BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 13 2011, pages 1839–1845
doi:10.1093/bioinformatics/btr282

Systems biology
Advance Access publication May 6, 2011

Creating views on integrated multidomain data

Hendrik Rohn[1,*], Christian Klukas[1] and Falk Schreiber[1,2]
[1]Department of Molecular Genetics, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben and [2]Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle, Germany

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Modern data acquisition methods in biology allow the procurement of different types of data in increasing quantity, facilitating a comprehensive view of biological systems. As data are usually gathered and interpreted by separate domain scientists, it is hard to grasp multidomain properties and structures. Consequently, there is a need for the integration of biological data from different sources and of different types in one application, providing various visualization approaches.

Results: In this article, methods for the integration and visualization of multimodal biological data are presented. This is achieved based on two graphs representing the meta-relations between biological data and the measurement combinations, respectively. Both graphs are linked and serve as different views of the integrated data with navigation and exploration possibilities. Data can be combined and visualized multifariously, resulting in views of the integrated biological data.

Availability: http://vanted.ipk-gatersleben.de/hive/.

Contact: rohn@ipk-gatersleben.de

Received on 12 October, 2010; revised on April 16, 2011; accepted on May 2, 2011

1 INTRODUCTION

Biological systems are complex systems, which are investigated separately on different levels and with different biological approaches and methods. The cycle of knowledge acquisition in biology starts usually with biological experiments to investigate interesting phenomena. The generated data is used to create a model and build hypothesis, which then can be tested by further experiments. All experiments performed in this cycle result in large datasets from different data domains, such as structural or functional image-based investigations, graph-based modeling approaches, numeric measurements at different time points and spatial resolution such as cell, tissue and organ level. Such data is gathered from different groups around the world in increasing quantity and quality. There is need for an integrated view on several biological data domains to analyze manifold cause-and-effect in the systems. All data gathered from a biological system is in context of each other and therefor data integration may provide a more complete view on the system, facilitating knowledge beyond the intended results of single experiments (Ball *et al.*, 2004). It could be shown that the combination of networks and omics data is useful and leads to novel insight into data and biological systems (Gehlenborg *et al.*, 2010). We expand such approaches by considering spatial data, in order to be able to account for additional spatial systemic properties and often occurring data types.

There have been numerous database-oriented integration approaches developed (for example, Ameur *et al.*, 2006; Shah *et al.*, 2005; Sujansky, 2001; Töpel *et al.*, 2008), which partly provide world-wide integration of multidomain biological data, but many biological researchers are overwhelmed by the number of specialized databases, sheer data flood and unhandy web interfaces. Most researchers seek for broad combination of their own data together with data from selected databases and collaborators, in the best case in one application. Consequently, there is need for an approach, which allows researchers to access and handle their data on their own computer and import data from different sources and collaborators. Complex data and several data types shall be combined and visualized in many ways, not only providing overview and insight, but also publication-ready illustrations. On the one hand, there are quite a number of applications available integrating data on the database level, such as Köhler *et al.* (2006), local applications combining networks and experimental data [for a list see Gehlenborg *et al.* (2010)] and 2D/3D data integration tools (Abramoff *et al.*, 2004; Hjornevik *et al.*, 2007; Moore *et al.*, 2007; Stalling *et al.*, 2005). On the other hand, flexible integration of more than two different data types in one application, running locally on standard computers, is sparse. Therefore, we propose a method, which formalizes and supports the biologists workflow from the measurements to a proper visualization of the data (Fig. 1), together with navigation and exploration possibilities, allowing researchers to reproduce and revisit experimental results.

There are two main requirements: first, it should be possible to integrate data from different experiments, realizing the multipersonal, multimethodological nature of biological data acquisition. Navigation, filtering and exploration possibilities will help to understand how the experiments are structured. Second, it should be possible to combine and visualize data resulting in integrated views on data from different sources and types. Visualization and data mapping approaches will help to understand what the data represent in the real system. In this article, we will focus mainly on methodological and theoretical aspects of the integration and visualization of multidomain data in an application.

The structure of this article is as follows: we propose a method to integrate multimodal data at the metadata (MetadataGraph) and measurement (MappingGraph) level and show how this allows to combine, analyze and navigate through the data by user interaction. After this, IntegrationViews to visualize combined data are explained and an application to barley data is described. Finally,

[*]To whom correspondence should be addressed.

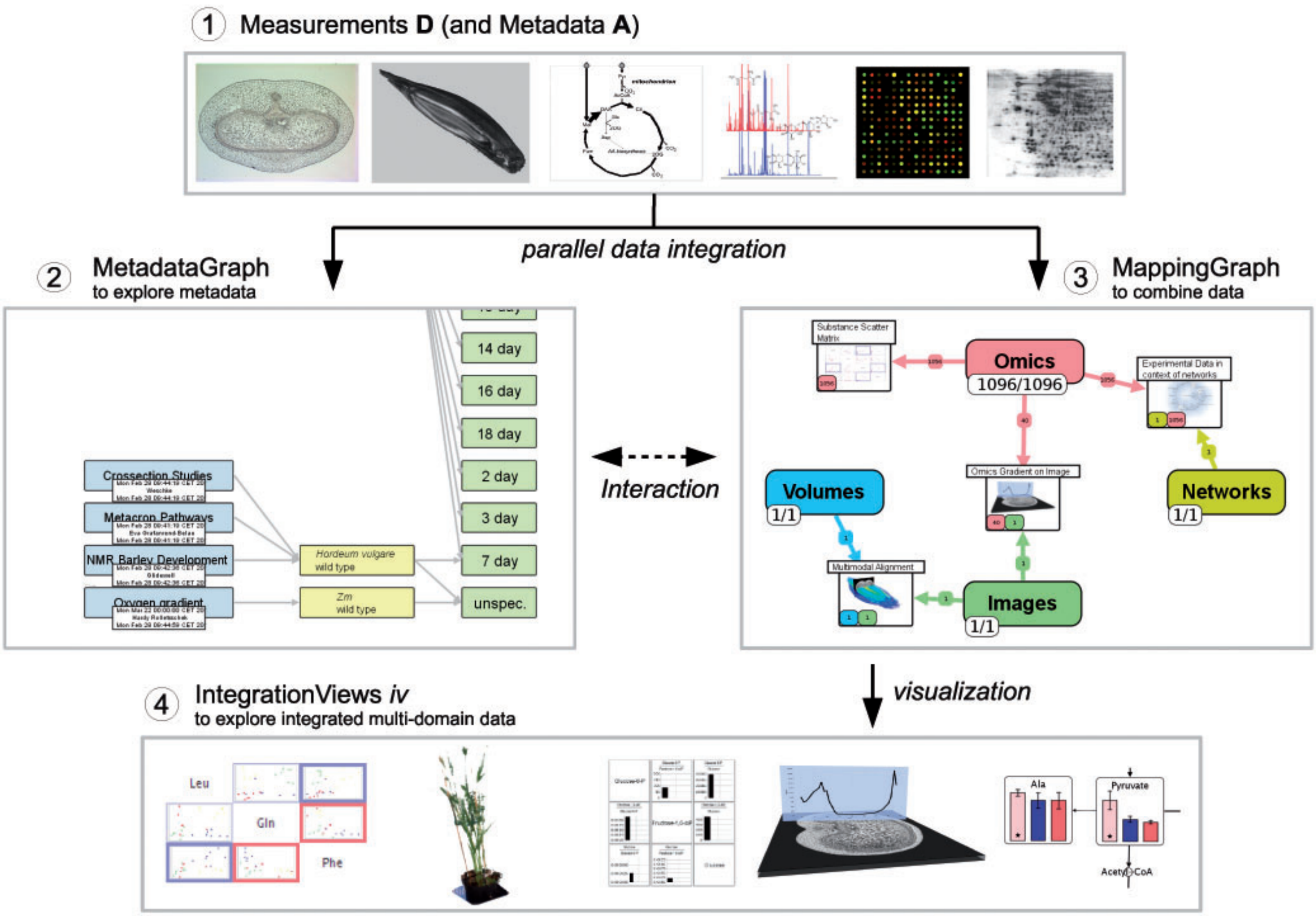


Fig. 1. Basic workflow of combining multimodal biological data. Step 1) starting with a number of measurements, which are annotated with metadata, the data get integrated into the MetadataGraph, which allows to visualize and explore the relationship between the data on the metadata level (Step 2). Step 3: All measurements from different data domains can be combined together into 'mappings' using the MappingGraph. Mappings can be visualized, allowing to create views on integrated biological data. Step 4: These IntegrationViews promote understanding of complex biological system and also serve as publication-ready images.

we discuss the implementation and some properties of the presented application.

2 METHODS

In order to investigate the complex biological systems, many different data acquisition approaches are used. Additionally, data come from different data domains, e.g. spaceless numeric information about the metabolome, transcriptome and proteome (so-called *omics data), spatial information of compound distribution and structural information (e.g. pathways or bidirectional protein interactions).

2.1 Data model for measurements and metadata

The structure of biological multidomain data can be expressed in a small data model, described in detail in Rohn *et al.* (2009). It models four types of measurements (exemplary data are shown in Step 1 in Fig. 1) and metadata annotating these measurements. To formalize this model, we define the set of measurements as $\mathbb{D}=\{d|d\in\mathbb{O}\cup\mathbb{I}\cup\mathbb{V}\cup\mathbb{N}\}$, with $\mathbb{O}$ = set of numeric measurements, $\mathbb{I}$ = set of images, $\mathbb{V}$ = set of volumes and $\mathbb{N}$ = set of networks. The metadata (defined as set $\mathbb{A}$) is used to enable search and navigation through all measurements and provides information about the experiment, contained conditions, taken samples, and measured substances (compare also Fig. 2). We define an object path $op(d)$ with $d\in\mathbb{D}$ as the set of metadata for a certain measurement object, see also Figure 2.

2.2 Combination of data on metadata level

Based on the metadata $\mathbb{A}$, the measurements $\mathbb{D}$ can be integrated in two attributed graph structures called MetadataGraph (Fig. 2) and MappingGraph (which will be described in detail in Section 2.3). The MetadataGraph combines data from several experiments on the metadata level and allows an integrated view on the available metadata.

As the cardinalities of the data model indicate, each experiment may be connected to several condition objects. Analogue, each condition is connected to several samples. Samples again are connected to a number of measurement objects. The measurement objects are also connected to substance objects, indicating the measured substance, e.g. an image may represent the 2D distribution of a metabolite. For conditions and samples, equal objects have equal attribute values. For substances or experiments, equal objects have equal 'name' attributes.

To build up the MetadataGraph, we iteratively add representative nodes for all objects to the graph. First, one experiment node is added. Then all condition nodes of this experiment and edges connecting the experiment node

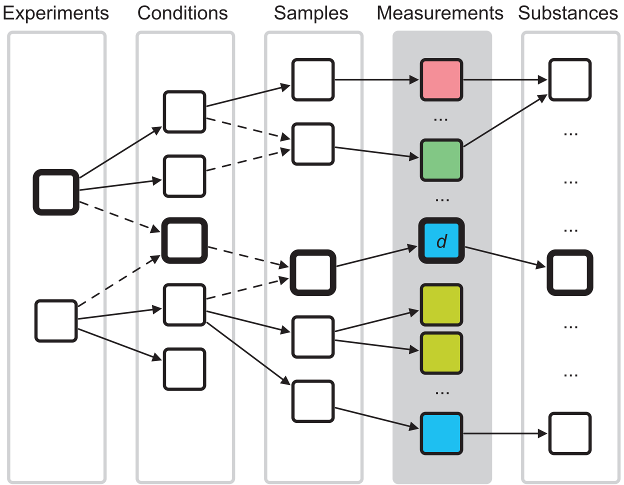


Fig. 2. Conceptional MetadataGraph structure (see also Step 2 in Fig. 1 for an instance). The MetadataGraph consists of five layers: four layers representing the metadata $\mathbb{A}$ and one layer representing the measurements $\mathbb{D}$ (gray background). The MetadataGraph allows to visualize the relationships between biological data on metadata level and supports navigation through and exploration of the data by selection and filtering operations. Dashed edges indicate a merging operation and highlighted nodes depict one object path $op(d)$ (defined in Section 2.1). Measurement nodes are colored according to the measurement types.

with each condition node are added, indicating the predecessor/successor relation and allowing one to visually track dependencies. In the following, successor sample nodes for each condition are created, together with edges between the condition and sample nodes. Subsequently, for each measurement object a node and edge from the sample node to this node is built. If the measurement object is connected to a substance object, a substance node is created and an edge to the data object node is drawn. This procedure is repeated for all experiments. Every time a node will be added, all nodes representing the same class are checked for equality by the new object node. If there is already a node representing an object with equal attribute values as the new object, no new node is created but the existing node is taken. Then all successor nodes of the new object are added to this existing node. The procedure is performed on all but the measurement object nodes, as they lack any equals operator.

Due to the merging operation, the resulting graph is a directed acyclic graph (Fig. 2), representing all integrated measurements at any time and built up using the metadata of all integrated measurements. As the set of measurements may increase over time, the graph grows accordingly. Additional data enter the graph by sequential user imports or combination of data resulting in new one (see Section 2.3.3). To prevent cluttering because of the increasing number of nodes created over time, the measurement object and substance object nodes are not shown in a view of the MetadataGraph. An instance of such a MetadataGraph can be seen in Step 2 in Figure 1. The tree-like layout was chosen, because it is a simple and convenient layout for the underlying data structure. As the graph structure depicts the relations between metadata objects of respective measurements, it allows to visually inspect the metadata and supports search functionality.

2.3 Combination of data on measurement level

Analogue to the MetadataGraph described in the previous section, we define and formalize the combination of the integrated data on the measurement level on the basis of MappingGraph.

2.3.1 Mappings Let $\mathcal{P}^{+}(\mathbb{D})$ be the non-empty power set of the measurement set. Let $op(\mathbb{D})$ be the set of metadata of all measurement objects in $\mathbb{D}$. Let $\mathbb{T}$ be any set of graph element attributes. Then a mapping $m \in \mathbb{M}$ (with $\mathbb{M}$ as the set of all mappings) is defined as the tuple

$$m=(\mathcal{P}^{+}(\mathbb{D}), op(\mathcal{P}^{+}(\mathbb{D})), \mathbb{T}).$$

By this definition, a mapping m represents a combination of measurements together with its metadata and a set of attributes. A mapping is realized as an attributed graph, in which nodes represent data objects. The nodes' attributes are set according to the object attribute values and eventually to mapping-specific attribute values of the set $\mathbb{T}$. These attributes might describe the position or size in 2D/3D space, coloring or other. Edges have mapping-dependent meanings, for example, in the 'Linked Pathway Integration' mapping (Table 1): this mapping is represented by an overview graph with a node for each pathway and edges represent substances appearing in two different pathways [overview graphs are defined and explained in Klukas and Schreiber (2007)]. Alternatively, for a mapping describing an assembled anatomic body from several 3D volumes, edges may represent connected bodily parts.

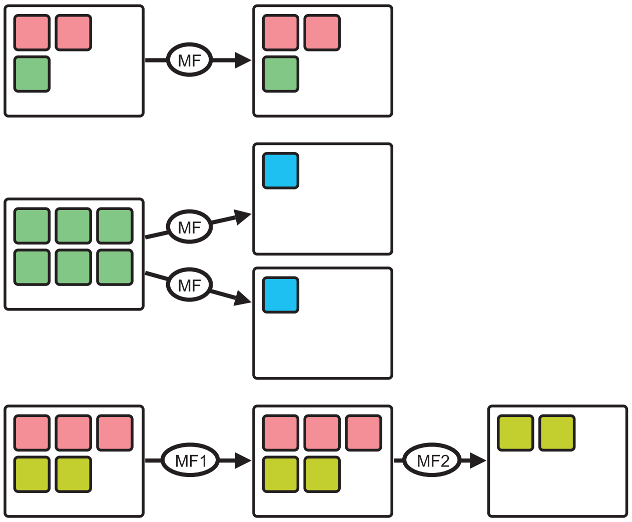


Fig. 3. Conceptual overview of mapping and remapping in the MappingGraph. Mappings (white rectangles) contain different measurements (small rectangles, color represent measurement types) and may be used by a mapping function (edge with label 'MF') to create new mappings. Mapping functions may be executed several times in order to investigate different parameters. Complex use cases can be realized by remapping the output of mapping functions. Note that mapping functions can change the measurement/data type, e.g. a set of images to one volume.

Mapping functions are able to create mappings based on the data of other mappings:

$$map: \mathcal{P}^{+}(\mathbb{M}) \mapsto \mathbb{M}.$$

In other words, a mapping function is a projection from at least one mapping to another mapping and, consequently, mappings can be mapped again (see next section). Examples of mapping functions are listed in Table 1. During the mapping procedure, new measurements may be created, which will result in an implicit update of the MetadataGraph by the new measurements' metadata.

2.3.2 MappingGraph To be able to create, visualize, analyze and select mappings, a simple attributed graph called MappingGraph was defined in Rohn *et al.* (2009). Here, the definition of the MappingGraph is extended, which now represents all integrated data on the measurement level. The four initial nodes of the MappingGraph integrate all imported data $d \in \mathbb{D}$ and are called measurement import (MI) nodes. If metadata and measurements are integrated into the MetadataGraph, the measurements are also integrated into the MappingGraph. Additional nodes of the MappingGraph can be created by the user and represent mappings (so-called mapping nodes). Edges in the MappingGraph represent the measurement flow. Edges point from a node, which was used as the measurement source, to the mapping node, which's mapping combines this measurements. Consequently, MI nodes have only outgoing edges, as they serve only as sources of measurements for possible mapping functions. It is possible to create a mapping under use of any number of mapping nodes and MI nodes, but at least one edge has to be created (which implies, that a mapping function needs at least one measurement $d \in \mathbb{D}$ as input). Therefore, edges represent the work flow/information flow when creating mappings and the resulting data are used again as input for a mapping function. This directed and successive application of mapping functions as functional modules enables users to realize complex workflows and use cases (Fig. 3). It is, for example, possible to apply a mapping function several times to the same mappings using different parameter sets, in order to compare the resulting mappings. The 'Volumetric Reconstruction' mapping may be, for example, tested according to the number of rotation images needed, in order to generate an sufficient or appealing reconstruction.

2.3.3 Generation of mappings Mappings are generated by the user selecting a number of nodes in the MappingGraph. A temporary set of measurements $\mathbb{D}_{\text{sel}}$ is generated from this nodes by extracting all measurements, which are part of the nodes' mapping. The full list of mapping functions is reduced by the mapping functions, which are not able to map at least a non-empty subset of $\mathbb{D}_{\text{sel}}$. The user chooses one mapping function, which takes a subset of the extracted measurements as an input and creates a new mapping, represented by a new mapping node in the MappingGraph. If the new mapping contains new measurements created by the mapping function, the MetadataGraph is updated with the measurements' metadata.

2.4 IntegrationViews

IntegrationViews *iv* are views on combined measurements and are defined as a tuple $(\mathbb{M}, \mathbb{R}^i)$, with $i \in \{1,2,3\}$ and $\mathbb{R}$ = set of real numbers. A visualization function

$$vis: \mathbb{M} \mapsto \mathbb{R}^i, i \in \{1,2,3\}$$

projects mappings in the Euclidean space. This means, an IntegrationView is a visualization of a mapping $m \in \mathbb{M}$, generated by the visualization function *vis*. The function uses the graph of the mapping *m* as input. A user creates an IntegrationView by selecting a mapping node in the MappingGraph and choosing a visualization function, which is able to handle this kind of mapping. There are several visualizations possible for one mapping, e.g. the IntegrationView 'Network Stacking' (a stacking of several networks) can be drawn as graphs in a 2D plane (nodes are rectangles, edges are lines) or in 3D space (nodes are cuboids, edges are cylinders). An 'Image Stacking' can be displayed as ordinary 2D drawings or planes lying in 3D space. Table 1 shows a number of possible IntegrationViews, of which three examples are explained in the context of barley data in detail in Section 3. Most of the presented IntegrationViews combine only one or two types of data, but combining more than two types of measurements is also possible. As this can be quite complex, usually the IntegrationViews depend on the requirements of biologists for special problems and are often use-case oriented.

IntegrationViews realize both interactive and explorative work with combined multimodal biological data and publication-ready illustrations. Some of them were already developed and implemented in other systems, and also novel IntegrationViews have been developed. Our approach allows to combine the previously separated combination of data together in one application and in one workflow, supported by navigation and exploration possibilities.

3 APPLICATION EXAMPLE

Barley is an important crop plant for food and feeding. Consequently, several interdisciplinary projects are investigating the metabolism from a system-wide perspective, in order to find targets to promote resistance and increased yield. Different data acquisition methods result in multidomain data, which has to be integrated and visualized manifold. We tackle these requirements by applying our approach to a set of barley grain data and demonstrate how this data of one object but of different sources and modalities may be exemplarily brought into context of each other using the described approach. The dataset includes a gradient of oxygen concentration, several measurements of metabolites of primary metabolism, topology-based metabolic pathway information, a cross-section image and an NMR volume imaging the water distribution of barley grain.

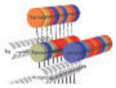
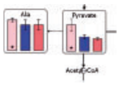
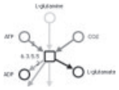
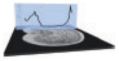
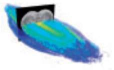
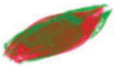
The oxygen gradient describes the oxygen concentration across the barley grain, which was measured by a probe moving through the tissue [compare Rolletschek *et al.* (2004)]. We represent the gradient as numeric measurements with a position attribute. It is useful to show this gradient in the context of the surrounding grain structure [see Fig. 4 in Rolletschek *et al.* (2004)], but as can be seen in the respective hand-made illustration, due to the 2D approach it is not really clear where the probe has been moved through and what the distributed points indicate. Therefore, we decided to present the gradient on top of the cross-section in 3D. The mapping function takes the numeric measurements and the cross-section as input. The gradient is visualized by a chart and is placed orthogonal to the cross-section in such a way, that it cuts the cross-section image along the line the probe was moved. The position and rotation is set by the user according to the biological experiment. The resulting IntegrationView can be seen in Figure 4Ca.

Other metabolite measurements were also taken from barley at different timepoints (Borisjuk *et al.*, 2005). These data include sugars and amino acids and is modeled as numeric measurements with different samples (representing different time points). To understand the behavior of metabolite concentration changes during barley development at a systemic level, it is useful to bring these time-course data into context of the metabolic pathway structure. This allows one to visually track dependencies between metabolites in different parts of the pathway. The mapping function takes as input numeric measurements and pathways of the primary metabolism of barley from the MetaCrop database (Grafahrend-Belau *et al.*, 2008). As pathway nodes represent single substances, the mapping function integrates all numeric measurements of one substance in the corresponding node of the pathways. A 2D graph visualization function paints the pathway as usual, but shows the integrated numeric measurements of nodes as diagrams painted onto the node. In our case, these diagrams visualize the time-course of one substance. The resulting IntegrationView can be seen in Figure 4Cb (only the TCA cycle pathway is shown). It is easy to image a new mapping based on this mapping, by mapping similar time-course measurements of different species to species-specific pathways and stack these data-enriched pathways in 3D, allowing visual comparison of different species data. This would be an example of the remapping procedure described in Section 2.3.2.

Biologists are often interested in substances, which behave similar for different conditions or time points. As the metabolite measurements represent quantitative concentration values, scatter plots are widely used to visually observe potentially correlated substance. Therefore, scatter plots are created from the set of metabolic measurements. The mapping function takes all metabolite measurements as input and builds up a pairwise substance measurement matrix. One element of the matrix contains all measurements of the two substances. The visualization function displays each element of the matrix in a well-known scatter plot visualization, by plotting points for pairwise measurement values. If the pairwise measurements of the matrix element are part of different conditions, the points will be colored differently. The resulting IntegrationView can be seen in Figure 4Cc. In this figure, only those substances are shown, which were integrated into the nodes of the TCA cycle in Figure 4Cb.

Finally, structural information such as the cross-section may be combined with functional information, allowing to compare quantitative information in the context of the grain structure. Therefore, NMR imaging experiments were carried out (Glidewell, 2006) to obtain information about the water content, which is an important indicator of grain maturity (Bewley and Black, 1994,

Table 1. Overview of some implemented IntegrationViews

IntegrationView		Involved datasets	Description
	Image Stacking	$\mathbb{I}^{(k)} \rightarrow \mathbb{I}^{(k)}$, $k \geq 2$	Several images are aligned according to their position in the biological object.
	Network Stacking	$\mathbb{N}^{(k)} \rightarrow \mathbb{N}^{(k)}$, $k \geq 2$	Several networks are aligned based on the nodes labels in 3D space allowing visual comparison of network structure (Brandes *et al.*, 2004).
	Omics Network Context	$\mathbb{N}^{(k)} \times \mathbb{O}^{(j)} \rightarrow \mathbb{N}^{(k)} \times \mathbb{O}^{(j)}$, $k \geq 0, j \geq 1$	Numeric measurements as diagrams in their biological context, such as metabolic or gene regulatory networks (Junker *et al.*, 2006), also described in Section 3.
	Linked Pathway Integration	$\mathbb{N}^{(k)} \rightarrow \mathbb{N}^{(k+1)}$, $k \geq 2$	Generates an overview-graph (Klukas and Schreiber, 2007), where each node represents a network, allowing navigation through all linked networks.
	Network Comparison	$\mathbb{N}^{(k)} \rightarrow \mathbb{N}^{(1)}$, $k \geq 2$	Generates a new graph, consisting of all graph elements of the input graphs, which are colored according to the frequent occurrence in the input graphs.
	Condition Log Ratio	$\mathbb{O}^{(k)} \rightarrow \mathbb{O}^{(k)}$, $k \geq 2$, conditions $=2$	Compares a condition with another condition according to their substance concentrations as $\log_2$ ratio.
	Substance Ratio Matrix	$\mathbb{O}^{(k)} \rightarrow \mathbb{O}^{(k)}$, $k \geq 2$, substances ≥ 2	Creates a matrix displaying the pairwise ratio of substances as bar charts.
	Substance Scatter Matrix	$\mathbb{O}^{(k)} \rightarrow \mathbb{O}^{(k)}$, $k \geq 2$, substances ≥ 2	Correlates numeric measurements of pairwise substances in a scatter plot matrix, also described in Section 3.
	Gradient on Image	$\mathbb{I}^{(1)} \times \mathbb{O}^{(k)} \rightarrow \mathbb{I}^{(1)} \times \mathbb{O}^{(k)}$, $k \geq 2$, substances $=1$, $o \in \mathbb{O}$ has position attribute	Visualizes a gradient in context of structural properties, also described in Section 3.
	Multimodal Alignment	$\mathbb{V}^{(1)} \times \mathbb{I}^{(k)} \rightarrow \mathbb{V}^{(1)} \times \mathbb{I}^{(k)}$, $k \geq 1$	Registers 2D images into 3D volumes to be able to compare information of different resolution and methods (Scharfe *et al.*, 2010), also described in Section 3.
	Volume Registration	$\mathbb{V}^{(k)} \rightarrow \mathbb{V}^{(k)}, k \geq 2$	Registers two or more 3D volumes onto each other to be able to compare differences in substance distribution or bring functional information into structural context.
	Volumetric Reconstruction	$\mathbb{I}^{(k)} \rightarrow \mathbb{V}^{(1)}$, $k \geq 1$	Reconstructs the 3D structure of an object from rotation images using space carving similar to Kutulakos and Seitz (2000).

IntegrationViews represent combined and visualized multimodal biological data, which provide insight into the data and can be used for publication purposes. Several of the listed IntegrationViews are well-known data visualizations, but also new combinations appeared due to our generalizing approach. All IntegrationViews have been further formalized based on the theoretical foundations in this article. The second column describes the involved input and result measurements of the mapping function, where $\mathbb{N}^{(k)}$ represent a set of k networks, $\mathbb{V}^{(k)}$ represent a set of k volumes and so on.

pp. 136–138). It is possible to combine both spatial information, volumetric and image based, by using a multimodal alignment algorithm: the mapping function takes the volume and the cross-section as input. The cross-section image is positioned in the volume according to the position in the real grain by the user via rotation and translation of both models. To support the visual analysis, the 3D view additionally allows to modify the view (zooming, panning, rotating, etc.) and models (transparency, cutting and clipping planes, color maps, etc.). It is easy to imagine a new mapping, which bases on this mapping and additionally maps a gradient as shown in Figure 4Cd.

The described IntegrationViews realize different views on the integrated barley data of different data domains, promoting understanding and linking knowledge from different experiments. Using this approach, the data can be processed faster and visualized in a more intuitive way. The IntegrationViews may also be part of the publication process, as the complete project may be exported and interactively explored by reviewers and interested scientific staff using our application. As indicated, even more IntegrationViews could be generated from this example dataset. Continuative genomic information and more spatial data can be included, allowing to investigate tissue- or compartment-specific properties during development of barley grain.

4 IMPLEMENTATION

The approach described in the previous sections is implemented as an Add-on for the VANTED system (Junker *et al.*, 2006), called HIVE (**H**andy **I**ntegration and **V**isualisation of multimodal **E**xperimental data). It is written in Java and Java3D, utilizes the Model-View-Controller concept and other Software Design Patterns. Together with the plugin-based structure of VANTED, it is easy to extend the functionality even at run time, such as additional database access, e.g. MetaCrop (Grafahrend-Belau *et al.*, 2008) and DBE (Borisjuk *et al.*, 2005), simulation capabilities, e.g. flux balance analysis (Kauffman *et al.*, 2003), mapping functions and visualization algorithms. The Add-on is available at http://vanted.ipk-gatersleben.de/hive/ and is released under the GPL 2.0 License.

Every type of measurement is assigned to a color (see for example, Fig. 4: color of edges in the MappingGraph). This color code is used throughout the application to provide hints about the data domains tackled. For example, when selecting measurements to be mapped, the description of the measurements is color coded according to the measurement type. The list of available mapping functions (see Tab at the right side of Fig. 4) is also color coded according to the type of measurements, which are needed as input for the mapping functions. Mapping nodes give a hint about the measurements, which are part of the mapping. This is implemented by showing a number of small rectangles colored according to the type of measurements and with a number describing the size of the measurement set of this type (see Fig. 4B and Step 3 in Fig. 1).

To be able to keep an overview of all views and interaction possibilities, the application window layout is divided into three parts (Fig. 4): both views representing the MetadataGraph or MappingGraph, respectively, are aligned stacked at the left part of the application window, which allows the user to quickly comprehend the integration and combination workflow at any time. The middle part provides space for the IntegrationViews and the right part situates the controls working on the active view. Although it is a well-arranged layout, the user is able to maximize, minimize and move the views at own will. During work with these views, several optical indicators help the users to track their actions. For each IntegrationView shown to the user, the mapping node in the MappingGraph will be highlighted. Simultaneously, for all measurements d, which are part of the mapping, the object nodes of $op(d)$ will be highlighted in the MetadataGraph. This allows a user to quickly comprehend the tackled metadata and measurements in context of the Metadata- and MappingGraph. As this highlighting of object nodes links the IntegrationViews to the Metadata- and MappingGraph, it is also possible to use this link backwards, such as selecting highlighted object or mapping nodes and navigate to the corresponding IntegrationViews.

5 DISCUSSION

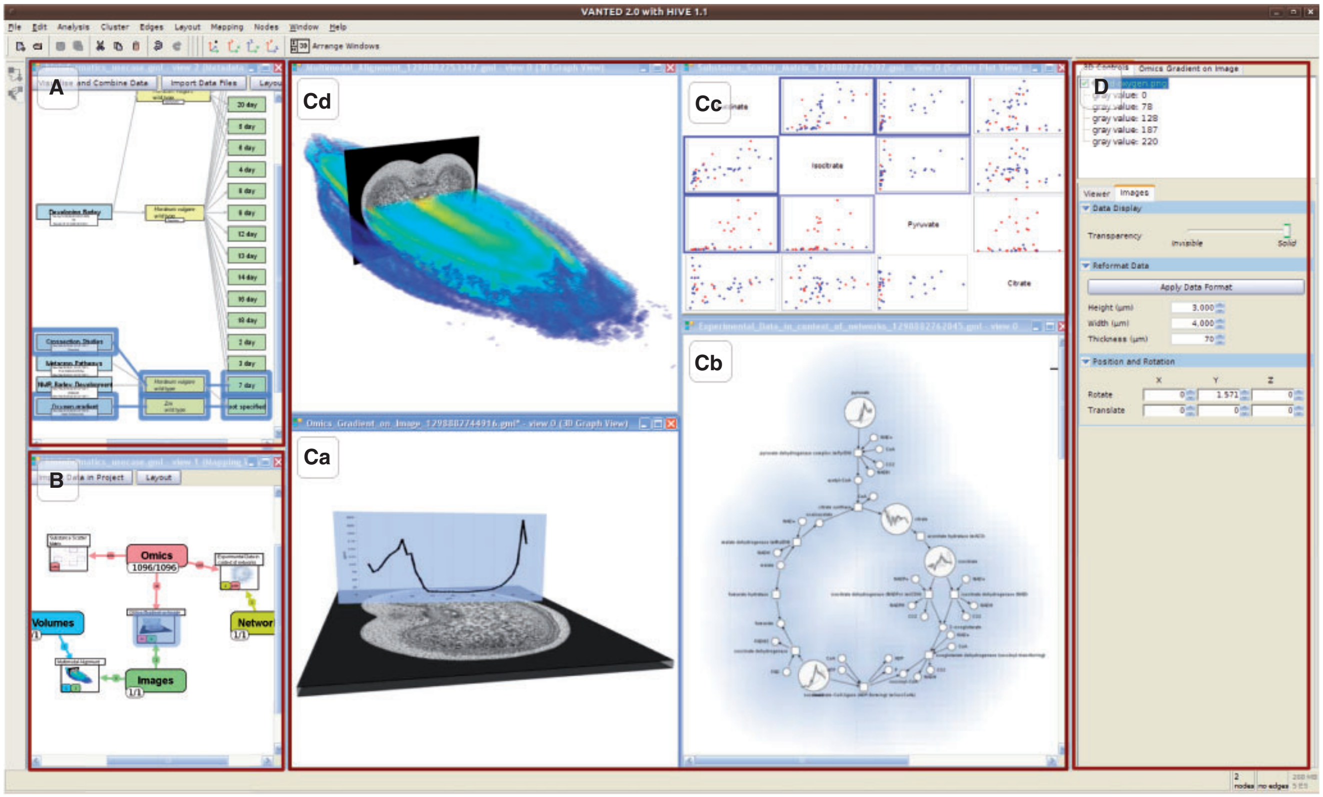
We presented the theoretical and methodological foundations of an application supporting the integration and visualization of biological multidomain data. Several users with biological and computer science background had time to get used to the workflow and work with a prepared project and example data. After this, personal discussions with the scientists revealed that after a short period of vocational adjustment,. the complex application workflow is logical and productive, despite the unfamiliar graph-based interface, which is similar to the pool window in the AMIRA application (Stalling *et al.*, 2005). They rated it as a good approach capable of realizing complex use cases given by biological researchers and supporting them in their daily work. At the moment, our scientific partners do not possess comprehensive datasets consisting of all the different data types from the same species. We will evaluate our integration approach when such datasets emerge in more detail.

The scalability of the application is good with the currently considered datasets. Often biological experiments consists of 100–1000 data records, and even tens of such experiments can be handled with standard PCs at any time. As our approach supports measurement working sets (by hiding object nodes in the MetadataGraph), researchers are able to temporarily hide parts of the experimental data they are not interested in, while performing specific tasks. Prospective increase of data quantity can be faced with optimization of the performance and memory consumption of the underlying framework VANTED, improved data storage and increased screen size. Management and persistent storing of hundred of experiments is not intended by our approach, but can be realized by connecting to the DBE database (Borisjuk *et al.*, 2005) using the respective VANTED Add-on. The network sizes are limited by the underlying framework and work up to 50 000 nodes. The implementation of the 3D view provides a variety of interaction and model alteration possibilities, but needs especially for high-resolution volumetric data (>50 million voxels) much memory at the moment. Improvements, such as swapping the data to the graphics card, may be included, but would end in reduced interaction possibilities [similar to Schmid *et al.* (2010)].

6 CONCLUSION

We described a methodology and an application, which is able to bring together data from different data domains. The integration is realized by the MetadataGraph and the MappingGraph, focusing on different aspects of the data integration. Both graphs support user interaction, allowing to navigate through the data and realizing a workflow from the imported data to combined and visualized data. The application is able to bring several biological data domains together in a formalized way, guiding the user to explore, understand and visualize the data. This approach can also be useful for researchers of different fields of science, such as theoretical biology, systems biology and medicine.

Further extensions are to link the application to other tools, in order to realize complex mapping functions by utilizing problem-specific solutions. We will develop more mapping functions and extent the set of visualization functions, to be able to provide a broader range of biological data combinations and visualizations. This includes working with different domain scientists in order to integrate data of different sources and modalities together. By doing



Height (µm) 3,000
Width (µm) 4,000
Thickness (µm) 70

	X	Y	Z
Rotate	0	1.571	0
Translate	0	0	0

Fig. 4. Screenshot of the application HIVE, which allows to integrate, combine and visualize data from different biological domains, resulting in views on integrated data. (**A**) MetadataGraph integrating data at metadata level; (**B**) MappingGraph integrating data at measurement level, (**C**) IntegrationViews (**a**) 'Gradient on Image', (**b**) 'Omics Network Context', (**c**) 'Scatter Plot Matrix' and (**d**) 'Multimodal Alignment' of the application example in Section 3; (**D**) Tab with controls working on the active view.

this, the application example can be extended and the approach also validated for other species, such as human or bacterial data.

Funding: BMBF grant (0315044A) (partly).

Conflict of Interest: none declared.

REFERENCES

Abramoff,M. *et al.* (2004) Image processing with ImageJ. *Biophotonics Int.*, **11**, 36–42.

Ameur,A. *et al.* (2006) The LCB data warehouse. *Bioinformatics*, **22**, 1024–1026.

Ball,C.A. *et al.* (2004) Funding high-throughput data sharing. *Nat. Biotechnol.*, **22**, 1179–1183.

Bewley,J.D. and Black,M. (1994) *Seeds - Physiology of Development and Germination*. Plenum Press, New York, London.

Borisjuk,L. *et al.* (2005) Integrating data from biological experiments into metabolic networks with the DBE information system. *In Silico Biol.*, **5**, 93–102.

Brandes,U. *et al.* (2004) Visual triangulation of network-based phylogenetic trees. In *Proceedings of Joint Eurographics - IEEE TCVG Symposium on Visualization*, vol. 2912 of *Lecture Notes in Computer Science*, pp. 75–84.

Gehlenborg,N. *et al.* (2010) Visualization of omics data for systems biology. *Nat. Methods*, **7**, S56–S68.

Glidewell,S.M. (2006) NMR imaging of developing barley grains. *J. Cereal Sci.*, **43**, 70–78.

Grafahrend-Belau,E. *et al.* (2008) Metacrop - a detailed database of crop plant metabolism. *Nucleic Acids Res.*, **36**, D954–D958.

Hjornevik,T. *et al.* (2007) Three-dimensional atlas system for mouse and rat brain imaging data. *Front. Neuroinformat.*, **1**, 1–12.

Junker,B.H. *et al.* (2006) VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, **7**, 109.

Kauffman,K.J. *et al.* (2003) Advances in flux balance analysis. *Curr. Opin. Biotechnol.*, **14**, 491–496.

Klukas,C. and Schreiber,F. (2007) Dynamic exploration and editing of KEGG pathway diagrams. *Bioinformatics*, **23**, 344–350.

Köhler,J. *et al.* (2006) Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, **22**, 1383–1390.

Kutulakos,K.N. and Seitz,S.M. (2000) A theory of shape by space carving. *Int. J. Comput. Vis.*, **38**, 199–218.

Moore,E.B. *et al.* (2007) Mindseer: a portable and extensible tool for visualization of structural and functional neuroimaging data. *BMC Bioinformatics*, **8**, 389.

Rohn,H. *et al.* (2009) Integration and visualisation of multimodal biological data. In *Proceedings of the German Conference on Bioinformatics*, Vol. 157 of *Lecture Notes in Informatics*, pp. 105–115.

Rolletschek,H. *et al.* (2004) Energy state and its control on seed development: starch accumulation is associated with high ATP and steep oxygen gradients within barley grains. *J. Exp. Bot.*, **55**, 1351–1359.

Scharfe,M. *et al.* (2010) Fast multi-core based multimodal registration of 2D cross-sections and 3D datasets. *BMC Bioinformatics*, **11**, 20.

Schmid,B. *et al.* (2010) A high-level 3D visualization API for java and ImageJ. *BMC Bioninformatics*, **11**, 274.

Shah,S. *et al.* (2005) Atlas - a data warehouse for integrative bioinformatics. *BMC Bioinformatics*, **6**, 34.

Stalling,D. *et al.* (2005) *Amira: A Highly Interactive System for Visual Data Analysis*, chapter 38. Academic Press, Inc. Orlando, FL, USA, pp. 749–767.

Sujansky,W. (2001) Heterogeneous database integration in biomedicine. *J. Biomed. Informat.*, **34**, 285–298.

Töpel,T. *et al.* (2008) BioDWH: a data warehouse kit for life science data integration. *J. Integr. Bioinform.*, **5**, 93.