Representing Experimental Biological Data in Metabolic Networks

Tim Dwyer[1], **Hardy Rolletschek**[2], **Falk Schreiber**[3]
[1]School of Information Technologies
University of Sydney
Sydney NSW 2006, Australia

`dwyer@cs.usyd.edu.au`

[2]Department of Molecular Genetics and [3]Bioinformatics Center
Institute of Plant Genetics and Crop Plant Research
Corrensstraße 3, 06466 Gatersleben, Germany

`{rollet,schreibe}@ipk-gatersleben.de`

Abstract

This paper describes a novel approach to representing experimental biological data in metabolic networks. The aim is to allow biologists to visualise and analyse the data in the context of the underlying processes. Biological networks can be modelled as graphs and visualised using graph drawing methods. We present a general method for mapping experimental data onto nodes and edges of a graph and to visualise the data-enriched networks in 2½ dimensions such that the data is easy to understand. Our focus is on time series data occurring during developmental analysis. We demonstrate the utility of our approach by a real world example from the seed development of barley *(Hordeum vulgare).*

Keywords: Metabolic networks, Visualisation, Graph drawing, Metabolic profiling, Time series data

1 Introduction

Detailed knowledge of metabolic functionality and control represents an essential basis for plant functional genomics. To analyse major metabolites of primary metabolism (e.g. sugars, sugar alcohols, amino acids, intermediates of glycolysis and citrate cycle, nucleotides and their sugars) both enzymatic and chromatographic methods are most widely used. More recently, new tools for metabolic profiling have become available. Mass spectroscopy is coupled to liquid and gas chromatography (Buchholz *et al.* 2001, Fiehn *et al.* 2000, Roessner *et al.* 2000). This novel methodology offers the fascinating possibility of analysing hundreds of metabolites simultaneously in a quantitative and comprehensive manner, finally aiming to determine the entire metabolome.

These experimental methods output huge amounts of data and biologists require tools to assist in the analysis of these large data sets. To make sense of the experiments this data has to be interpreted in the context of the underlying biological processes. Biological processes are often represented as complex networks such as metabolic networks (Michal 1999), signal transduction pathways (Schacherer *et al.* 2001), and protein-protein interaction networks (Mayer & Hieter 2000). Dynamic visualisation methods have proven to be useful for understanding the relationships between the components of the networks and have been used in some sophisticated research tools (Demir *et al.* 2002, Širava *et al.* 2002, Friedrich & Schreiber 2003).

To visualise experimental data several standard methods are used (e.g. displaying data in tables, histograms and line-graphs) and new approaches have been developed such as visualisation of gene expression micro-array data (Zhou & Liu 2003). Comparison of single metabolites may give a detailed view of individual aspects. However, to recognize causal relationships within the metabolic network the experimental data has to be represented within this network and more sophisticated methods are necessary. Existing approaches use static visualisations (fixed, pre-computed or hand-drawn drawings) of metabolic networks and map the experimental data onto predefined positions. In (Nakao *et al.* 1999) the metabolic pathway diagrams of the KEGG system (Kanehisa & Goto 2000) were linked to the EXPRESSION database for integration with DNA micro-array data. Wolf *et al.* (Wolf *et al.* 2000) use static visualisations of metabolic pathways from KEGG and in-house sources and display protein and mRNA expression data in these diagrams such that colours indicate the relative change in expression level and reproducibility. However, static visualisation has several drawbacks (see Schreiber 2002), which negatively impact on this particular application, and dynamic visualisation is necessary. Dynamic visualisation is the computation of a drawing of the network at the time the diagram is needed.

In this work we combine dynamic network visualisation with mapping of experimental data, especially time series data about metabolites, onto the networks. This novel approach allows an easy visual analysis of processes in organisms and may assist users in the analysis of large data sets from biological experiments.

The paper is structured as follows: In Section 2 we define the graph model and graph representation we are working with, describe how we derive our metabolic networks and how the experimental data is mapped onto them. Section

---

Copyright © 2004, Australian Computer Society, Inc. This paper appeared at the 2nd Asia-Pacific Bioinformatics Conference (APBC2004), Dunedin, New Zealand. Conferences in Research and Practice in Information Technology, Vol. 29. Yi-Ping Phoebe Chen. Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.

3 deals with our dynamic visualisation method and some implementation aspects. Finally, in Section 4 we use our approach for the visual analysis of experimental data from the seed development of barley.

2 Representation of Metabolic Networks

2.1 Graph Modelling

A graph $G=(V,E)$ is a mathematical structure that consists of a finite set of nodes V and a finite set of edges E. Each edge $e\in E$ connects at most two nodes $u,v\in V$. A hyper-graph $G_H= (V,E_H)$ consists of finite sets of nodes V and finite set of hyper-edges E_H. Each hyper-edge $e\in E_H$ connects a set of nodes $u_1,\dots,u_n\in V$. A bipartite graph $G_B= (V_1\cup V_2,E_B)$ consists of finite sets of nodes V_1 and V_2 and a finite set of edges E_B. Each edge $e\in E_B$ connects exactly two nodes, one node from the set V_1 with one node from the set V_2.

From a formal point of view a metabolic network is a hyper-graph. The nodes represent the metabolites and the hyper-edges represent the reactions. Each hyper-edge connects all metabolites of a reaction. A hyper-edge can be seen as a *n*-ary relation between nodes. Hyper-graphs can be represented by bipartite graphs, and often metabolic networks are modelled as bipartite graphs, especially for simulations of such networks (Hofestädt & Thelen 1998, Reddy *et al.* 1993). Here the reactions themselves are nodes, and edges are binary relations connecting metabolites of reactions with reaction nodes.

Our interest in this paper is not in simulating metabolic networks, rather in simply using graphs to model them. Again, in graph $G=(V,E)$ nodes $v\in V$ represent metabolites and edges $e\in E$ represent reactions. We use additional reaction nodes $v\in V$ only for reactions that connect more than two metabolites or that involve intermediate metabolites which are not represented.

2.2 Metabolic network data

The sources of our metabolic network data are the KEGG LIGAND database (Goto *et al.* 2002) and the BioPath system (Forster *et al.* 2002). To build the networks of interest two steps were applied:

1. The data from these databases was transformed into graphs and stored as GML files. GML (Graph Modelling Language) is a widely used an easily extendible exchange format for graphs (Himsolt 1997, Himsolt 2000). All networks were merged into one big network of metabolic reactions.

2. We identified the partial network of interest. This network is defined by the metabolites for which experimental data exists and the main connections between these metabolites.

As part of these steps the network was manually edited for the following reasons:

- The merging of data from different sources introduced inconsistencies into the network such as different names for the same metabolite. There are also errors in the data from the databases such as the direction of some reactions.

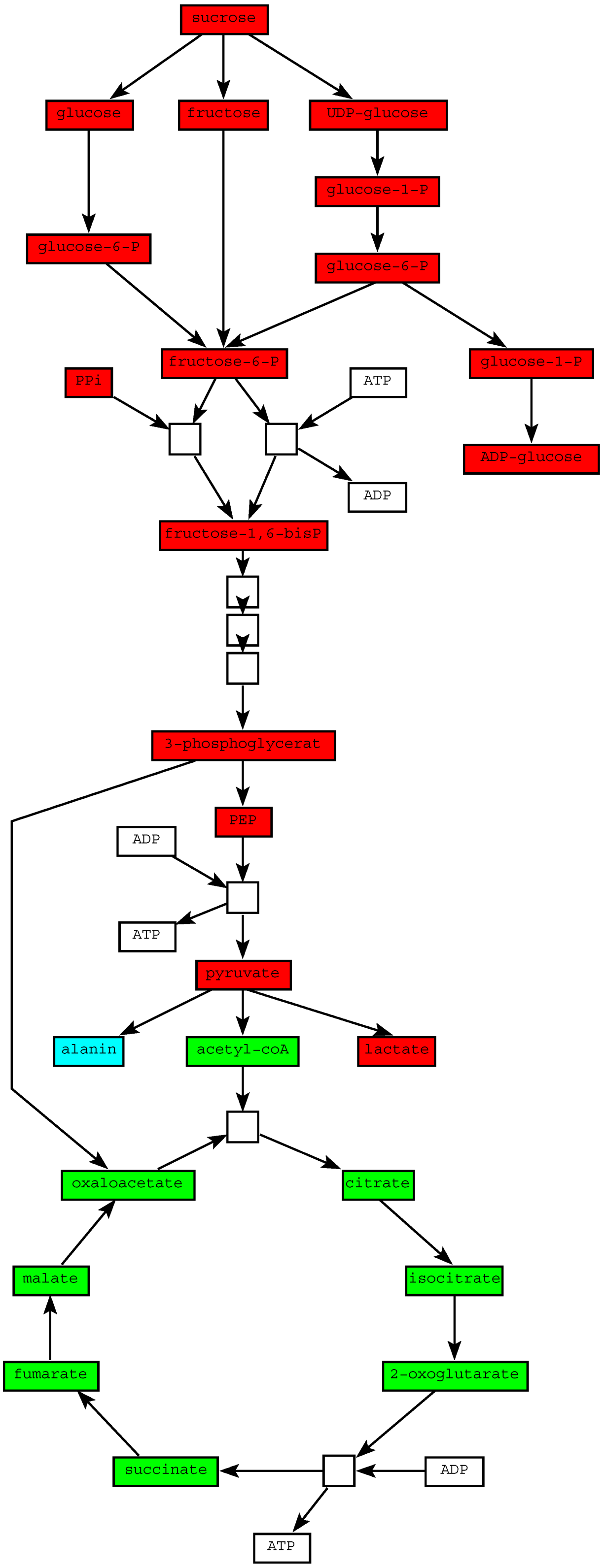


Figure 1: The metabolic network for the experimental data described in Section 4. Different colours are used to distinguish different pathways in the network such as glycolysis and TCA cycle.

- We have to deal with some plant specific physiological aspects that are not correctly represented in the above-mentioned general metabolic pathway databases.

As an example for the resulting network, Figure 1 shows the network for the data discussed in Section 4.

2.3 Experimental biological data

Given the metabolic network of interest as a GML file we can add the experimental data to the metabolites and reactions. A GML file consists of a hierarchical key-value list: a key is a sequence of alphanumeric characters (e.g. *graph*, *id*), and a value is a *string*, an *integer*, a *floating point number*, or a *list of key-value pairs*. Using GML it is possible to attach additional information to every object.

We add information about our experimental time series data to the graph. If, for example, for a specific metabolite such information exists (the metabolite was measured and quantified over several days), its node is assigned the data about the measured amount on the different days. An example of a GML file representing a network and additional time series data looks like:

```
graph [
  node [
    id 1
    label "Sucrose"
    experimental_data [
      day "day 0"
      value 1534
      day "day 2"
      value 2801
      day "day 4"
      value 2914
    ]
  ]
  node [
    id 2
    label "Fructose"
    experimental_data [
      day "day 0"
      value 2341
      day "day 2"
      value 2894
      day "day 4"
      value 2786
    ]
  ]
  edge [
    id 1
    label ""
    experimental_data [
      day "day 0"
      value 54
  …
    ]
  ]
]
```

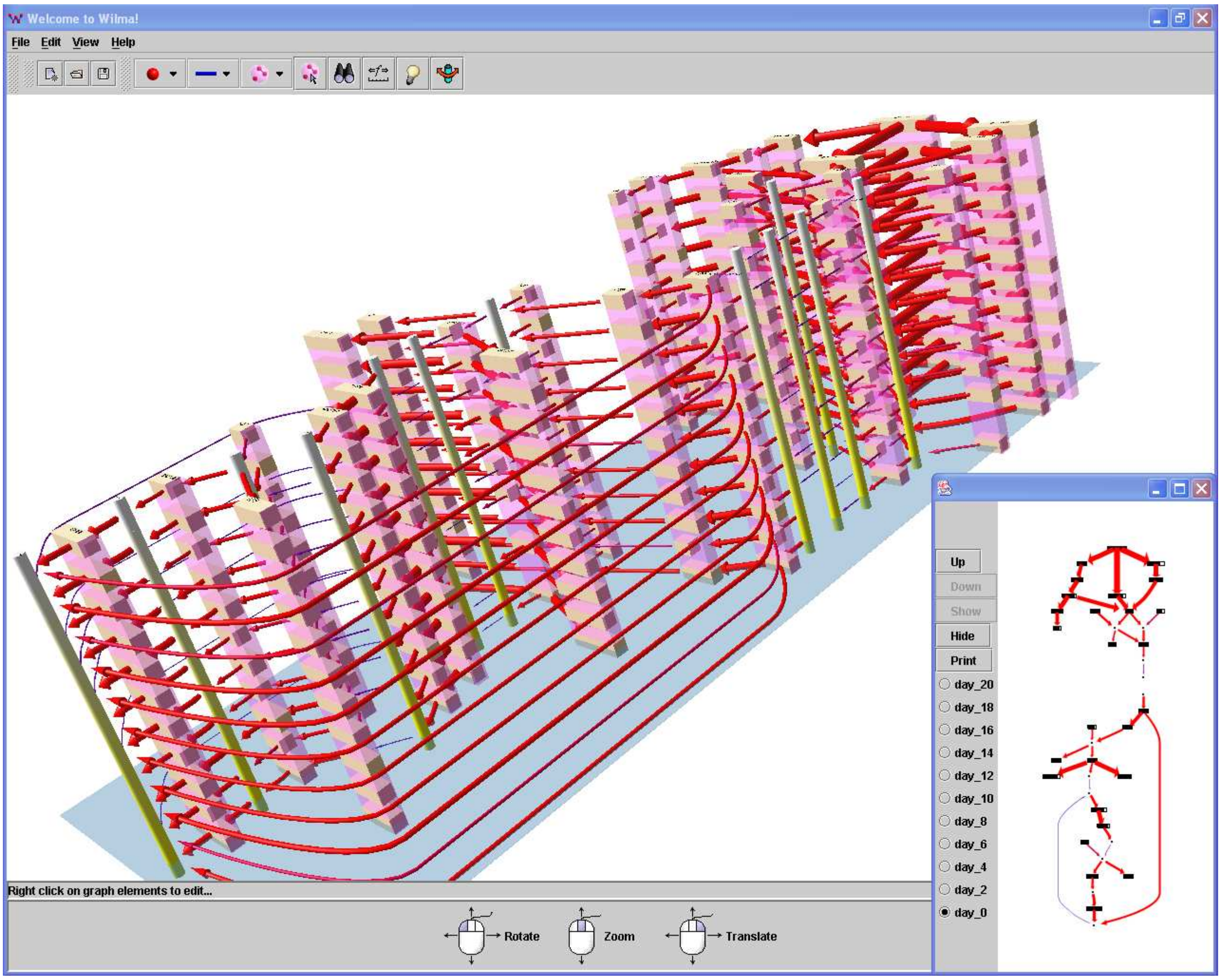
3 Dynamic visualisation method

The mathematical model, which we call a “graph”, is not related to a visual representation of the graph (and therefore to a visualisation of the underlying metabolic network). The problem of visualising these structures can be formulated as a graph drawing problem (Di Battista *et al.* 1999). Some biological networks and hierarchies have been successfully visualised using standard graph drawing algorithms: protein-protein interaction networks are often laid out by force-directed graph drawing algorithms (Basalaj & Elbeck 1999) and hierarchies are usually visualised using tree drawing algorithms (Reingold & Tilford 1981).

The structural characteristics of metabolic networks make them particularly amenable to *layered graph drawing* methods (Becker & Rojas 2001, Schreiber 2002). Special extensions of these methods have been developed for the dynamic visualisation of metabolic pathways (Becker & Rojas 2001, Karp & Paley 1994, Mendes 2000, Schreiber 2002, Širava *et al.* 2002). However, these approaches were concerned with 2D representations of the networks. Note that the notion of layered graph drawing, as used above, usually means arranging the nodes in the network onto layers in 2D such that edges always connect nodes on different layers. Also, the layering is chosen such that as many edges as possible point downward in order to better show flow from sources to sinks.

In (Brandes *et al.* 2003) an additional type of layering was introduced. The 3rd dimension available in modern computer graphics was utilised to stack a set of related biochemical pathways for easy comparison. We call this style of constrained 3D graph visualisation “2½D”, the distinction being that only two dimensions are used to arrange the graphs and the third dimension is used for a different purpose. Our focus for this paper is to use a similar technique to visualise a metabolic network’s structure as well as showing the amounts of metabolites and the flow within the network as they change over time. That is, the third dimension is now mapped to the ordinal variable of time.

The visualisation shown in Figure 2 was produced using a tool developed by the authors called WilmaScope (http://wilma.sourceforge.net). The 2D layered layout used is produced with a modified version of the *dot* program (Koutsofios & North 1993) distributed with the *GraphVis* graph visualisation system (http://www.graphviz.org). WilmaScope then extrudes the 2D drawing into the third dimension in order to show the time series on each level of the stack. WilmaScope provides many facilities for interactively browsing a graph, for example users can zoom into an individual node, select only a portion of the graph to show or they can set a weight threshold for edges to be shown. These facilities are very useful in browsing the metabolic networks discussed in this paper, and in some of the images shown parts of the network are hidden in order to better show detail of time series.



Right click on graph elements to edit...

Figure 2: A screenshot of the WilmaScope system examining a 2½D visualisation of the metabolic network from Figure 1

In the visualisation shown in Figure 2 each level or slice in the 2½D structure corresponds to sample data taken on different days, with the oldest data (day 0) on the bottom slice and the most recent (day 20) at the top. The small window floating before the "Wilma" window shows a cross-section through the lowest slice of the stack and corresponds with the semi-transparent blue plane that can be seen in the 2½D view. This "water-level" can be moved up or down to highlight an individual cross section of the graph.

Using this approach we are able to visualise the time series data for each metabolite in-situ. That is, each metabolite can be shown as a histogram giving the measured values at each point in time. Where time series data for a particular metabolite is not available, or is not interesting to the user, the metabolite is shown as a narrow column. Figure 3 shows two possible 2½D representations for a single metabolite with associated time series data. Each section represents the measured value of the metabolite on a particular day. In the representation on the left a square-root scale is used to determine the radius such that the area of the disc is directly proportional to the amount of the metabolite. In the histogram-like representation on the right a log scale is used to better fit the bars into the space available. Note that the transparent pink box makes it easy to compare each bar against the maximum value. Figure 5 shows the complete network using the disc representation. 3D graphics is most effective in an interactive environment in which the user is able to freely rotate or fly around the model. Figure 6 shows the network projected in parallel to better show the slices in the 2½D structure and rotated to give a feeling of depth.

The diameter and colour of the reaction edges can also give an indication of flow given by the activity and quantity of enzymes (this could, for example, be derived from expression data). Note that in the application example (Section 4) such information is not yet available from our experiments. This data is an estimate based on the quantity of source and sink metabolite and is shown here purely to illustrate the potential of the visualisation paradigm.

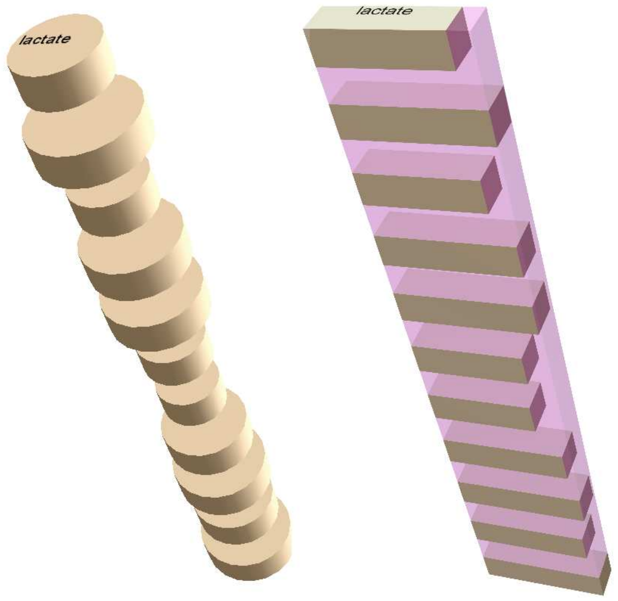


Figure 3: A close-up of two possible 2½D representations for a single metabolite. Each section represents the measured value of the metabolite on a particular day. In the representation on the left a square-root scale is used to determine the radius such that the area of the disc is directly proportional to the amount of the metabolite. On the right hand side a log scale is used and the transparent pink box makes it easy to compare each bar against the maximum value.

4 Application example

4.1 Metabolic data

To analyse major metabolites of primary metabolism both enzymatic and chromatographic methods are most widely used, and new tools for metabolic profiling have become available. Mass spectroscopy is coupled to liquid and gas chromatography (Fiehn *et al.* 2000, Roessner *et al.* 2000, Buchholz *et al.* 2001). We used this approach together with conventional chromatographic techniques (ion chromatography, HPLC) to investigate the metabolite pattern of growing barley caryopses (*Hordeum vulgare*).

The agronomical importance of cereal seeds is principally based on their accumulation of storage products, mainly starch and proteins. Despite extensive studies on the structure, biochemistry and genetics of developing grains (Duffus & Cochrane 1982, Olsen 1992, Bewley & Black 1994) the regulatory mechanisms underlying their high storage capacity are largely unknown. During their development, caryopses undergo distinct growth phases and differentiation events. These in turn are reflected in changes of the metabolic state. Biosynthetic fluxes increase in a specific temporal and spatial manner.

To investigate these patterns, time series analyses of metabolites are required. Caryopses were harvested every 2 days over a growth period of about 20 days post anthesis. Seed development was analysed from 0 to 20 days post anthesis (DPA), covering the pre-storage, intermediate and storage phase. In this period the endosperm enlarges, becoming the main storage organ of cereal seeds. Within the pre-storage phase caryopsis consists mainly of pericarp tissue, embedding the liquid endosperm. Increase in the fresh weight and starch accumulation is low. The subsequent intermediate phase begins after endosperm cellularisation at 4-5 DPA and proceeds with the differentiation of endosperm tissues. Starch accumulation starts, although with low synthesis rates. During the main storage phase (from 10-11 DPA onwards), the high starch synthesis rate is evident. About 70 metabolites were measured and quantified. A typical example of time series data is given in Figure 4.

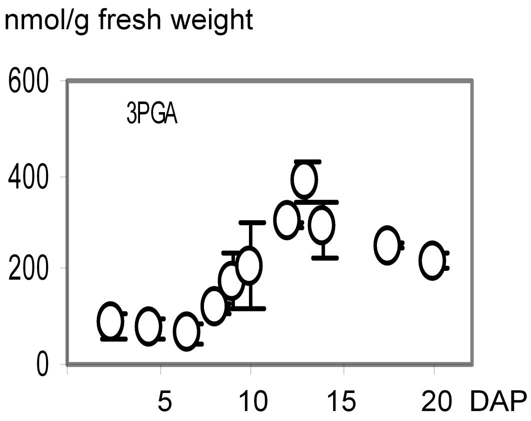


Figure 4: A typical example of time series data showing the metabolite 3PGA at several days post anthesis (DPA).

4.2 Visual analysis of the experimental data

We used two different approaches within our visualisation method to analyse the experimental data in the context of metabolic networks.

Firstly, the sizes of all nodes were fixed. This size represents 100% and the amount of a specific metabolite on a specific day was given as a percent of the maximal value of this metabolite over all days. This approach emphasises the relative change of metabolites over time and the relative flow through the network. An example of this visualisation style is given in Figure 7.

Secondly, the sizes of nodes were variable, depending on the amount of the corresponding metabolites and were directly proportional to this amount. In this type of diagram the whole amount of metabolites in the tissue and the absolute ratio of different metabolites can be seen. Whereas the first approach shows relative changes this visualisation can be also very useful as often the ratio of two metabolites can be important (such as the ratio of the absolute amounts of ATP and ADP). An example of this approach is shown in Figure 8.

As we currently have data from only one such experiment, the evaluation and comparison of these visualisation styles requires further work. But we believe that both approaches will be quite useful for the visual analysis of experimental data.

5 Discussion

We have presented a novel method for mapping experimental data onto nodes and edges of a graph, and for visualising these data-enriched networks in 2½ dimensions. This allows biologists to visualise and analyse experimental data in the context of the underlying biological processes. Our focus was on time series data and we demonstrated the utility of our approach by an example from the seed development of *Hordeum vulgare*.

This work is under continuing development and is carried out in close cooperation between biologists and computer scientists. We have already implemented the visualisation system, whereas the construction of the network of interest and the mapping of the data onto the network is currently a semi-automatic process which needs manual editing. The next step will be the development of more automatic methods for these parts. We also look forward to evaluating our method with more data.

As mentioned in Section 3, the ability to interact with and "fly" around the 2½D graph structures is very important in facilitating a users understanding of the model. The static screen shots in this paper do not do them justice. A quantitative user study evaluating the effectiveness of the 2½D graph visualisation paradigm is planned for the near future. Subjects will complete speed and accuracy tests using both a physical model of a 2½D graph structure and a set of 2D graphs printed on cards. Furthermore, we will compare our approach with 2D graph drawings where the temporal changes are displayed with graphical elements (e.g., histograms) associated with the nodes and edges.

In this work we used metabolite data which is mapped onto metabolic networks. In general the described approach can be used for other experimental data (e.g. transcriptome data) and for mapping of data onto other biological networks (e.g. gene regulatory networks).

Acknowledgement

This work was supported by the German Ministry of Education and Research (BMBF) under grant 0312706A and the Land Sachsen-Anhalt under grant MK-LSA 0031KL/1002L. We would like to thank Ljudmilla Borysyuk and Mohammed Hajirezaei for fruitful discussions.

References

Basalaj, W. and Elbeck, K. (1999): Straight-line drawings of protein interactions. *Proc. 7th Intl. Symposium on Graph Drawing (GD'99)*, Springer LNCS **1731**:259-266.

Becker, M. Y. and Rojas, I. (2001): A graph layout algorithm for drawing metabolic pathways. *Bioinformatics*, **17**(5):461-467.

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

Brandes, U., Dwyer, T., and Schreiber, F. (2003): Visualizing Related Metabolic Pathways in Two and a Half Dimensions. *Proc. 11th Intl. Symposium on Graph Drawing (GD'03)*, Springer LNCS, to appear.

Buchholz, A., Takors, R., and Wandrey C. (2001): Quantification of intracellular metabolites in Escherichia coli K12 using liquid chromatographic-electrospray ionization tandem mass spectrometric techniques. *Analytical Biochemistry* **295**:129-137.

Demir, E., Babur, O., Dogrusoz, U., Gursoy, A., Nisanci, G., Cetin-Atalay, R., and Ozturk, M. (2002): PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics* **18**(7):996-1003.

Di Battista, G., Eades, P., Tamassia, R., and Tollis, I. G. (1999): *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, New Jersey.

Duffus, C. M. and Cochrane, M. P. (1982): Carbohydrate metabolism during cereal grain development. *The physiology and biochemistry of seed development, dormancy and germination*, Elsevier Biomedical press, 43-66.

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R. N., and Willmitzer, L. (2000): Metabolite profiling for plant functional genomics. *Nature Biotechnology* **18**:1157-1161.

Forster, M., Pick, A., Raitner, M., Schreiber, F., and Brandenburg, F. J. (2002): The System Architecture of the BioPath system. *In Silico Biology* **2**(3):415-426.

Friedrich, C. and Schreiber, F. (2003): Visualization and navigation methods for typed protein-protein interaction networks. *Applied Bioinformatics*, to appear.

Goto, S., Okuno, Y., Hattori, M., Nishioka, T., and Kanehisa, M. (2002): LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Research* **30**:402-404.

Himsolt, M. (1997): GML: A portable Graph File Format. *Technical report*, University of Passau.

Himsolt, M. (2000): Graphlet: Design and Implementation of a Graph Editor. *Software - Practice and Experience* **30**(11):1303-1324.

Hofestädt, R. and Thelen, S. (1998): Qualitative Modeling of Biochemical Networks. *In Silico Biology* **1**(1):39-53.

Kanehisha, M. and Goto, S. (2000): KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleid Acids Research* **28**:27-30.

Karp, P. D. and Paley, S. M. (1994): Automated Drawing of Metabolic Pathways. *Proc. 3rd Intl. Conference on Bioinformatics and Genome Research*:225-238.

Koutsofios, E and North, S. C. (1993): Drawing Graphs with Dot. *Technical report AT&T Bell Laboratories*, Murray Hill, NJ.

Luyf, A. C. M., de Gast, J., and van Kampen, A. H. C. (2002) : Visualizing metabolic activity on a genome-wide scale. *Bioinformatics* **18**(6):813-818.

Mayer, M. L. and Hieter, P. (2000): Protein networks - built by association. *Nature Biotechnology* **18**(12): 1242-1243.

Mendes P. (2000): Advanced Visualization of Metabolic Pathways in PathDB. *Proc. 8th Conference on Plant and Animal Genome*.

Michal, G. (1999): *Biochemical Pathways*. Spektrum Akademischer Verlag.

Nakao, M., Bono, H., Kawashima, S., Kamiya, T., Sato, K., Goto, S., and Kanehisa, M. (1999): Genome-scale gene expression analysis and pathway reconstruction in KEGG. *Genome Informatics* **10**:94-103.

Olsen, O. A., Potter, R. H., and Kalla, R. (1992): Histo-differentiation and molecular biology of developing cereal endosperm. *Seed Science Research* **2**:117-131.

Reddy, V. N., Mavrovouniotis, M. L., and Liebman, M. N. (1993): Petri Net Representations of Metabolic Pathways. *Proc. 1st Intl. Conference on Intelligent Systems for Molecular Biology (ISMB'93)*:328-336.

Roessner, U., Wagner, C., Kopka, J., Trethewey, R. N., Willmitzer, L. (2000): Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant Journal* **23**:131-142.

Schacherer, F., Choi, C., Gotze, U., Krull, M., Pistor, S., and Wingender, E. (2001): The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics* **17**(11):1053-1057.

Schreiber, F. (2002): High Quality Visualization of Biochemical Pathways in BioPath. *In Silico Biology* **2**(2):59–73.

Širava, M., Schäfer, T., Eiglsperger, M., Kaufmann, M., Kohlbacher, O., Bornberg-Bauer, E., and Lenhof, H. P. (2002): BioMiner – modeling, analyzing, and visualizing biochemical pathways and networks. *Bioinformatics* **18**(Suppl.2):S219-S230.

Wolf, D., Gray, C. P., and de Saizieu, A. (2000): Visualising gene expression in its metabolic context. *Briefings in Bioinformatics* **1**(3):297-304.

Zhou, Y. and Liu, J. (2003): AVA: visual analysis of gene expression microarray data. *Bioinformatics* **19**(2):293-294.

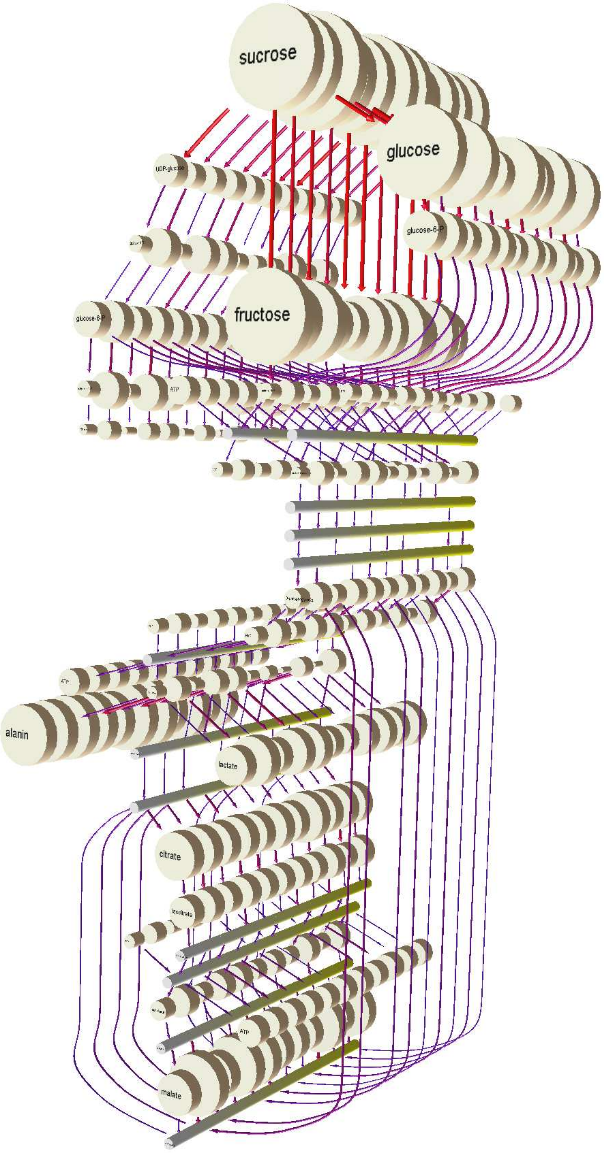


Figure 5: A view of the 2½D visualisation of the network using the "disc" representation for the metabolites.

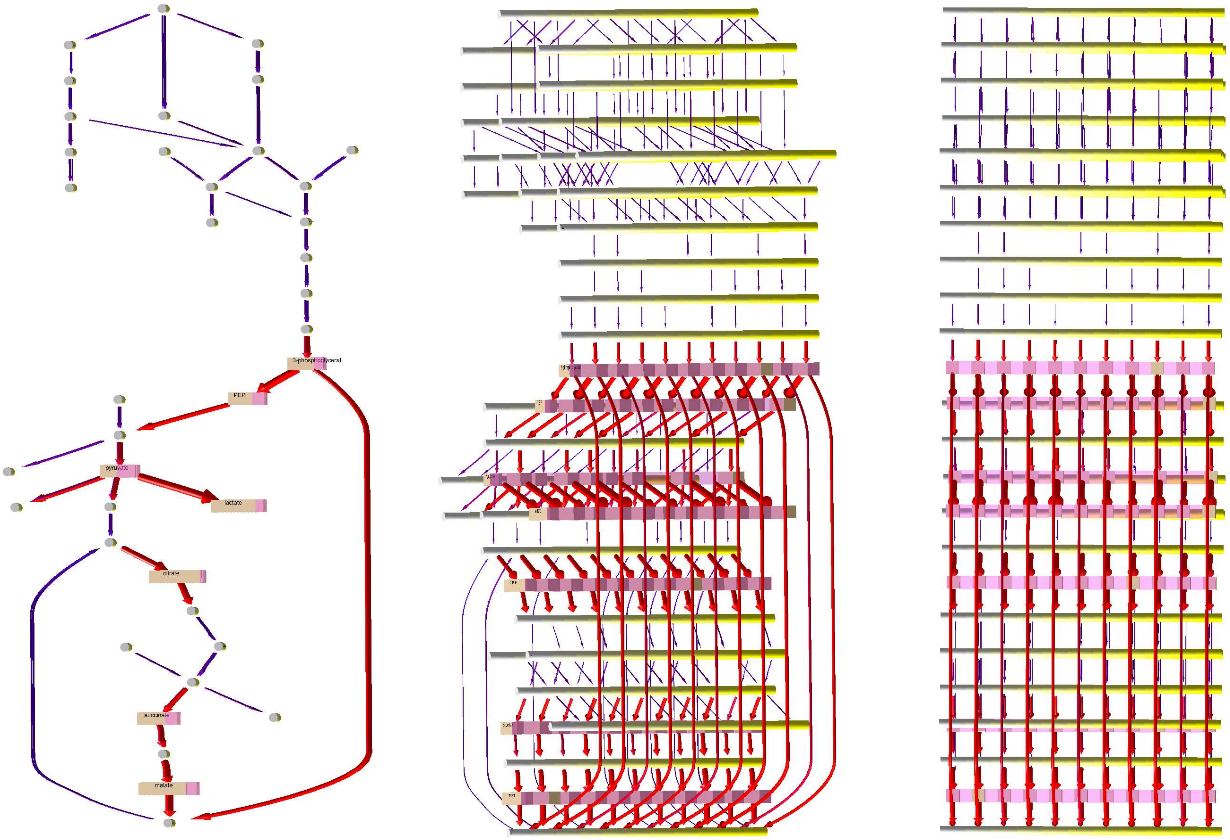

3-phosphoglycerat
PEP
pyruvate
lactate
citrate
succinate
malate

Figure 6: 3 parallel projected views of the network, viewed from three different orientations to give the reader a better understanding of the 2½D stacking method. WilmaScope allows interactive navigation to explore the stacked network from all directions.

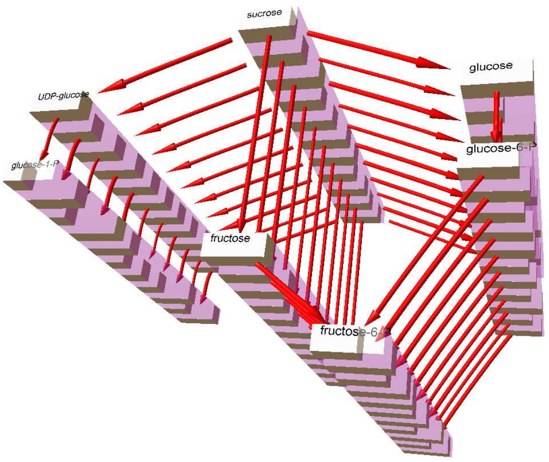

sucrose
UDP-glucose
glucose
glucose-1-P
glucose-6-P
fructose
fructose-6-P

Figure 7: Detail of the neighbourhood around sucrose using the fixed node width style.

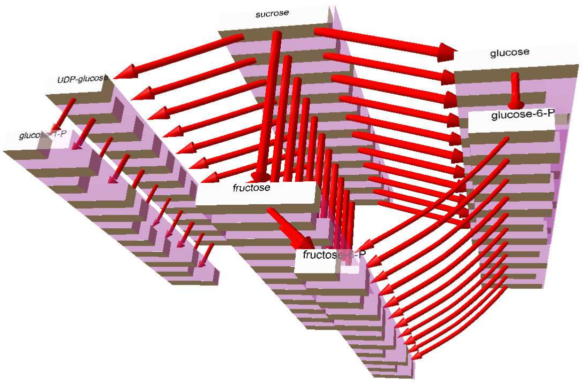

sucrose
UDP-glucose
glucose
glucose-6-P
fructose

Figure 8: The sub-network from Figure 7 shown with absolute node size.