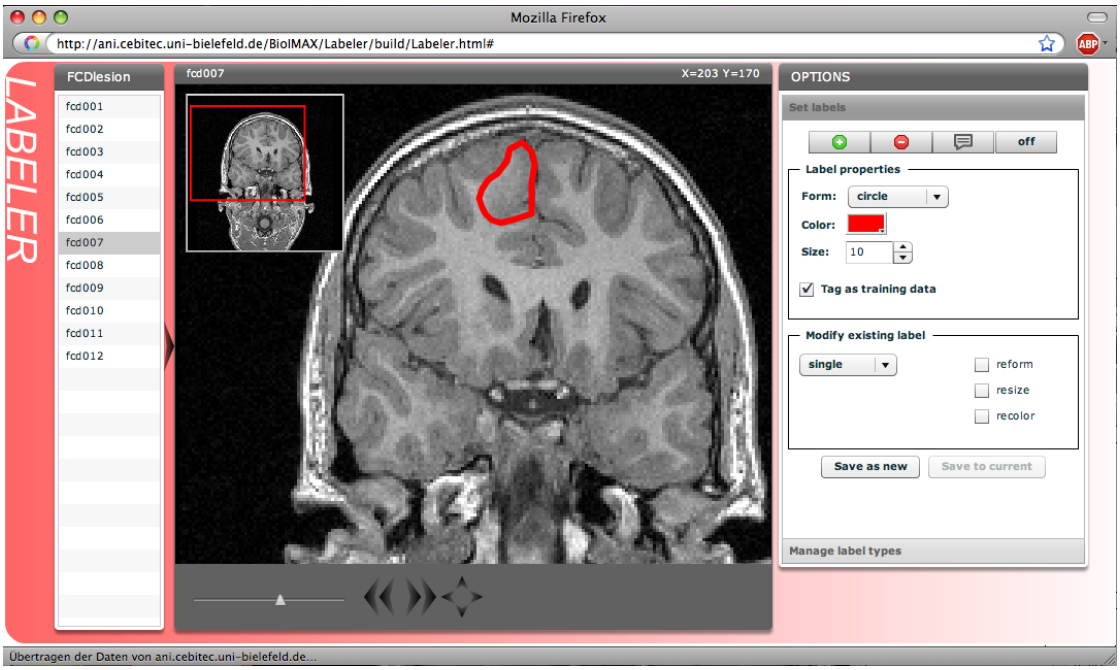


Figure 7.10: Outlining FCD lesions for automatic lesion detection. This figure examplary illustrates the precise labeling of dysplastic brain regions with the *BioIMAX Labeler*. In this way, several experts could manually label a large amount of MRIs in a collaborative way with the objective to create a data basis of lesional regions. These regions could be used by experienced image analysis experts as training samples for the development of an automatic lesion detection algorithm. In addition, the *Labeler* provides the possibility to discuss the quality of the labeled regions via the chat functionality, e.g., how precise the label describes the lesional tissue. The *Labeler* would allow radiologic experts to set up a gold standard of FCD positive image regions with a high quality for the development and evaluation of an automatic FCD lesion classifier.

confirmed pathological cases, in order to allow generalized analysis results. This data basis, which needs to include manually labeled FCD lesions, has to be provided by experienced neuroradiologists. Both, the establishment of a large data basis of histologically proven FCD cases and FCD-negative cases as control group as well as the process of manually labeling the lesional image regions are critical parts for the experts. This is the point where *BioIMAX* could be of assistance:

Creation of MRI data basis The creation of comprehensive data basis by one radiologist is a laborious task, which usually has to be accomplished in addition to her/his daily work. Therefore, it is more reasonable that several radiologists provide limited sets of MRI data instead of single individuals providing a huge amount of sets. Here, particular attention has to be paid on gathering image data that has been acquired with the same image acquisition parameters, in order to guarantee a standard of comparable image data. Via *BioIMAX* several experts at different locations would be able collect their

MRI data in centrally organized projects being available to collaborating experts, e.g., researchers from the fields of mathematics, statistics, or computer science experienced in advanced image analysis. As soon as new MRI data would be available the existing data basis could be extended on the fly. In this way, the time and effort for each expert would be reduced considerably and it would result in a large data basis in less time. Another important aspect is, that a time-consuming and complicated exchange of MRI data from one expert to another would be avoided by using the shared *BioIMAX* database.

Labeling of FCD lesions Second, the development of algorithms and methods for automatic identification of lesional brain tissue requires a substantial number of training samples, i.e., manually labeled image regions identified as FCD lesions. Pixels of these lesional brain regions are used for one input class and pixels of healthy brain regions for the other class of training samples for the SVM classifier. The labeling of the lesions in MRIs is a tedious procedure and has to be carefully performed, since the more precise the malformations are outlined, the more accurate the results of an automatic detection or classification algorithm are. With *BioIMAX* expert users would be able to upload and store MRI data as well as to manually label lesional regions in a sequence of 3D-MRI slices, thereby directly linking and storing labels to the respective images. As a consequence, the radiologists would not require any additional software for labeling image data. Collaborating *BioIMAX* users could directly inspect MRI data with linked labels without installing the same software or routines as the creator of the label. Another major advantage of *BioIMAX* would be, that the *Labeler* provides the opportunity to communicate about labeled image regions. Via the chat functionality of the *Labeler*, several experts could discuss about the quality of labels characterizing FCD lesions. They could communicate about how precise the labels describe the lesions, which is often a difficult task. As a result, it would be possible to set up a gold standard with a high quality for the training of an automatic FCD lesion classifier. In the same way, results of a classification procedure, i.e., automatically generated labeled regions, could be stored to the *BioIMAX* system and could be shared, discussed, and evaluated with collaborating experts.

The above-mentioned example scenarios illustrate, that *BioIMAX* could also be a supporting medium for other fields of application in addition to the purpose of merely analyzing multivariate image data, e.g., in the diagnostic and research of epilepsy based on medical image data. In this context, the application of *BioIMAX* could strengthen the network of experts in radiology and fosters the exchange of diagnostic knowledge and observations regarding epilepsy. Furthermore, *BioIMAX* could serve as a platform for the communication of non-expert and experienced radiologists. Finally, current research topics such as automatic analysis and aided diagnosis of epilepsy could be accelerated and simplified by bringing experts of different disciplines together via the standardized *BioIMAX* functionalities and principles. In sum, *BioIMAX* could be considered as a combination of an image database system and a data sharing platform in epilepsy, as proposed in (Siadat et al., 2005), and a

social media network, such as the *sermo*[1] online platform tailored to US physicians with the objective to collaborate on various clinical issues.

[1] http://www.sermo.com/

CHAPTER 8

Discussion

In this work the Web2.0 system *BioIMAX* for the collaborative exploration and analysis of various types of multivariate and high-content bioimage data is presented. The major objective of *BioIMAX* is to provide an easy-to-use Web-based workbench that augments essential aspects of bioimage analysis in modern life science projects. *BioIMAX* is a novel information technology approach that fosters the integration of different views and perspectives regarding different aspects in bioimage interpretation and analytics. This ranges from manual annotation based on direct visual exploration to fully automatic datamining techniques, within a collaborative and user-shaped Web-based environment. In this chapter the essential cornerstones of the *BioIMAX* system are discussed.

8.1 Bioimage Data Analysis

Due to the increased complexity of high-content and multivariate image data, where a growing number of n variables or signals is associated to each spatial location (pixel) of the sample, the generation of analysis strategies or specific hypotheses is a difficult task and typically not straightforward, since little *apriori* knowledge is available for the data at hand. Therefore, the *BioIMAX* system provides several tools and methods that allow researchers to gain initial visual insights into the structural characteristics of the multivariate signal domain.

After a short registration process, researchers have direct access to the full functionalities of *BioIMAX* . The careful design of the *BioIMAX* data model, which includes a user and data management in combination with a rights privilege management, facilitates users to

easily import new image data to the *BioIMAX* database and to search, retrieve, manage and remove their data throughout the entire analysis process. This is graphically supported by the *BioIMAX Data Browser* that allows users to interactively "browse" their data using various search or filter operations. Additionally, the *Data Browser* serves as the central starting point for any data manipulation or exploration tasks.

Using the *BioIMAX VIStoolBox* users have several options to visually inspect and explore the multivariate image domain gaining different views and perspectives of the data. This tool does not explicitly include predefined analysis workflows regarding a special analytical or biological question, as other bioimage informatics tools typically do. In fact, with the *VIStoolBox* human experts are directly involved in the knowledge discovery process, while shifting and adjusting the exploration goals themselves using interactive visualization displays that provide methods from the fields of exploratory data analysis, visual datamining and information visualization, following Ben Shneiderman's information seeking mantra: *Overview first, zoom in and filter, and details on demand* (Shneiderman, 1996). Such a visual data exploration usually allows a faster and a more intuitive access to the data and provides a much higher degree of confidence in the findings from the exploration, since the human directly interact with the raw image data (Keim, 2002). This kind of data analysis fosters the hypothesis generation process by integrating different perspectives of the data into the human expert's mental model of the data.

It is evident that, with increasing number of variables, i.e., number of image channels, a sole manual evaluation and exploration is unfeasible to get a larger picture of the *n*-variate data domain. Thus, methods are required, which process the image content in such a way that it is comprehensible by the human expert. Typically, sophisticated datamining and unsupervised learning techniques are well suited to represent the high-dimensional data space in a more compact way, e.g., using dimension reduction or clustering methods, allowing for the generation of descriptive models based on selected subsets of image data. For this reason, the *BioIMAX* architecture provides the flexibility to integrate such algorithms. In Chapter 6.6.1 and 6.6.2 two prototype tools are described that allow the application of sophisticated datamining and unsupervised learning algorithms that run on a powerful server as these algorithms are typically time-consuming and computationally expensive. The results of these algorithms are directly integrated in the *BioIMAX* database and can subsequently be visualized and manually explored with specially designed *BioIMAX* graphical interfaces. This is of great benefit for scientists, since it supports the visual exploration of higher-dimensional image data without the need to install external machine learning toolboxes on user's desktop.

8.2 Collaboration

As image data produced by modern multivariate and high-content imaging modalities is increasingly getting richer and incorporates much more complex information, it is impossible to access, quantify and extract all relevant image information in one session by one researcher. As stated in the motivation of this thesis, image data needs to be evaluated by scientists from different disciplines regarding different analysis aspects. Thus, collaboration is a crucial

point in bioimage analysis and therefore, has played an important role in the realization of the *BioIMAX* system.

Since all image data and its potential analysis results are stored and organized in a centralized data repository, the *BioIMAX* system facilitates easy sharing of the expert's data with collaborating researchers. With the *BioIMAX* project concept users can easily build up small communities of collaborating experts, sharing a specific subset of their data regarding a defined biological or analytical question. User-defined projects support a clear organization of data allowing involved members to quickly access, read and analyze project-relevant data. This kind of data sharing is the first and necessary step towards collaborative work in *BioIMAX*.

Another important collaborative aspect in *BioIMAX* is data and analysis reproducibility. Since *BioIMAX* is a purely Web-based software solution with a centralized data repository, all users work on the same copy of an image using the same analysis and exploration methods. This can potentially prevent ambiguity or misinterpretations in the analysis process, which are the problems frequently occurring, when researchers apply their individual exploration procedures and analysis routines on the same data that result in different findings and outcomes. Thus, the usage of *BioIMAX* will lower these general hurdles and trigger a convergence of the mental models, different experts have for the same data.

With the *Labeler* the *BioIMAX* system provides a powerful tool that allows users to graphically and semantically annotate image regions in single image channels via the Web. This is a valuable tool, which fosters crucial communicative processes between collaborating researchers from different institutes that are geographically distributed and simplifies and speeds up essential aspects in data analysis. First, the *Labeler* allows a group of researchers to collaboratively annotate image regions with semantic categories using the same graphical Web-based interface. In particular in bioimage analysis, focussing on specific image regions of interest is often necessary, e.g., regarding the development of new data processing methods or to quantify and evaluate analysis results regarding the accuracy of applied algorithms. The application example in Chapter 7.1 illustrates a potential scenario, where a group of independent experts apply the *Labeler* to generate a gold standard, in order to evaluate an automatic classification algorithm.

In addition to annotating regions with semantic labels, the option to link chat-like discussions to image regions is of particular value for geographically distributed scientists. They can easily communicate about selected image regions, e.g., by adding comments, questions or even high-level discussions to each graphical label, preventing complicated and time-consuming exchange of data and information, e.g., via e-mail or CDs/DVDs. This facilitates collaborative work on one image in a Web2.0 fashion, while the stored states of communication content are directly linked to image regions.

8.3 Architecture

BioIMAX was designed as a Rich Internet Application using the Adobe Flex framework. Adobe Flex allows developers to create powerful Web applications with graphic and interac-

tivity capabilities usually featured only in desktop applications. In this way, it is possible to combine the Web's lightweight collaboration and distribution architecture with the interface interactivity and computation power of desktop applications. This implies the key feature of *BioIMAX*, which is not owned by any other existing bioimage informatics tool. Except for the installation of the Flash Player plugin, users do not need to go through cumbersome installation routines of software packages or libraries and have direct access to sophisticated and powerful bioimage exploration and analysis facilities via the Web. This is independent of their whereabouts and the computing platform they are using, only a valid account with *username* and *password* is required.. The benefit of the *BioIMAX* architecture is, that it integrates both, the application of sophisticated visual exploration and datamining methods and essential collaboration and communication aspects regarding several facets of the analysis of multivariate and high-content bioimage data in one powerful and interactive Web-based infrastructure.

The basic design of the *BioIMAX* data model, with the View concept as central entity, allows developers to flexibly integrate new or extended data types, which have been generated by new analysis or exploration tools, into the existing system without adapting or changing the entire model. This aspect has already been confirmed by the successful and easy integration of the TICAL and WHIDE tools (described in Chapter 6.6.2) within student projects. The development of new analysis tools is carried out independently from the existing *BioIMAX* system and does not need to be integrated into the present *BioIMAX* graphical user interface. They are implemented and tested as stand-alone Flash applications, which are connected to *BioIMAX* via defined interfaces and running within external browser windows. Therefore, a tedious and complicated embedding of the implemented routines into the existing *BioIMAX* source code is avoided. The structural partitioning of the *BioIMAX* platform into modular and almost independent software components has further advantages. The extension or upgrades of *BioIMAX* is less prone to error and the software is easier to maintain. Furthermore, it provides a clearly arranged graphical interface structure via single browser windows, which is of particular interest to users having available multiple computer displays. Finally, the performance of the entire system is considerably increased, since selected tools are invoked and loaded at runtime as required and not all at once during system startup.

The potential drawback of the *BioIMAX* system regarding the insecurity of storage of scientific data on remotely hosted central data servers via the Web is weakened considerably through the realization of user-specific and -controlled data access rights. The project concept of *BioIMAX* allows users to grant read/write privileges to individuals for selected datasets. In this way, confidential data is protected from other (unauthorized) users as far as possible. However, there will always be a residual risk, which usually persists by using the Internet. The general problem of storing confidential data, results, and scientific findings on remote servers is still a sensitive issue in modern science. However, a change in thinking is recognizable in many research fields in the last years. As an example, in genome research, new identified sequences or array data is commonly logged in central databases when a scientific paper is submitted. Thus, the confidential information is now available to the public prior to its publication, even if the paper is not accepted.

8.4 Summary

The aim of *BioIMAX* is neither to provide a Web-based LIMS (Laboratory Information Management System) offering a set of features that support modern laboratory's operations such as workflow or meta data management nor to realize a comprehensive analysis platform covering all aspects in bioimage data analysis. The design of *BioIMAX* is rather focussed on a quick collaborative visual exploration and analysis of a large variety of complex multivariate bioimages ranging from spectral data to multi-tag fluorescence images. With *BioIMAX* distributed researchers from different disciplines are able to easily share their data and expertise regarding several analytical or biological questions. This allows a much earlier integration of their knowledge in research as it is usually done in many distributed projects fostering the whole knowledge discovery and hypotheses finding process.

CHAPTER 9

Conclusion and Outlook

With *BioIMAX* the potential of Web2.0 technologies for the analysis of complex bioimage data has been demonstrated in this thesis. It has been shown, that modern Web technologies in the form of RIAs are extremely powerful and allows the development of sophisticated scientific Web applications running in the Web browser with the computation power of desktop applications. *BioIMAX* exploits this advanced interface interactivity and combines it with the communication and collaboration capabilities offered by the very nature of the Internet. In life science, collaboration between experts of different disciplines located at geographically distributed locations is a crucial part in scientific reasoning, in particular regarding bioimage analysis, and is still a time-consuming and laborious task. Using *BioIMAX*, the collaboration hurdles can considerably be lowered while shifting sophisticated analysis and interactive visual exploration of bioimage data from the user's desktop to the Web. Thus, *BioIMAX* permits a faster and easier evaluation and discussion of ideas, hypotheses, data or results via the Internet independently of the user's whereabouts. We believe, that our approach allows new perspectives in the analysis of complex high-content image data, complement to the existing bioimage informatics solutions.

9.1 Perspectives

Based on the existing system and its benefits that have been reported in this thesis, several extensions and future development potentials of the *BioIMAX* system can be envisioned.

In a first step, the datamining tools described in Chapter 6.6, which are currently available

only as prototype versions, should be fully build into the *BioIMAX* system as soon as the testing phase has been completed. In addition to this necessary step, existing and already integrated tools could be extended with advanced functionalities, in order to increase their usability or flexibility regarding bioimage analysis and collaborative exploration.

In their current status, the visualization displays provided by the *VIStoolBox* are independent units as described in Section 6.5.4. Interactivity of each n-variate visualization technique is limited only to the image domain, in which pixels of image regions are highlighted corresponding to a selection of data point representations in a respective plot. All other plots that are visualizing actually the same image information in another way are currently not affected by interactive selections in the first plot. An important extension of the *VIStoolBox* might be the development and implementation of advanced interactivity facilities that allow the linkage of different data displays via Link-and-Brush. Thus, data points selected or highlighted in one display triggers highlighting of corresponding data points in all other open plots, e.g., as it is applied in the CellProfiler Analyst software (Jones et al., 2008). In this way, users are able to explore the image domain with multiple visualization displays from different perspectives and at different levels simultaneously. The connection of different plot types fosters the examination of relationships in the data, in particular when exploring a large number of data points.

Another aspect future developments of *BioIMAX* should be focussed on is the integration of advanced communication and collaboration features. The steadily developing and improving Internet technologies allow developers to easily integrate state-of-the-art Web communication facilities such as instant messaging clients, video conference tools, or synchronous remote collaboration facilities into rich-client Web applications. In this way, a number of scientists could meet in a kind of a virtual conference room, where they could interactively share, edit, and discuss their data, visualizations, or results in real-time multi-user sessions independent of their whereabouts. Thus, the existent asynchronous communication functionalities of the *Labeler* interface could be extended by these modern communication media, which fosters a more dynamic and faster communication between scientists being at different locations in real-time.

A further question concerns the integration of *meta data* into the *BioIMAX* system. Bioimage data is often associated with additional meta data or meta information. Meta data usually describes facts about the image data at hand, i.e., imaging parameters such as settings of the used imaging modality, biological or histological information about the imaged sample, or records about patients or organisms the imaged samples are taken from. The ability to store meta data is of particular importance for image retrieval tasks, for specific analysis issues, or for protocoling observations prior to the actual analysis. Depending on the biological or analytical questions and the type of image data at hand, meta data is often structured differently and its complexity can vary from experiment to experiment or even from image to image. Different types of meta data are available as plain text, XML, tables, lists, or additional documents in proprietary formats containing texts, images, or graphics. Since *BioIMAX* is a platform for arbitrary types of multivariate bioimage data, it is hardly feasible to develop and integrate a standardized meta data format into the *BioIMAX* system, which meets the dynamic requirements of all types of bioimage meta data. Rather, scientists need

an appropriate graphical toolkit that allows them to generate and manage their individual meta data structure and to link this structure to the respective image data. Such a toolkit should provide a set of graphical widgets, containers, or controls that can be arbitrarily arranged in a window using a flexible layout manager. In this way, *BioIMAX* users can design their own meta data layout using text boxes, lists, table grids, tree views, or hyperlinks to additional documents centrally stored in *BioIMAX* database. Such a new meta data tool is currently under development within a student project and will be integrated into *BioIMAX* when it is finished.

Finally, the increasing emergence and the greater use of specialized Web-based information systems in various life science disciplines, such as the *BioIMAX* system or diverse genome, metabolome, or proteome information systems, lead to a continuously growing amount of data and information with various complexity and structure. This poses the question, how to systematically capture, structure, retain, and reuse this data and information, in order to draw a comprehensive picture of how particular systems works, and subsequently how to convey the generated knowledge meaningfully to other information systems. This is especially important regarding research fields that necessarily depend on knowledge integration, such as systems biology. In recent years, tremendous effort has been spent in the field of biological knowledge management and knowledge representation such as formulating ontologies and syntaxes, that allow to semantically describe biological knowledge in a standardized and machine-readable way, so that it can be exchanged more easily and universally between different systems. As a consequence a new concept of the Web has been carried out: the Semantic Web (Lee et al., 2001). Semantic Web technologies aim at meeting the challenges of knowledge management, promising an infrastructure that comprises machine understandable content, and, therefore, a World Wide Web of semantically linked data (Antezana et al., 2009). In sum, knowledge management and representation will play a crucial role in future developments of Web-based applications and information systems in life science, combining both, the advances of Web2.0 and the ideas of the Semantic Web resulting in a new area of the Web: the Web3.0. The successful development of Web3.0 applications taking full advantage of the integrative and analytical potential of knowledge management will support the consolidation of all facets of the knowledge discovery process in life science research.

List of Figures

Bibliography

Abramoff, M. D., Magelhaes, P. J., and Ram, S. J.: Image processing with ImageJ. *Biophotonics international*, 11(7):36–42 (2004).

Agrawal, R., Imieliński, T., and Swami, A.: Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, 207–216 (1993).

Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, volume 1215, 487–499 (1994).

Albaum, S., Neuweger, H., Fränzel, B., Lange, S., Mertens, D., Trötschel, C., Wolters, D., Kalinowski, J., Nattkemper, T., and Goesmann, A.: Qupe—a Rich Internet Application to take a step forward in the analysis of mass spectrometry-based quantitative proteomics experiments. *Bioinformatics*, 25(23):3128 (2009).

Allaire, J.: Macromedia Flash MX—A next-generation rich client. *Macromedia white paper*, 1–2 (2002).

Angeletti, C., Harvey, N. R., Khomitch, V., Fischer, A. H., Levenson, R. M., and Rimm, D. L.: Detection of malignancy in cytology specimens using spectral–spatial analysis. *Laboratory investigation*, 85(12):1555–1564 (2005).

Antezana, E., Kuiper, M., and Mironov, V.: Biological knowledge management: the emerging role of the Semantic Web technologies. *Briefings in bioinformatics*, 10(4):392–407 (2009).

Arif, M., Rajpoot, N. M., Nattkemper, T. W., Technow, U., Chakraborty, T., Fisch, N., Jensen, N. A., and Niehaus, K.: Quantification of cell infection caused by Listeria monocytogenes invasion. *Journal of Biotechnology* (2011).

Bairoch, A., Apweiler, R., Wu, C., Barker, W., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al.: The universal protein resource (UniProt). *Nucleic acids research*, 33(suppl 1):D154–D159 (2005).

Baumbach, J. I.: Ion mobility spectrometry coupled with multi-capillary columns for metabolic profiling of human breath. *Journal of Breath Research*, 3:034001 (2009).

Baumbach, J. I. and Eiceman, G. A.: Ion mobility spectrometry: arriving on site and moving beyond a low profile. *Applied spectroscopy*, 53(9):338A (1999).

Bödeker, B., Vautz, W., and Baumbach, J.: Visualisation of MCC/IMS-data. *International Journal for Ion Mobility Spectrometry*, 11(1):77–81 (2008).

Becker, R. A. and Cleveland, W. S.: Brushing scatterplots. *Technometrics*, 29(2):127–142 (1987).

Bunkowski, A.: Software tool for coupling chromatographic total ion current dependencies of GC/MSD and MCC/IMS. *International Journal for Ion Mobility Spectrometry*, 1–7 (2010).

Card, S., Mackinlay, J., and Shneiderman, B.: *Readings in information visualization: using vision to think*. Morgan Kaufmann (1999).

Carpenter, A. E., Jones, T. R., Lamprecht, M. R., Clarke, C., Kang, I. H., Friman, O., Guertin, D. A., Chang, J. H., Lindquist, R. A., Moffat, J., et al.: CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, 7(10):R100 (2006).

Chen, C.: *Information visualization: Beyond the horizon*. Springer-Verlag New York Inc (2004).

Colliot, O., Antel, S., Naessens, V., Bernasconi, N., and Bernasconi, A.: In Vivo Profiling of Focal Cortical Dysplasia on High-resolution MRI with Computational Models. *Epilepsia*, 47(1):134–142 (2006).

Cortes, C. and Vapnik, V.: Support-vector networks. *Machine learning*, 20(3):273–297 (1995).

Cottrell, W. J., Wilson, J. D., and Foster, T. H.: Microscope enabling multimodality imaging, angle-resolved scattering, and scattering spectroscopy. *Optics letters*, 32(16):2348–2350 (2007).

Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N., and Weerawarana, S.: Unraveling the Web services web: an introduction to SOAP, WSDL, and UDDI. *Internet Computing, IEEE*, 6(2):86–93 (2002).

Fox, P. and Hendler, J.: Changing the equation on scientific data visualization. *Science*, 331(6018):705 (2011).

Fraternali, P., Rossi, G., and Sánchez-Figueroa, F.: Rich internet applications. *Internet Computing, IEEE*, 14(3):9–12 (2010).

Garrett, J.: Ajax: A new approach to web applications. *essay, Adaptive Path, San Francisco, CA* (2007).

Geladi, P. and Grahn, H.: *Multivariate image analysis*. Wiley Online Library (1996).

Goldberg, I., Allan, C., Burel, J. M., Creager, D., Falconi, A., Hochheiser, H., Johnston, J., Mellen, J., Sorger, P., and Swedlow, J.: The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging. *Genome Biology*, 6(5):R47 (2005).

Herold, J.: *A data mining approach for high content fluorescence microscopy images of tissue samples*. PhD thesis, Bielefeld University, Bielefeld (2010).

Herold, J., Loyek, C., and Nattkemper, T. W.: Multivariate image mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):2–13 (2011).

Hiraoka, Y., Shimi, T., and Haraguchi, T.: Multispectral imaging fluorescence microscopy for living cells. *Cell structure and function*, 27(5):367–374 (2002).

Huppertz, H., Grimm, C., Fauser, S., Kassubek, J., Mader, I., Hochmuth, A., Spreer, J., and Schulze-Bonhage, A.: Enhanced visualization of blurred gray–white matter junctions in focal cortical dysplasia by voxel-based 3D MRI analysis. *Epilepsy research*, 67(1):35–50 (2005).

Inselberg, A. and Dimsdale, B.: Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the 1st conference on Visualization'90*, 361–378 (1990).

Ireton, K.: Entry of the bacterial pathogen Listeria monocytogenes into mammalian cells. *Cellular microbiology*, 9(6):1365–1375 (2007).

Johnston, J., Nagaraja, A., Hochheiser, H., and Goldberg, U.: A flexible framework for Web interfaces to image databases: supporting user-defined ontologies and links to external databases. In *Biomedical Imaging: Nano to Macro, 2006. 3rd IEEE International Symposium on*, 1380–1383 (2006).

Jones, T. R., Kang, I. H., Wheeler, D. B., Lindquist, R. A., Papallo, A., Sabatini, D. M., Golland, P., and Carpenter, A. E.: CellProfiler Analyst: data exploration and analysis software for complex image-based screens. *BMC bioinformatics*, 9(1):482 (2008).

Keim, D. A.: Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, 8(1):1–8 (2002).

Kvilekval, K., Fedorov, D., Obara, B., Singh, A., and Manjunath, B. S.: Bisque: a platform for bioimage analysis and management. *Bioinformatics*, 26(4):544 (2010).

Lamprecht, M. R., Sabatini, D. M., and Carpenter, A. E.: CellProfiler^{TM}: free, versatile software for automated biological image analysis. *Biotechniques*, 42(1):71 (2007).

Langenkämper, D., Kölling, J., Abouna, S., Khan, M., Niehaus, K., and Nattkemper, T.: TICAL-a web-tool for multivariate image clustering and data topology preserving visualization. In *MIAAB - Microscopic Image Analysis with Applications in Biology*. Heidelberg, Germany (2011).

Laurent, S., Johnston, J., and Dumbill, E.: *Programming web services with XML-RPC*. O'Reilly Media (2001).

Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. *Scientific American*, 284(5):34–43 (2001).

Letovsky, S., Cottingham, R., Porter, C., and Li, P.: GDB: the human genome database. *Nucleic Acids Research*, 26(1):94 (1998).

Levenson, R. M. and Hoyt, C. C.: Spectral imaging and microscopy. *American Laboratory*, 32(22):26–33 (2000).

Lew, M., Sebe, N., Djeraba, C., and Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(1):1–19 (2006).

Long, F., Peng, H., Liu, X., Kim, S., and Myers, E.: A 3D digital atlas of C. elegans and its application to single-cell analyses. *Nature methods*, 6(9):667–672 (2009).

Loyek, C., Bunkowski, A., Vautz, W., and Nattkemper, T.: Web2. 0 paves new ways for collaborative and exploratory analysis of Chemical Compounds in Spectrometry Data. *Journal of integrative bioinformatics*, 8(2):158 (2011a).

Loyek, C., Kölling, J., Langenkämper, D., Niehaus, K., and Nattkemper, T.: A Web2. 0 Strategy for the Collaborative Analysis of Complex Bioimages. *Advances in Intelligent Data Analysis X*, 258–269 (2011b).

Loyek, C., Rajpoot, N., Khan, M., and Nattkemper, T.: BioIMAX: A Web 2.0 approach for easy exploratory and collaborative access to multivariate bioimage data. *BMC bioinformatics*, 12(1):297 (2011c).

Loyek, C., Woermann, F., and Nattkemper, T.: Detection of Focal Cortical Dysplasia Lesions in MRI using Textural Features. *Bildverarbeitung für die Medizin 2008*, 432–436 (2008).

Manders, E. M., Stap, J., Brakenhoff, G. J., Van Driel, R., and Aten, J. A.: Dynamics of three-dimensional replication patterns during the S-phase, analysed by double labelling of DNA and confocal microscopy. *Journal of cell science*, 103(3):857 (1992).

Martinetz, T., Berkovich, S., and Schulten, K.: Neural-gas' network for vector quantization and its application to time-series prediction. *Neural Networks, IEEE Transactions on*, 4(4):558–569 (1993).

Megason, S. G. and Fraser, S. E.: Imaging in systems biology. *Cell*, 130(5):784–795 (2007).

Meyer, F., Goesmann, A., McHardy, A., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R., et al.: GenDB—an open source genome annotation system for prokaryote genomes. *Nucleic acids research*, 31(8):2187 (2003).

Micheva, K. D. and Smith, S. J.: Array tomography: a new tool for imaging the molecular architecture and ultrastructure of neural circuits. *Neuron*, 55(1):25–36 (2007).

Murphy, R. F.: Putting proteins on the map. *Nature biotechnology*, 24(10):1223–1224 (2006).

Nattkemper, T. W.: Multivariate image analysis in biomedicine. *Journal of Biomedical Informatics*, 37(5):380–391 (2004).

Neuweger, H., Albaum, S., Dondrup, M., Persicke, M., Watt, T., Niehaus, K., Stoye, J., and Goesmann, A.: MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics*, 24(23):2726–2732 (2008).

Ontrup, J. and Ritter, H.: Large-scale data exploration with the hierarchically growing hyperbolic SOM. *Neural networks*, 19(6-7):751–761 (2006).

OReilly, T.: What is Web 2.0: Design patterns and business models for the next generation of software (2007). URL http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-isweb-20.html.

Peng, H.: Bioimage informatics: a new area of engineering biology. *Bioinformatics*, 24(17):1827 (2008).

Peng, H., Long, F., and Myers, E. W.: VANO: a volume-object image annotation system. *Bioinformatics*, 25(5):695 (2009).

Pepperkok, R. and Ellenberg, J.: High-throughput fluorescence microscopy for systems biology. *Nat Rev Mol Cell Biol*, 7(9):690–696 (2006).

Pepperkok, R., Simpson, J. C., and Wiemann, S.: Being in the right location at the right time. *Genome Biol*, 2(9):1024.1–1024.2 (2001).

Ramaswamy, V., Cresence, V. M., Rejitha, J. S., Lekshmi, M. U., Dharsana, K. S., Prasad, S. P., and Vijila, H. M.: Listeria-review of epidemiology and pathogenesis. *JOURNAL OF MICROBIOLOGY IMMUNOLOGY AND INFECTION*, 40(1):4 (2007).

Saalfeld, S., Cardona, A., Hartenstein, V., and Tomančák, P.: CATMAID: collaborative annotation toolkit for massive amounts of image data. *Bioinformatics*, 25(15):1984 (2009).

Schubert, W., Bonnekoh, B., Pommer, A. J., Philipsen, L., B\textbackslashöckelmann, R., Malykh, Y., Gollnick, H., Friedenberger, M., Bode, M., and Dress, A. W.: Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nature biotechnology*, 24(10):1270–1278 (2006).

Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, 336–343 (1996).

Shneiderman, B.: Science 2.0. *Science*, 319(5868):1349–1350 (2008).

Siadat, M., Soltanian-Zadeh, H., Fotouhi, F., and Elisevich, K.: Content-based image database system for epilepsy. *Computer methods and programs in biomedicine*, 79(3):209–226 (2005).

Simons, D. J. and Rensink, R. A.: Change blindness: Past, present, and future. *Trends in Cognitive Sciences*, 9(1):16–20 (2005).

Spence, R.: Information visualization: Design for interaction. In *Proceedings of CHI 2005 Conference on Human Factors in Computing Systems* (2007).

Starkuviene, V. and Pepperkok, R.: The potential of high-content high-throughput microscopy in drug discovery. *British Journal of Pharmacology*, 152(1):62–71 (2007).

Swedlow, J. R., Goldberg, I., Brauner, E., and Sorger, P. K.: Informatics and quantitative analysis in biological imaging. *Science*, 300(5616):100 (2003).

Swedlow, J. R., Goldberg, I. G., and Eliceiri, K. W.: Bioimage Informatics for Experimental Biology*. *Annual review of biophysics*, 38:327–346 (2009).

Tassi, L., Colombo, N., Garbelli, R., Francione, S., Lo Russo, G., Mai, R., Cardinale, F., Cossu, M., Ferrario, A., Galli, C., et al.: Focal cortical dysplasia: neuropathological subtypes, EEG, neuroimaging and surgical outcome. *Brain*, 125(8):1719 (2002).

Taylor, D., Falconer, M., Bruton, C., and Corsellis, J.: Focal dysplasia of the cerebral cortex in epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry*, 34(4):369 (1971).

Tukey, J.: Exploratory data analysis. *Reading, MA* (1977).

Turner, M., Budgen, D., and Brereton, P.: Turning software into a service. *Computer*, 36(10):38–44 (2003).

Unser, M.: Advanced image processing and analysis using ImageJ. In *8th European Light Microscopy Initiative Meeting*, 27–30 (2008).

Vicens, Q. and Bourne, P. E.: Ten simple rules for a successful collaboration. *PLoS Comp Biol*, 3:e44 (2007).

Vinegoni, C., Ralston, T., Tan, W., Luo, W., Marks, D. L., and Boppart, S. A.: Multimodality imaging of structure and function combining spectral-domain optical coherence and multiphoton microscopy. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 6079, 226–233 (2006).

Waldrop, M. M.: Science 2.0 - Great new tool, or great risk? *Scientific American* (2008).

Walter, T., Shattuck, D. W., Baldock, R., Bastin, M. E., Carpenter, A. E., Duce, S., Ellenberg, J., Fraser, A., Hamilton, N., Pieper, S., et al.: Visualization of image data from cells to organisms. *Nature methods*, 7:S26–S41 (2010).

Ware, C.: *Information visualization: perception for design*, volume 22. Morgan Kaufmann (2004).

Woermann, F., Brandt, C., and Schaumann-von-Stosch, R.: Neuroradiologische Diagnostik in der Epileptologie Clinical Neuroimaging in Epilepsy. *Akt Neurol*, 31(2):60–72 (2004).

Wolkenhauer, O., Kitano, H., and Cho, K. H.: Systems biology. *Control Systems Magazine, IEEE*, 23(4):38–48 (2003).

Yoo, T. S., Ackerman, M. J., Lorensen, W. E., Schroeder, W., Chalana, V., Aylward, S., Metaxas, D., and Whitaker, R.: Engineering and algorithm design for an image processing api: a technical report on itk-the insight toolkit. *Studies in health technology and informatics*, 586–592 (2002).

Zimmermann, T., Rietdorf, J., and Pepperkok, R.: Spectral imaging and its applications in live cell microscopy. *FEBS Letters*, 546(1):87–92 (2003).