Web2.0 paves new ways for collaborative and exploratory analysis of Chemical Compounds in Spectrometry Data

Christian Loyek[1*], Alexander Bunkowski[2], Wolfgang Vautz[3], Tim W. Nattkemper[1]

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

[2]Genome Informatics Group, Faculty of Technology, Bielefeld University, Germany

[3]Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V., Dortmund, Germany

Summary

In nowadays life science projects, sharing data and data interpretation is becoming increasingly important. This considerably calls for novel information technology approaches, which enable the integration of expert knowledge from different disciplines in combination with advanced data analysis facilities in a collaborative manner. Since the recent development of web technologies offers scientific communities new ways for cooperation and communication, we propose a fully web-based software approach for the collaborative analysis of bioimage data and demonstrate the applicability of *Web2.0* techniques to ion mobility spectrometry image data. Our approach allows collaborating experts to easily share, explore and discuss complex image data without any installation of software packages. Scientists only need a username and a password to get access to our system and can directly start exploring and analyzing their data.

1 Introduction

The quality of semi-automatic analysis of raw biodata such as spectra, images, etc. is a crucial point in nowadays life sciences, which cannot be fully automatized. In this work, we consider ion mobility spectrometry (IMS), which is a method to characterize chemical compounds on the basis of gas-phase ions in an electrical field [1]. It has been proven to be a powerful technique to screen complex mixtures like samples from the headspace of cell cultures and even more complex mixtures like human breath [2]. Together with the usage of a multi-capillary column for pre-separation, the resulting data is typically visualized as heat-map images facilitating the detection, quantification, and comparison of chemical compounds within one or more samples. 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. Nowadays, scientific visualization more and more is becoming an integral part of the scientific analysis process, instead of being an end product only illustrating analysis results [3]. Various facets in IMS research leads to challenges at different levels in

*To whom correspondence should be addressed. Email: cloyek@techfak.uni-bielefeld.de

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 collaborating experts. In general, collaboration is nowadays more important than ever before in life science projects [4]. However, such a collaborative analysis of IMS data is a complicated and time-consuming task, since the collaborating scientists are often spread across several research institutes. A typical scientific collaboration scenario looks like the following:

Two collaborating experts at different locations are working on the same IMS data pool regarding the same biological question. Both produce specific results with their individual exploration procedures and analysis routines. These results have to be discussed in regularly scheduled scientific meetings, where each of the experts present their respective results and findings. In the meeting, the experts possibly find out, that their IMS analysis strategies and exploration results led to different findings. At this point, apart from the considerable time spent on such meetings, the first practical problems in collaborative IMS analysis arise. Figuring out the reasons for the different findings is often a difficult and complicated task, since the experts have applied their own complex exploration and analysis strategies and usually have developed specialized tools for their needs, which are not directly available for the other expert, making it impossible to reproduce the results. A short-term solution for this problem would be, that the experts install their respective tools on the workstation of the other expert. However, this leads to further technical problems: What happens, if some parts of the tools change? What if upgrades of depending libraries are available, which are incompatible with the current version of the tools? What if experts are forced to change their operating system?

This example scenario only points out some of the frequent problems occurring in many research projects related to all aspects of molecular biology ranging from genomics to metabolomics and they illustrate general hurdles in collaborative analysis. As a consequence, scientists in IMS research projects need alternatives and new opportunities for collaboration and communication during the exploration and analysis process of IMS data. Since the web is getting more collaborative and user-shaped (effects which are referred to as *Web2.0*) and offers more and more powerful graphics applications, it paves new ways for data analysis and collaboration in scientific projects. Therefore, we are taking advantage of this development of web technologies and demonstrate its potential for collaborative analysis of IMS data.

In this paper, we propose and demonstrate a fully web-based approach for IMS data analysis, which is called *BioIMAX* (**B**io**I**mage **M**ining, **A**nalysis and e**X**ploration). *BioIMAX* allows the integration of both formerly separated aspects, individual exploratory data analysis of complex image data in combination with essential collaboration and communication issues in scientific projects by moving both aspects to the web. *BioIMAX* is the attempt to explore the potential of social network technologies regarding scientific research projects, which is referred to as *Science2.0* [5, 6]. The main objective of *BioIMAX* is, that collaborating scientists can explore and analyze IMS image data and share and discuss their data and findings within one software solution, independent from their whereabouts by using a standard web browser. Scientists only need a username and password to have access to the platform without any installation of additional software or libraries. Such a web-based collaborative work on the same data domain and on the same scientific question is a fundamental step towards integrative bioinformatics in the context of bioimage data analysis. *BioIMAX* can be accessed at `http://ani.cebitec.uni-bielefeld.de/BioIMAX` with the username and the password "testIMS" for testing

purposes.

2 Architecture

BioIMAX has been designed as a *Rich Internet Application* (RIA), implemented with *Adobe Flex* [7]. A RIA is a web application whose performance and look-and-feel is comparable to a standard desktop application, but will be executed in a web browser allowing for platform independency and avoiding additional installation and maintenance costs. The application of RIAs is becoming an increasingly important part of the change of the World Wide Web towards *Web2.0*. After a short registration process, which generates a unique user account, the users directly have access to all functionalities of the *BioIMAX* platform. Through an easy-to-use interface users can import arbitrary sets of IMS heat-map images into the system. All user-generated content, i.e., original IMS image data and derived exploration and analysis results within *BioIMAX* will be stored in a central data repository clearly organized by the relational database management system *MySQL* [8]. For the management of data, *BioIMAX* provides a data browser that allows the user to search, browse, filter and modify own datasets and datasets from other users, provided that they have access privileges. In addition, the data browser is the central component, in order to initiate any processing of selected datasets with integrated exploration or analysis tools. In the following, we describe several aspects of *BioIMAX* focussed on the collaborative exploration and analysis of IMS image data.

3 Collaborative work on IMS data

As mentioned before, scientific collaboration takes place at different levels. With *BioIMAX* the most fundamental aspects in collaboration are covered:

- Sharing of data and results
- Reproducibility of data and analysis results
- Communication and discussion of particular image content

For the sharing of data and results the *BioIMAX* platform provides the concept of a *project*. Each user can create personalized projects with the objective to collect and organize a subset of IMS image data or its results by adding the data to the projects. Once a project has been created, the project owner can invite collaborating users to join her/his project, whereby identifying them as members of the respective project. This project concept is the first step towards collaborative work in *BioIMAX*, which supports sharing of specific datasets, e.g., regarding a defined biological or analytical question and allows the users to rapidly access project relevant data.

Since all accumulating data within *BioIMAX* is stored on a centralized data repository, each user works on the same copy of original IMS images and apply the same set of exploration and analysis routines or tools provided by *BioIMAX*. This prevents ambiguity and misinterpretations during an analysis process and enables a high degree of reproducibility of results or findings obtained from collaborating experts.

In many biodata analyses, it is necessary to evaluate, quantify or localize specific features of the data at hand. This is in particular important concerning bioimage analysis, since images are by their very nature unstructured and are often of high dimensionality. In many cases, image exploration and analysis is focussed on specific image regions of interest, which frequently needs to be discussed with collaborating experts.

As an example, during the analysis of IMS heat-map images, some of the regions cannot clearly assigned to known compounds. Such image regions need to be examined and discussed with experts from different disciplines, e.g., to quantify these regions or to avoid misunderstandings or problems for future analysis. Therefore, *BioIMAX* provides a tool, called *Labeler* that allows the user to graphically and semantically annotate and discuss image regions in single images. A graphical label is characterized by visual properties, e.g., shape, color, size and position, which can be adjusted by the user at any time. Annotations are placed as graphical objects on a layer belonging to each single image (see Figure 1). After saving a set of labels into the database and adding it to a project, it can be accessed and viewed by other project members.

In the context of collaborative evaluation of specific IMS compounds the *Labeler* provides an option to link chat-like discussions to image regions. Several users can communicate about one selected label via a chat window (see Figure 1) and the conversation will additionally be stored together with the label. This facilitates *Web2.0* style collaborative work on one image, while the stored states of communication content are directly linked to image regions.

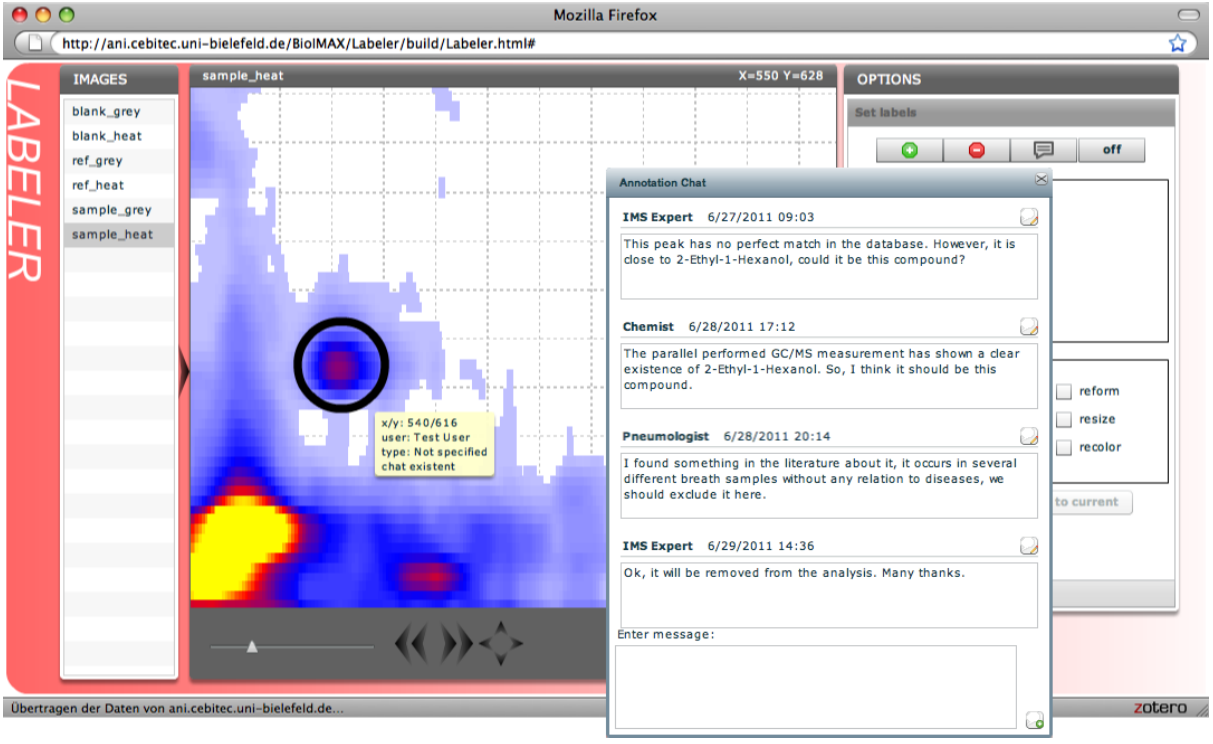


Figure 1: The *Labeler*, illustrating a discussion about a specific image region. One user draws a graphical label on an IMS heat-map and enters a question about the image region that needs to be discussed with collaborating experts from different disciplines in a chat window. The annotation and the chat-like discussion will be linked to the image and stored in the database, so other experts can load this image with the linked conversation and can directly answer or comment the question.

4 Exploratory analysis of IMS data

The generation of new hypotheses and analysis strategies in the biodata domain calls for advanced analysis approaches, which allows scientists to individually explore and visualize the data at hand. This refers in particular to those types of data, where the analysis goal is vague or the valuable information is not directly accessible, e.g., in the comparative exploration of multiple IMS samples. In addition to the identification and quantification of compounds in single IMS images, a typical challenge is the comparative analysis of sets of IMS samples, in order to detect structural differences or similarities between different samples, e.g., to evaluate the quality of a sample against reference data. For this type of comparative data analysis, which is referred to as multivariate image analysis [9], methods and techniques from the fields of *exploratory data analysis*, *information visualization*, and *visual datamining* [10] have proven to be powerful, in order to gain structural insights into the multivariate data domain. The benefit of those analysis techniques is, that the user is directly involved in the knowledge discovery process, while visually exploring the data space themselves following Ben Shneidermans information visualization mantra: *Overview first, zoom in and filter, details on demand.*

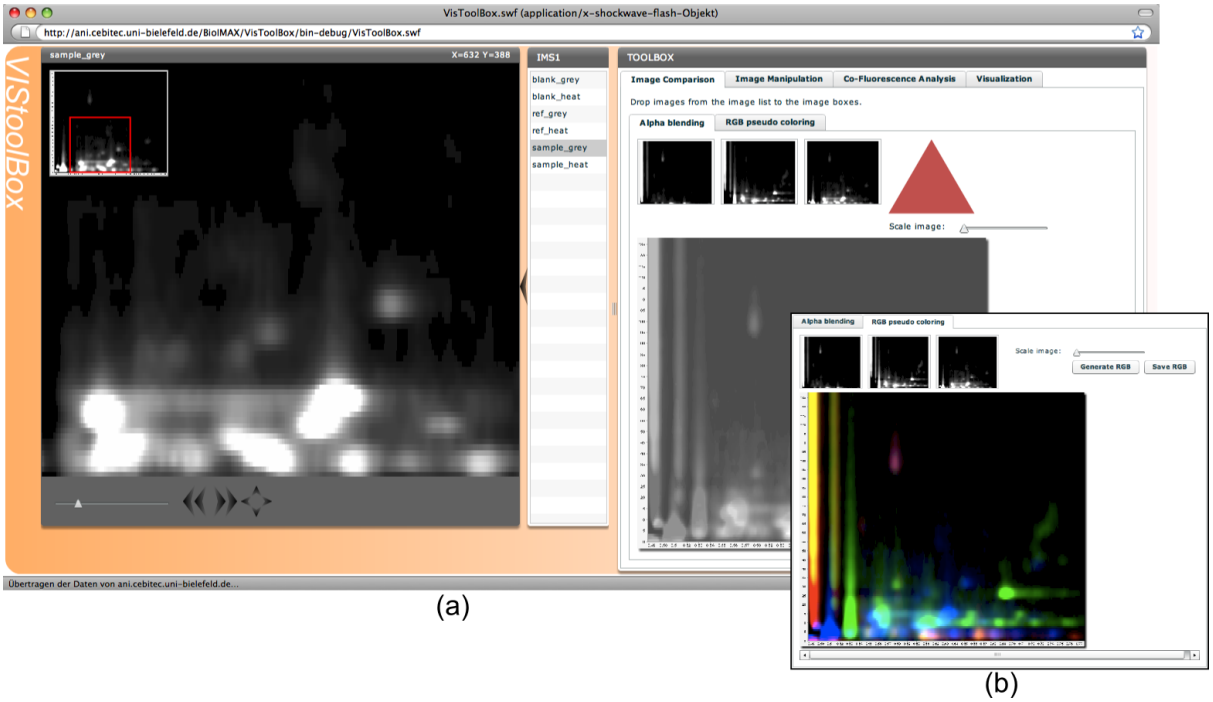


Figure 2: Screenshot of the *VisToolBox*. This interface includes methods for the visual exploration of the raw image data space. Here, two methods are illustrated, which enable a user to compare three images simultaneously on a structural level. The method shown in (a) aims at comparing three images, while superimposing them as layers an manually adjusting the opacity value of the respective layers. In (b) a RGB pseudo color fusion image is generated from three images. With such visualization methods users can immediately identify structural differences or similarities of selected images in a single display.

Since the *BioIMAX* platform is basically focussed on these type of data analysis, it provides several interfaces containing specialized exploration and analysis methods from the aforementioned fields, concentrating on different aspects of the data. In general, the exploration methods provided by *BioIMAX* can be divided into two categories. One category comprises methods for the visual exploration of the raw data domain, i.e., directly analyzing raw signals of single im-

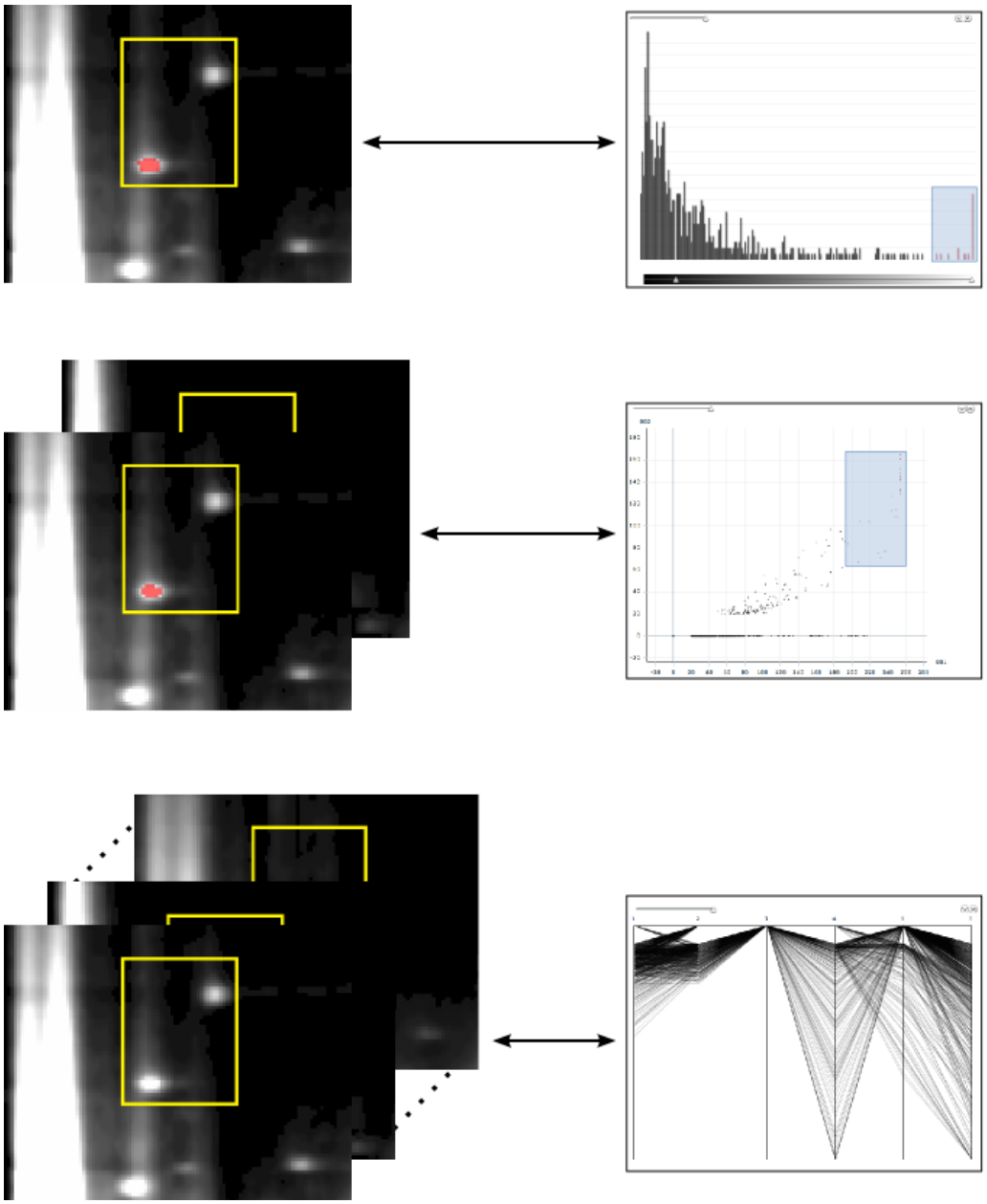
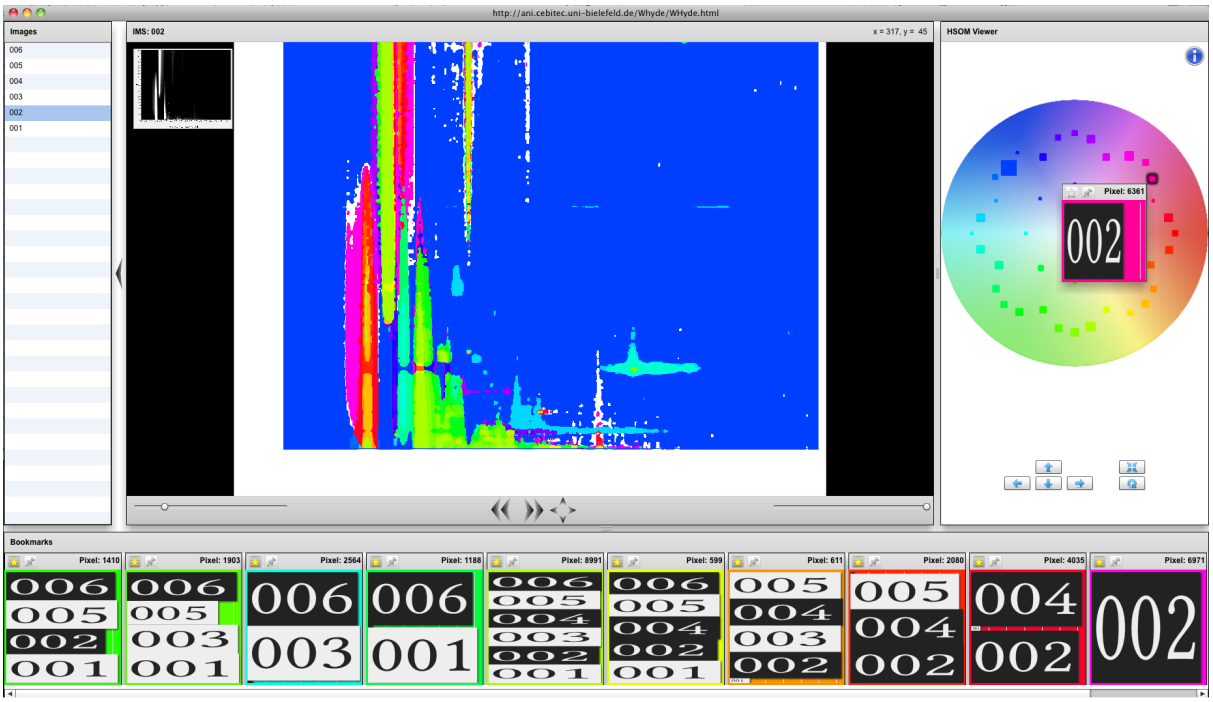


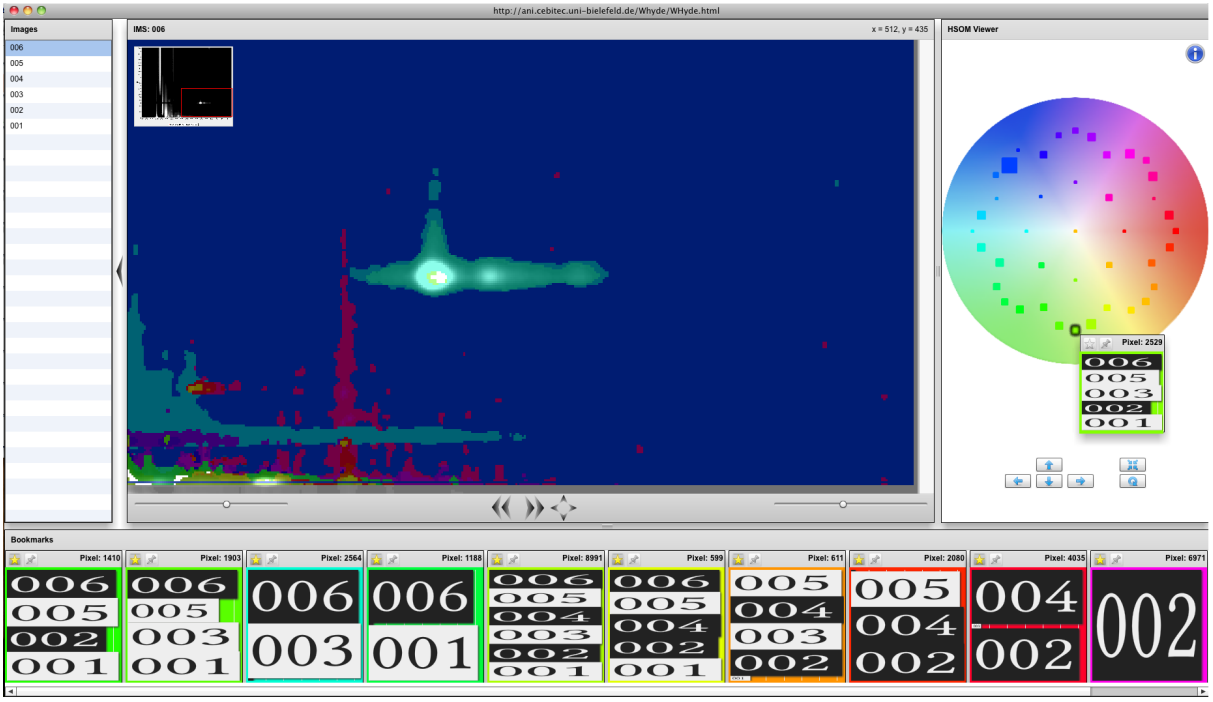
Figure 3: Using *BioIMAX' VisToolBox* users can apply a large variety of visualization tools to explore univariate (top), bivariate (middle) or multivariate (bottom) features of their data, independent from their whereabouts, provided with an internet connection. In this way, different views and perspectives on data are generated and can be integrated into the users' mental models of the data.

ages or exploring signals of multiple images in a comparative manner. Figure 2 and 3 illustrate application examples of exploration tools, which allow scientists to visually "browse through" the original IMS data space spanned by a number of selected images with the objective to extract and visualize the comparative information of image signals. Exploration can take place either on whole images (see Figure 2) or on selected image regions of interest (see Figure 3).

In the second category the visual exploration is associated with a previous reduction of the original image data complexity. *BioIMAX* allows web-based application of sophisticated data mining algorithms, e.g., clustering or dimension reduction, to generate descriptive models based on selected subsets of image data, whose computation can be initiated by directly using the



(a)



(b)

Figure 4: Screenshots of the visualization and exploration interface for clustering results. With *BioIMAX* the user has started a clustering process on the remote server based on selected images, which then can be visualized and explored with this interface. (a) and (b) illustrate different states while visually exploring the clustering result manually adjusted by the user.

BioIMAX interface. As such algorithms usually are computational expensive, they will be performed on a remote compute server, which is connected to the *BioIMAX* system as well as to

the centralized database.[1] Once a model is computed and stored in the database, its results can immediately be visualized and explored by any collaborating *BioIMAX* user, e.g., by adding the model and its results to a defined project. In Figure 4 we illustrate such a scenario with a visualization of six dimensional IMS data set using TICAL, which is BioIMAX' clustering tool. TICAL uses vector quantization clustering to group the D (here $D = 6$) intensity values at each coordinate pair in a spectrum into clusters of similar patterns. Each cluster is represented by a prototype that represents the average D-dimensional intensity pattern of a cluster. Using dimension reduction, the clusters are projected onto a two-dimensional color disc (see Figure 4 on the right with one single cluster prototype displayed). Now one spectrum can be visualized in pseudocolor by 1. mapping each coordinate pair to its cluster according to the best matching unit criterion and 2. draw this coordinate pair in the color of the cluster position on the disc (see Figure 4 in the middle). In the lower rows of the two screenshots, single selected cluster prototypes are displayed (e.g. the cluster on the far right shows a large signal only in the second image "002", the cluster next to it shows a strong signal in the 2nd and fourth images).

5 Typical Workflow Scenario

Existing IMS analysis tools allow the application of pre-processings operations like noise reduction, normalisation and alignment. [11, 12]. They also contain methods for automatic peak detection, quantification and functionalities to export the data as heat-map images. *BioIMAX* extends the standard IMS data analysis workflow with features for communication and exploration. In case of the analysis of exhaled air, experts from the fields of pneumology, chemistry and computer science need to communicate about the data, in order to discuss new or unexpected features of the data. One example subject which is frequently discussed is the origination of so far unknown peaks. To start such a discussion, a selection of pre-processed heatmap-images is exported using existing tools and uploaded to *BioIMAX*. Afterwards the respective region of interest is marked with the *BioIMAX Labeler* and the chat is started. 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. Aditionally the chemist can search existing databases if substances with matching characteristics exist and tries to identify the peak.

6 Discussion

We demonstrated the potential of *Web2.0* techniques to augment collaborative and explorative tasks in IMS data analysis. With the *BioIMAX* platform we illustrated advantages of RIAs for the application in the scientific context. The key feature of such a fully web-based platform is that users only need a username and a password to get access to the platform and can directly start uploading, exploring and sharing image data from any location with collaborating

[1] The BioIMAX architecture consists of three layers, modeling a client-server-architecture as mentioned before. First the user chooses several parameters in one selected application (e.g. in the clustering tool TICAL). When submitting the job afterwards, a HTTP-POST request is sent to the web server, where it triggers a PHP script on the web server. The script executes a XML-RPC (XML Remote Procedure Call). XML-RPCs are XML documents sent via a web protocol which are parsed at the server and contain an instruction to trigger a procedure server-side. This allows us to separate the web server and compute server for both performance and security benefits.

researchers at any location. No additional software packages or libraries have to be installed except for the standard *Flash Player*, so platform independence is achieved.

BioIMAX allows the users to build up small communities by creating projects with the aim to bring together experts from different disciplines collaborating in one research project. With the *BioIMAX Labeler* the scientists are able to focus image related discussions to specific image regions in a chat-like manner, which simplifies and speeds up the communication and collaboration in scientific research projects.

Data and analysis reproducibility is another major advantage of systems like *BioIMAX*, i.e., that all datasets and results are stored in a central data repository, so discussion and exploration of specific image aspects take place on the same copy of an image preventing ambiguity and misinterpretations. We believe, that tools such as *BioIMAX* will trigger a convergence of the mental models, different users have for the same data.

The benefit of the *BioIMAX* platform is, that not only different data with varying data structures such as images, semantic annotations, descriptive data models, etc., will be integrated in one web-based infrastructure, but also collaborative analysis on the same complex IMS image data and on the same biological or analytical question. This is a crucial step towards integrative bioinformatics with respect to many bioimage applications.

BioIMAX aims not at providing a web-based LIMS (Laboratory Information Management System) or a complete data editing system and it is in its current form not designed to perform the full spectrum of IMS specific data analysis. It is rather intended to be a general platform focussing on a quick collaborative visual exploration of various types of bioimage data such as data from 2D gel electrophoresis, microscopy or MALDI imaging. *BioIMAX* supports early data interpretation tasks in IMS analysis: Is there a misalignment? How much are two spectra correlated/identical? With *BioIMAX* scientists can get a rapid exploratory overview about the image data at hand and can easily exchange data and information without a complicated and time-consuming act via a single web-based platform.

To sum up, we expect that the ongoing development of web technologies in the age of *Web2.0* will have more and more impact on scientific work in the future, especially when several scientists from different disciplines or institutes has to collaborate, and therefore closes important gaps in IMS analysis left by standard desktop tools.

Acknowledgements

The financial support of the Bundesministerium für Bildung und Forschung, the Ministerium für Wissenschaft und Forschung des Landes Nordrhein-Westfalen, the fellowship of the University of Bielefeld (Stipendium aus Rektoratsmitteln) (CL) and the fellowship of the Genome Informatics Group, Bielefeld University (CL) is gratefully acknowledged.

References

[1] Baumbach, J.I. and Eiceman, G.A.: Ion Mobility Spectrometry: Arriving On Site and Moving Beyond a Low Profile. Appl Spectrosc, 53:338A-355A, 1999.

[2] Baumbach, J.I.: Ion mobility spectrometry coupled with multi-capillary columns for metabolic profiling of human breath. J Breath Res, 3:034001, 2009.

[3] Fox, P. and Hendler, J.: Changing the Equation on Scientific Data Visualization. Science, 331:705-708, 2011.

[4] Bourne, P.E. and Vicens, Q.: Ten Simple Rules for a Successful Colaboration. PLoS Comput Biol, 3(3):e44, 2007.

[5] Shneiderman, B.: Science 2.0. Science, 319:1349-1350, 2008.

[6] Waldrop, M.M.: Science 2.0 - Great new tool, or great risk? Scientific American, 2008.

[7] *Adobe Flex*. `http://www.adobe.com/products/flex/`

[8] *MySQL*. `http://www.mysql.com`

[9] Herold, J., Loyek, C., Nattkemper, T.W.: Multivariate image mining. Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery, 1(1):2-13, 2011.

[10] Keim, D.A.: Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and Computer Graphics, 7(1):100-107, 2002.

[11] Bödeker, B., Vautz, W., Baumbach J.I.: Visualisation of MCC/IMS-data. International Journal for Ion Mobility Spectrometry, 11:77-81, 2008.

[12] Bunkowski A.: Software tool for coupling chromatographic total ion current dependencies of GC/MSD and MCC/IMS. International Journal for Ion Mobility Spectrometry, 13:169-175, 2010.