Integration and Visualisation of Multimodal Biological Data

Hendrik Rohn[1], Christian Klukas[1], Falk Schreiber[1,2]

[1] Leibniz Institute of Plant Genetics and Crop Plant Research Gatersleben, Germany
[2] Martin Luther University Halle-Wittenberg, Germany
{rohn,klukas,schreibe}@ipk-gatersleben.de

Abstract: Understanding complex biological systems requires data from manifold biological levels. Often this data is analysed in some meaningful context, for example, by integrating it into biological networks. However, spatial data given as 2D images or 3D volumes is commonly not taken into consideration and analysed separately. Here we present a new approach to integrate and analyse complex multimodal biological data in space and time. We present a data structure to manage this kind of data and discuss application examples for different data integration scenarios.

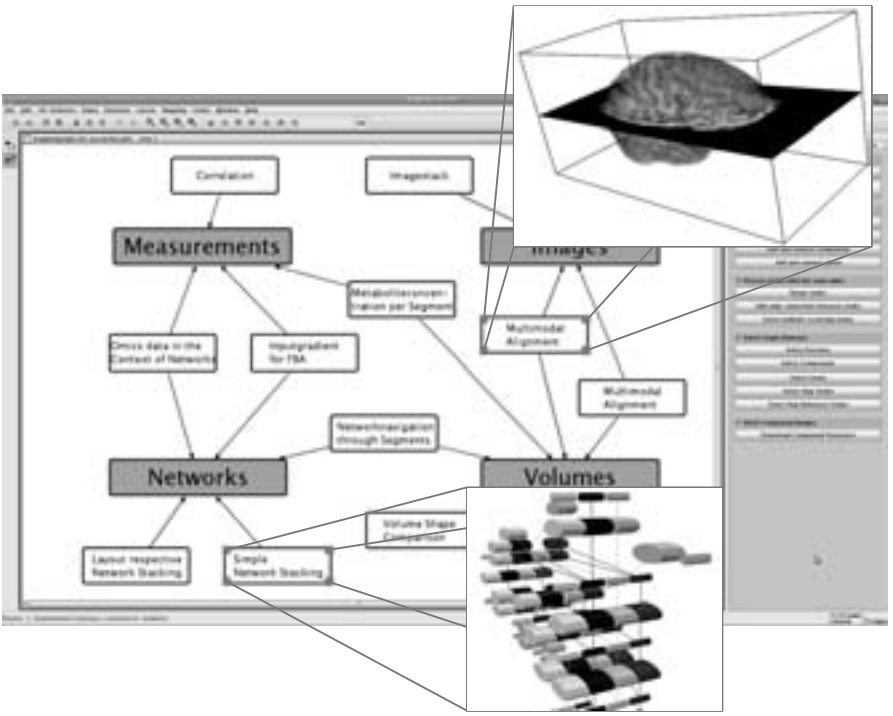

Figure 1: Preview of a prototypic system which integrates, analyses and visualises multimodal biological data based on a mapping graph.

1 Background

Modern life science researchers are able to acquire massive data by using high-throughput techniques. This leads to the accumulation of data from gene and protein activity, protein interaction and metabolite concentration, usually called -omics data. Additionally manifold *in silico* analysis such as flux balance analysis, kinetic modelling, network-centralities and -motifs can gather new information about the intrinsic properties of biological systems. To put this data into biological context network models describing the interactions and relations between biological objects are developed, such as gene regulatory or metabolic networks. Also spatial data, such as structural and functional NMR volume data, histological cross-sections, *in situ* hybridization and surface models, are measured and obtained in increasing quantity and quality and should be considered as valuable parts of models of biological systems.

To answer biological questions often different types of data have to be integrated and considered in spatial and temporal context. Using data mapping one can bring the multimodal data into context to each other, allowing more intuitive analysis, navigation and interpretation of the data. Currently there exist some tools for integration of -omics data into the context of networks [HMWD04, JKS06, KBT+06, Kol02, SMO+03, vIKP+08]. Also some 2D and/or 3D data integration tools exist [Bar06, HLD+07, MPLB07, SWH05]. However, integration of all datatypes in one application with complex mapping possibilities is not considered. In this paper we present a novel approach combining biological -omics data, 2D data, 3D data and network models under consideration of space and time.

The structure of this paper is as follows: First, we propose a data structure to represent and integrate such diverse data types. Second, we discuss different ways of mapping and visualising the multimodal data. Last, we show some example use cases for real-world data mapping applications. Fig. 1 gives an impression of such a system, which is able to intuitively integrate multimodal biological data.

2 Modelling Biological Data

Data is gathered from different parts of a biological system with different resolution. What is the structure of the data? How do we account for the spatial and temporal dimension?

The data structure for multimodal biological data can be seen in Fig. 2. It consists of two main parts: the measured data (highlighted in blue-grey) and annotation data. There are four types of measured data: “Simple measurements” standing for single values, such as the concentration of a metabolite without any further spatial information (-omics data is usually modelled by simple measurements). “Images” represents two-dimensional data such as histological cross-sections or *in situ* hybridisations. “Volumes” denote three-dimensional data such as structural and functional NMR imaging data. “Networks” stand for structural information of biological pathways expressed as a graph. Simple measurements, images and volumes have a “replicateID” to be able to distinguish experiments carried out several times helping to obtain statistical significant results. In addition to

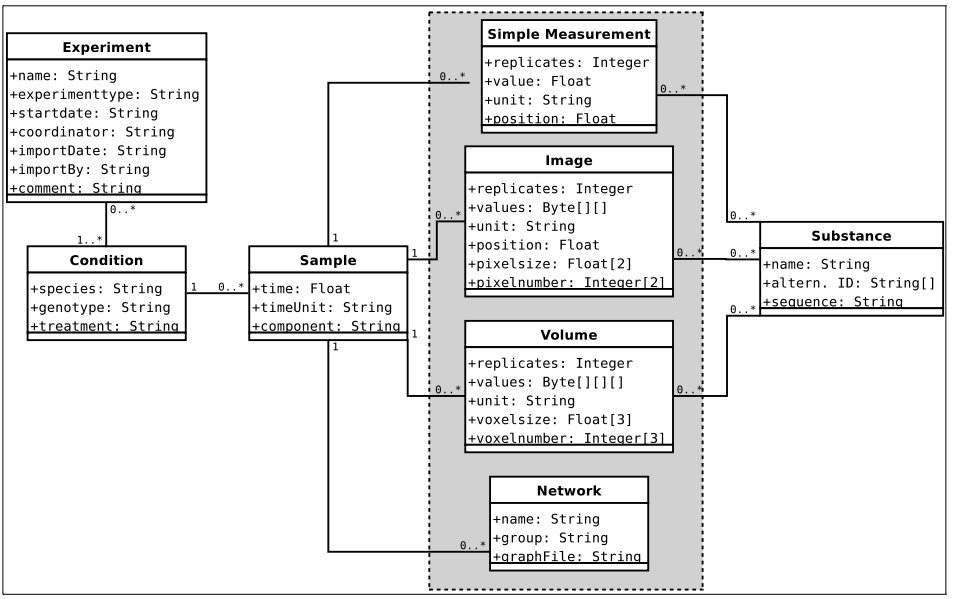

Figure 2: The model for data from experiments. Experiments are carried out under special conditions and consist of a number of samples. These include four different types of measurements: simple measurements, images, volumes and networks. Each measurement except networks may belong to a substance representing the measured biological object.

simple measurements images and volumes include information for their size in respective coordinate systems (pixel- and voxelsize and -numbers). Simple measurements and images also store position information, allowing to describe a vector of simple measurements (e.g. gradients) or images (e.g. position of image in the real biological object) in spatial context. Networks have a name and belong to a certain network group.

An biological experiment has some metadata such as name, coordinator of the experiment, date of import and who imported it. Additional information, for example, about the experiment setup, can be stored unstructured in the "comment" attribute. Each experiment has a number of conditions under which it was carried out: The name of the species is stored in the first attribute. The "genotype" attribute indicates a normal genotype or altered one (e.g. different transgenic lines). "Treatment" may be oxygen-depletion or other environmental properties. Under these conditions some samples are collected at a specific time-point, representing the temporal dimension. Measurements are also collected from a certain "component", for example, chloroplast (cell level) or brain (organ level). Each sample consists of a number of measurements, described above. All measurements but networks describe the quantity of a certain substance measured in the experiment. The substance will serve as an identifier in the data mapping, which will be described in detail in the next section. For simple measurement data the identifier is, for example, a metabolite or a protein, whereas the identifier for two-dimensional data may be the transcript measured in an *in situ* hybridization. For three-dimensional data the substance can be the metabolite the NMR image is based on, e.g., water or protons. Networks are not related to

substances because they only describe structural relations.

The proposed data model is simpler than that one used in the MIAME standard [BHQ+01] (microarray data), PEDRo database [TPG+03] (proteomics data) or ArMet framework [RJL+07] (metabolomics data). The reason is, that we do not want to model the complete experiment workflow. This would include experiment description, design and setup, normalisation methods, annotation methods, the raw and processed data, data standards and more. Instead the focus of our model is on already processed, filtered and normalised experimental data and metadata. Therefore we consider only data required for visualisation and analysis.

3 Integration of Multimodal Biological Data

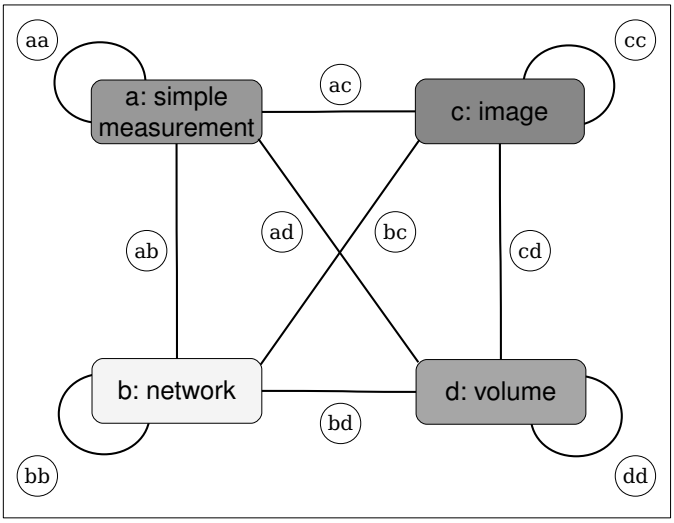

Figure 3: Mapping graph for integrating multimodal data. A node contains all biological data of one type (simple measurements, images, volumes and networks as shown in Fig. 2). An (hyper)edge represents a mapping between one, two or more types of biological data. There are several mappings possible, but for comprehensibility only one- and two-type mappings are shown.

The integration of multimodal biological data is achieved by a mapping graph, whose structure is shown in Fig. 3. The nodes represent different types of biological data, whereby the edges represent a possible mapping between these types. In the following we describe the kinds of data mapping. There are mappings between data of the same type (e.g. “aa”), mappings between data of two different types (e.g. “ac”), mappings between three types (e.g. “acd”), and mappings between all types of data (“abcd”). Note that the mapping usually allows several ways to be processed and visualised. For example, mapping one network on another could be represented as a stacking (see Fig. 4) or as one network showing the difference between both. Here we will give a typical example for some of the

mappings, but many more are possible.

aa: Mapping of simple measurements on simple measurements, for example, visualising the correlation of metabolite concentrations by scatter plots.

bb: Mapping of networks on other networks, for example, network stacking. A detailed use case can be found in Section 4.1.

cc: Mapping of images on images, for example, image stacking of cross-sections. This can be useful if several cross-sections of one object have been obtained and the images have to be placed according to their position in the real object.

dd: Mapping of volumes on volumes, which can be useful for comparing tissue shapes. Here researchers may acquire information how the shape of tissues differ for genetically altered systems.

ab: Mapping of simple measurements on networks, for example, concentration-dependent node colouring. A detailed use case can be found in Section 4.3

ac: Mapping of simple measurements on images, for example, combining high-resolution metabolite concentration data and low-resolution image data showing the concentration distribution in two dimensions.

cd: Mapping of images on volumes, for example multimodal alignment. High-resolution or special coloured cross-sections taken from a biological object are aligned into the three-dimensional representation of the object. A detailed use case can be found in Section 4.2

bd: Mapping of networks on volumes, for example, for navigation. Here segmented tissues of the volume can be used to navigate through the different networks obtained in experiments.

More complex mappings are also possible (e.g. “abc”), but depend on the requirements of life scientists and are therefore often purpose-built. By using this mapping graph the multimodal biological data, consisting of different data types, can be seamlessly combined and integrated into one system.

The data can be imported into the mapping graph using file open dialogs or drag and drop functionality. Such files can be exported from various tools and databases, e.g. KEGG, MetaCrop, AMIRA [SWH05]. Several data formats will be accepted, e.g. GML, GraphML, SBML and KGML for networks, CSV textfiles and Excel spreadsheets for simple measurements, VRML and Analyze 7.5 for volumes and PNG, JPEG and TIFF for images.

4 Use Cases

To show the functionality of the integration via a mapping graph we will highlight four exemplary use cases in detail.

4.1 Network Stacking

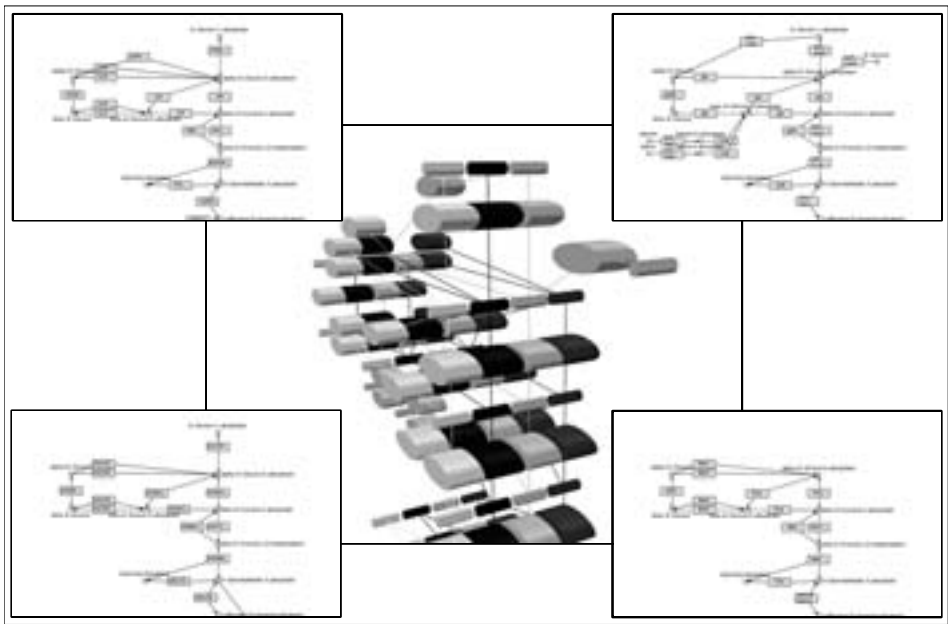

Figure 4: Use case network stacking: Four networks of glycolysis from different species are stacked in the three-dimensional space to support exploration of structural differences such as missing metabolites and interactions.

The first use case is network stacking (see Fig. 1 and 4), which is an instance of mapping case "bb" and represented in the mapping graph by an edge between networks (see Fig. 5). Here several networks are aligned allowing visual comparison of network properties: a network is mapped at one plane lying in a three-dimensional space. The next network and its plane are aligned in such a way that the corresponding nodes in the networks are stacked on top of each other respecting the layouts (see [BDS04] for further details). Additional networks can be aligned in the same way creating a $2\frac{1}{2}$D-stacking of networks. This representation allows to explore structural differences and similarities such as missing metabolites, unique interactions and conserved motifs for different species or genotypes.

4.2 Multimodal Alignment

Multimodal Alignment is a technique to align two-dimensional images into three-dimensional volumes. Often the images are high-resolution cross-sections through the biological object, allowing high detailed analysis and yield information obtained with specific methods such as *in situ* hybridisations. Volumes on the other hand represent lower-resolution three-dimensional information of an object. The idea is to combine both information,

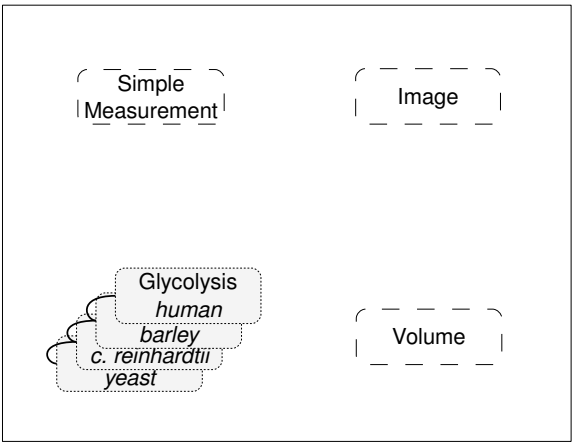

Figure 5: Instance of the mapping graph for use case network stacking. The mapping graph consists of four glycolysis networks of different species.

which means the image has to be moved to the correct position in the volume. To align the images one can use cross-correlation or other methods (e.g. some algorithms are implemented in the Insight Segmentation and Registration Toolkit [YAL02]). This means the second use case is an instance of mapping case “cd”. An example of the result of such a mapping can be seen in Fig. 1 on the first page.

4.3 Omics Data in the Context of Networks

For the analysis of biological data it is useful to apply an integrated view on the measured data and its related background information, such as metabolic pathways or regulative processes. For this purpose one can map the biological data (e.g. protein activity, metabolite concentration, etc.) to structural information such as glycolysis pathway, which represents a mapping of type “ab”. An automatic mapping of experiment data onto relevant network elements occurs if the measured data and the network nodes have common identifiers. For the visualisation of mapped data the display of multiple mapped datasets for a single network element is supported. Using line charts, bar charts and similar techniques the scientist is able to visualise more complicated datasets, such as data from different time points, experimental conditions and replicates. For further information about this mapping see [JKS06].

4.4 Oxygen Gradient and Flux Balance Analysis

The last exemplary use case consists of a mapping “ab” and can be seen in Fig. 6 and 7. At the top of Fig. 6 there is an oxygen gradient, which consists of a number of simple mea-

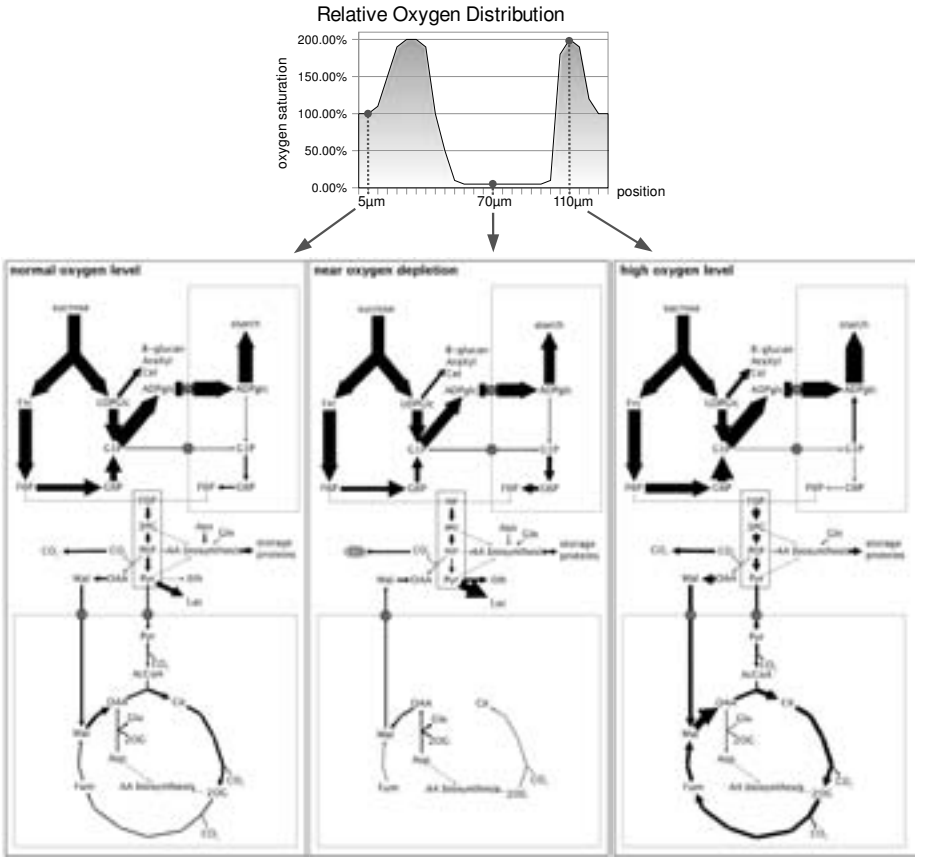

Figure 6: Use case flux balance analysis: One-dimensional oxygen gradient used as an input for flux balance analysis [KPE03]. The simulation results for different oxygen levels are mapped to the glycolysis network. The visualisation of the data shows, that the higher the oxygen concentration the higher the starch accumulating flux (middle, left, right).

surements. Such gradients could be obtained as time series or by a probe moving through a tissue and measuring the relative oxygen level for different positions (see [RWW+04]). The values of this gradient are used as an input for flux balance analysis [KPE03], which models fluxes in networks on basis of structural and stoichiometric information. Some starting concentrations are necessary, which are taken from the oxygen gradient as input for different scenarios: The middle network visualises the fluxes near oxygen depletion, the left one normal oxygen level and the right one higher oxygen level than in the air. In this way respective flux visualisations can be shown for different scenarios, based on simple measurements.

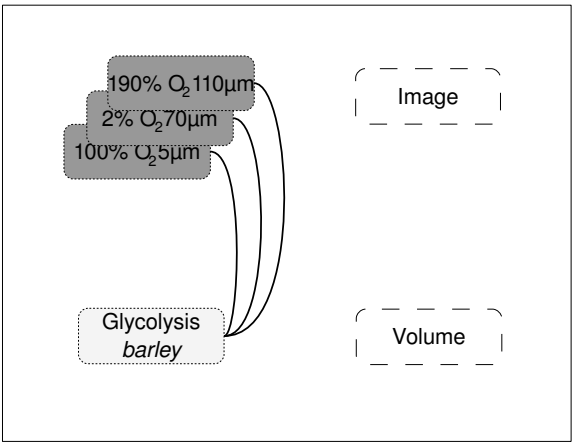

Figure 7: Instance of the mapping graph for use case flux balance analysis. The mapping graph consists of an oxygen gradient (simple measurement), which is mapped to a glycolysis network of barley and used for flux balance analysis.

5 Conclusion and Outlook

Using high-throughput methods biological researchers gather lots of data of different types from multiple -omics areas, network models and spatial data. For intuitive exploration of this data we propose a data structure representing the biological data and supporting all necessary mapping and data exploration methods. The biological data was integrated using a mapping graph, which allows intuitive combination of data. Its nodes represent the data types and its edges represent mappings between data types. Some mapping types were analysed and finally four exemplary use cases of data integration were described in detail.

The data structure and some of the mappings and use cases are implemented on the basis of the VANTED system [JKS06] in Java3D to provide scientists with the possibility to handle not only -omics data and network models, but also to account for two- and three-dimensional data in one system. We plan to complete the development and implementation of further mapping and interaction methods together with life scientists before releasing it as an Open Source Add-On for VANTED.

Acknowledgements

We would like to thank Rainer Pielot for help with the multimodal alignment and Eva Grafahrend-Belau for help with the flux balance analysis. This work was supported by grant BMBF *0315044A*.

References

[Bar06] K. U. Barthel. 3D-Data Representation with ImageJ. In *ImageJ Conference*, 2006.

[BDS04] U. Brandes, T. Dwyer, and F. Schreiber. Visual Understanding of Metabolic Pathways Across Organisms Using Layout in Two and a Half Dimensions. *Journal of Integrative Bioinformatics*, 1(1):119–132, 2004.

[BHQ+01] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nature Genetics*, 29(4):365–371, 2001.

[HLD+07] T. Hjornevik, T. B. Leergaard, D. Darine, O. Moldestad, A. M. Dale, F. Willoch, and J. G. Bjaalie. Three-Dimensional Atlas System for Mouse and Rat Brain Imaging Data. *Frontiers in Neuroinformatics*, 1:1–12, 2007.

[HMWD04] Z. Hu, J. Mellor, J. Wu, and C. DeLisi. VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics*, 5:17.1–8, 2004.

[JKS06] B. H. Junker, C. Klukas, and F. Schreiber. VANTED: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7:109.1–13, 2006.

[KBT+06] J. Köhler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Ruegg, C. Rawlings, P. Verrier, and S. Philippi. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22(11):1383–1390, 2006.

[Kol02] F. A. Kolpakov. BioUML - Framework for visual modeling and simulation of biological systems. In *International Conference on Bioinformatics of Genome Regulation and Structure*, pages 130–133, 2002.

[KPE03] K. J. Kauffman, P. Prakash, and J. S. Edwards. Advances in flux balance analysis. *Current Opinion in Biotechnology*, 14(5):491–496, 2003.

[MPLB07] E. B. Moore, A. V. Poliakov, P. Lincoln, and J. F. Brinkley. Mindseer: A Portable and Extensible Tool for Visualization of Structural and Functional Neuroimaging Data. *BMC Bioinformatics*, 8:389.1–12, 2007.

[RJL+07] D. V. Rubtsov, H. Jenkins, C. Ludwig, J. Easton, M. R. Viant, U. Günther, J. L. Griffin, and N. Hardy. Proposed reporting requirements for the description of NMR-based metabolomics experiments. *Metabolomics*, 3(3):223–229, 2007.

[RWW+04] H. Rolletschek, W. Weschke, H. Weber, U. Wobus, and L. Borisjuk. Energy state and its control on seed development: starch accumulation is associated with high ATP and steep oxygen gradients within barley grains. *Journal of Experimental Botany*, 55(401):1351–1359, 2004.

[SMO+03] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[SWH05] D. Stalling, M. Westerhoff, and H.-C. Hege. *Amira: A highly interactive system for visual data analysis*, chapter 38, pages 749–767. Academic Press, Inc. Orlando, FL, USA, 2005.

[TPG+03] C. F. Taylor, N. W. Paton, K. L. Garwood, P. D. Kirby, D. A. Stead, Z. Yin, E. W. Deutsch, L. Selway, J. Walker, I. Riba-Garcia, S. Mohammed, M. J. Deery, J. A. Howard, T. Dunkley, R. Aebersold, D. B. Kell, K. S. Lilley, P. Roepstorff, J. R. Yates, A. Brass, A. J. Brown, P. Cash, S. J. Gaskell, S. J. Hubbard, and S. G. Oliver. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature Biotechnology*, 21(3):247–254, 2003.

[vIKP+08] M. van Iersel, T. Kelder, A. Pico, K. Hanspers, S. Coort, B. Conklin, and C. Evelo. Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics*, 9:399.1–9, 2008.

[YAL02] T. S. Yoo, M. J. Ackerman, and W. E. Lorensen. Engineering and algorithm design for an image processing API: a technical report on ITK - the insight toolkit. In *Proceedings of Medicine Meets Virtual Reality*, pages 586–592, 2002.