Analysis and visualization of experiment data in the context of biological networks and classification hierarchies

Christian Klukas
Member of Plant Bioinformatics Group at the
Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany

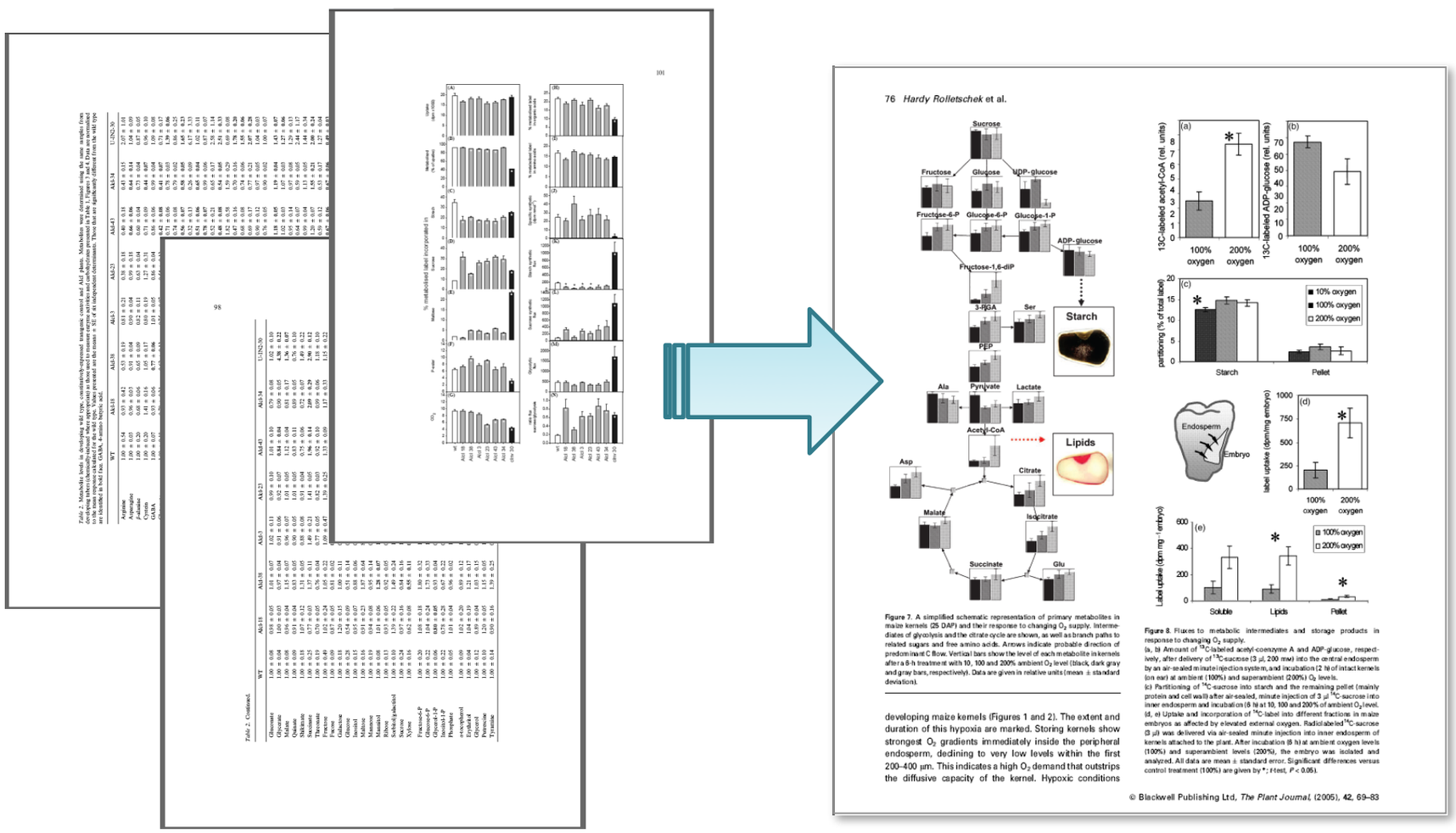
IPK GATERSLEBEN

Outline

1. Motivation
2. Data representation
    1. Experiment data
    2. Networks
    3. Classification hierarchies
    4. Alternative identifiers
3. Methods
    1. Data mapping
    2. Data visualization
    3. Data analysis (statistics)
    4. Navigation and interaction
4. Implementation
    1. Systems architecture
    2. Availability
5. Example use cases

Motivation

- High-throughput analysis techniques create difficult to handle large datasets
- Bioinformatics tools are essential for visualization and analysis, many tools exist but they are often specific for certain domains or linked to specific databases



76 *Hardy Rolletschek* et al.

developing maize kernels (Figures 1 and 2). The extent and duration of this hypoxia are marked. Storing kernels show strongest O_2 gradients immediately inside the peripheral endosperm, declining to very low levels within the first 200–400 μm. This indicates a high O_2 demand that outstrips the diffusive capacity of the kernel. Hypoxic conditions

© Blackwell Publishing Ltd, *The Plant Journal*, (2005), **42**, 69–83

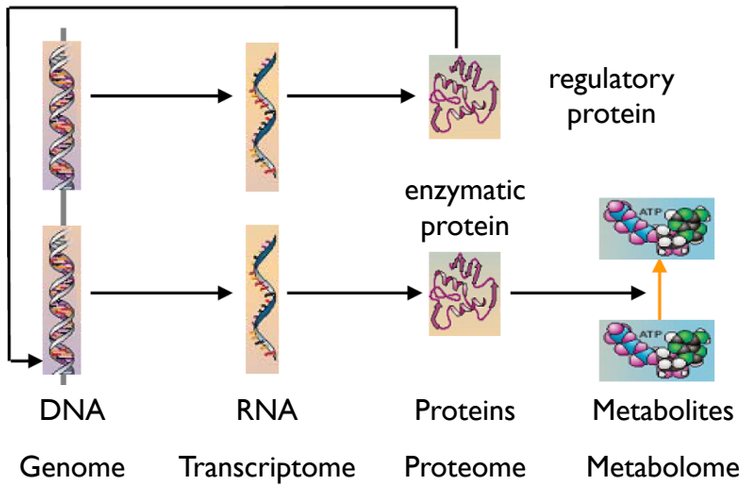
Data representation: experiment data

- Different "-omics" areas, e.g. proteomics and metabolomics, different (high-throughput) analysis techniques

- Data stored in databases or files

- Sometimes standardizied representation e.g. MIAME/MAGE for expression data, PEDRo for proteomics, ArMet for metabolite data

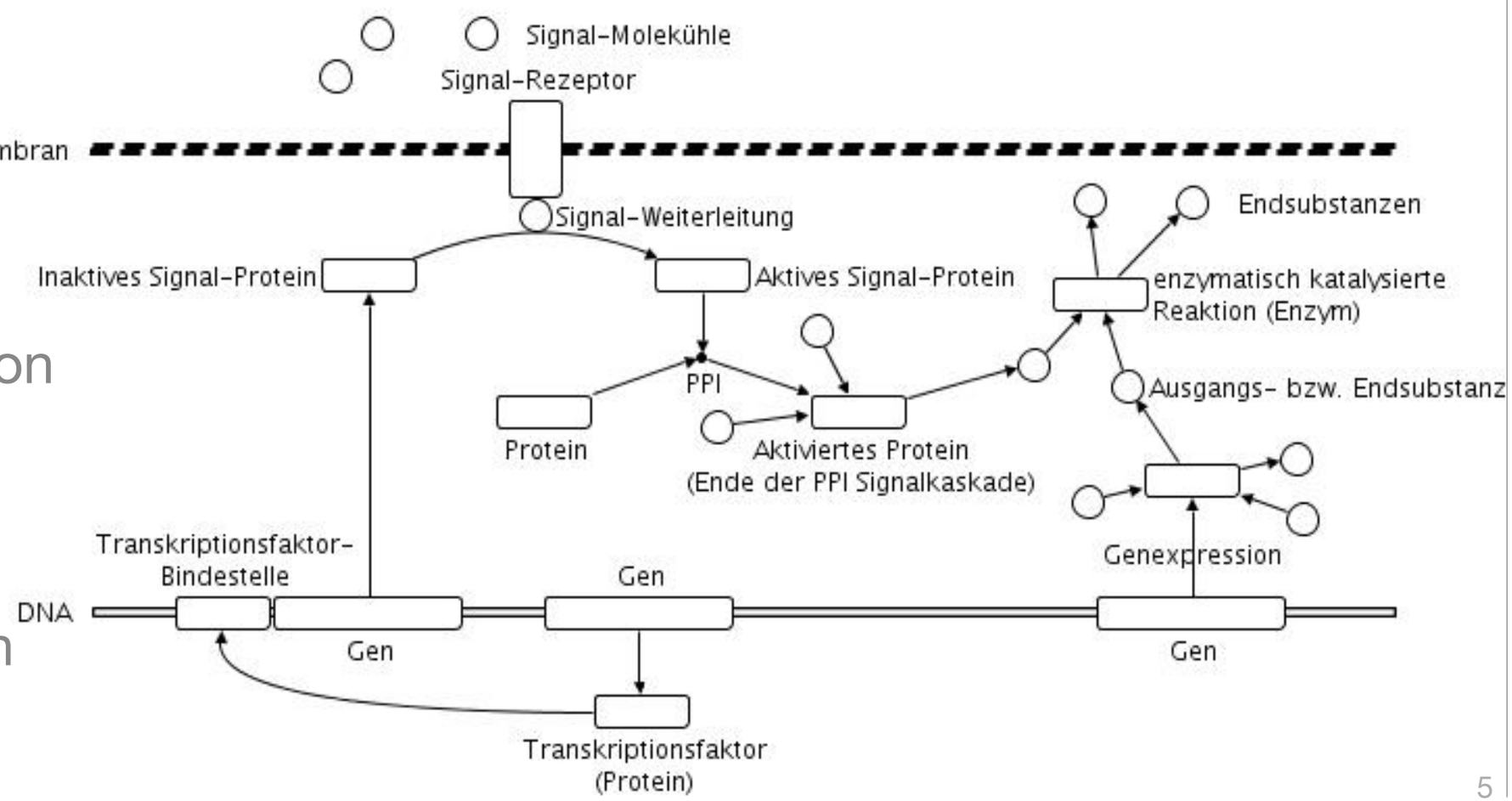
- Often unstructured (spreadsheet files, text files)



Data representation: biological networks

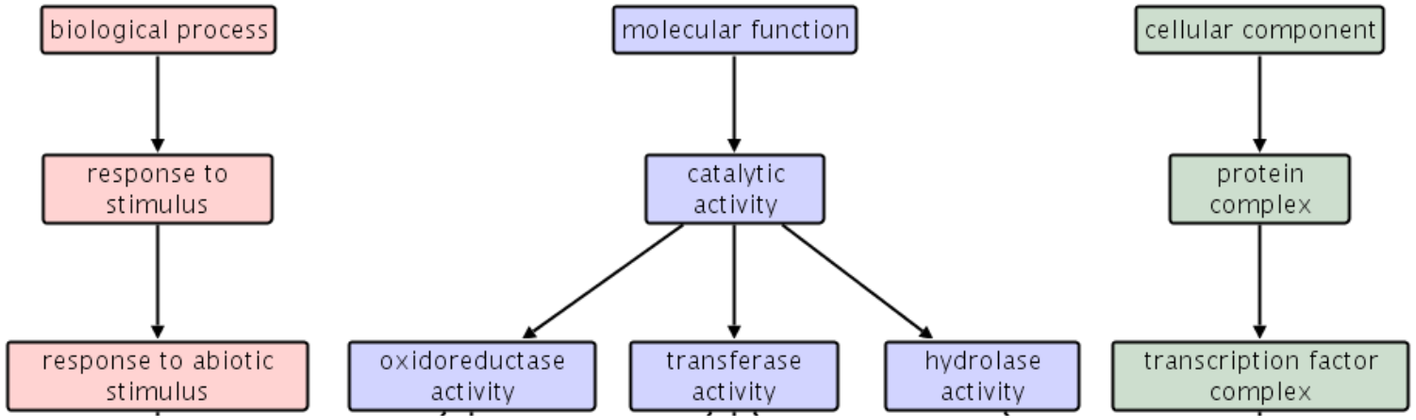
- Different biological domains and networks exist
- All of these networks are related or interconnected
- Use of general directed/undirected/mixed graph structure, node/edge attributes make identification of element type, visualization and analysis possible

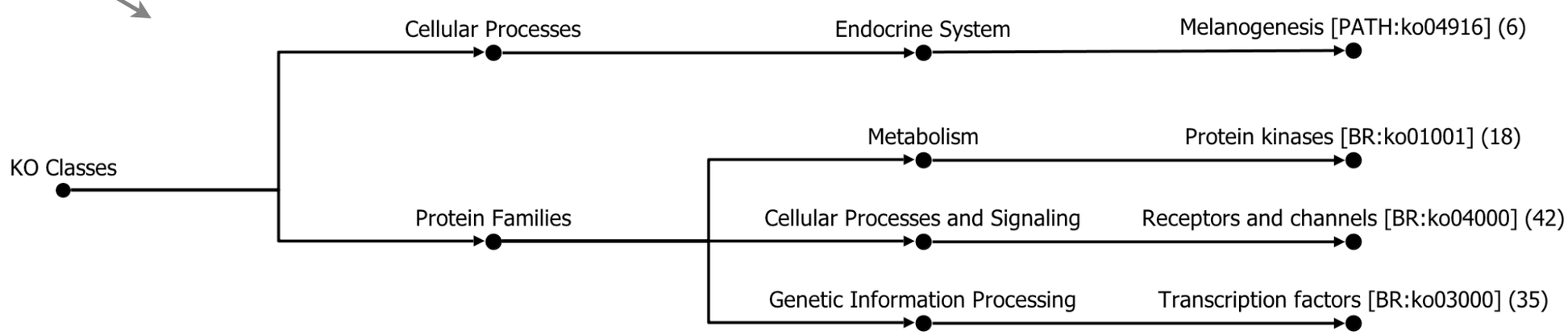
- PPI
- Biochemical reaction networks
- KEGG Pathways
- Signal transduction
- … and more



Data representation: classification hierarchies

- Classification hierarchies / ontologies are defined terms with hierarchical relationships
- Can be modelled as directed acyclic graphs or as trees
- Examples
  - Gene Ontology
  - KEGG Brite
  - MapMan





Data representation: alternative identifiers

- Knowledge about identifiers, synonyms, associated genes or general annotations

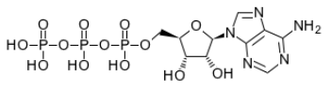
- Examples

  - KEGG Compound database contains compound names, synonyms and chemical formulas

  - ExPASy Enzyme database for enzyme names and synonyms

  - KEGG KO database contains orthologue genes, their gene IDs, names and links to other databases

COMPOUND: C00002

Entry	C00002 Compound
Name	ATP; Adenosine 5'-triphosphate
Formula	C10H16N5O13P3
Mass	506.9957
Structure	

ExPASy Home page | Site Map | Search ExPASy | Contact us | Swiss-Prot | ENZYME

Search ENZYME for Go Clear

NiceZyme View of ENZYME: EC 1.1.1.1

Official Name

Alcohol dehydrogenase.

Alternative Name(s)

Aldehyde reductase.

Reaction catalysed

An alcohol + NAD(+) <=> an aldehyde or ketone + NADH

Cofactor(s)

Zinc or iron.

Comment(s)

- Acts on primary or secondary alcohols or hemiacetals.
- The animal, but not the yeast, enzyme acts also on cyclic secondary alcohols.

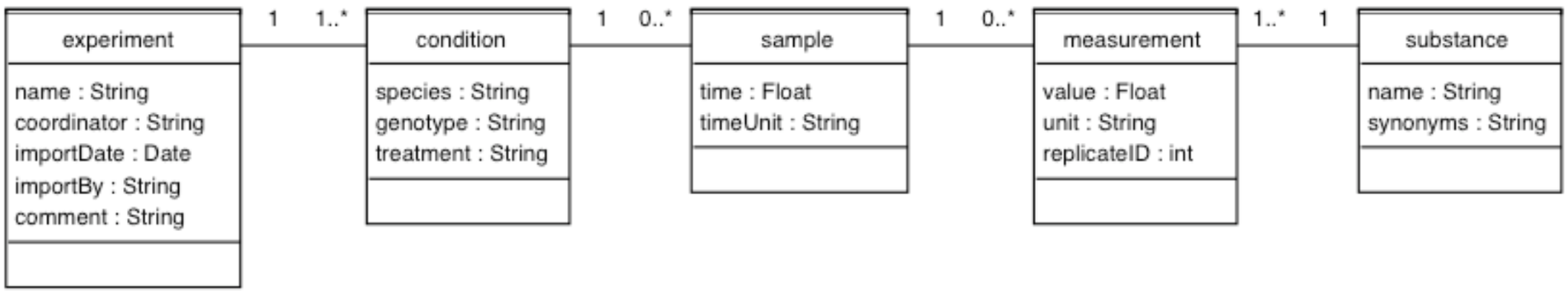
ORTHOLOGY: K00001

Help

Entry	K00001 KO
Name	E1.1.1.1, adh
Definition	alcohol dehydrogenase
Class	Metabolism; Carbohydrate Metabolism; Glycolysis / Gluconeogenesis [PATH:ko00010] Metabolism; Lipid Metabolism; Fatty acid metabolism [PATH:ko00071] Metabolism; Lipid Metabolism; Bile acid biosynthesis [PATH:ko00120] Metabolism; Lipid Metabolism; Glycerolipid metabolism [PATH:ko00561] Metabolism; Amino Acid Metabolism; Tyrosine metabolism [PATH:ko00350] Metabolism; Xenobiotics Biodegradation and Metabolism; 3-Chloroacrylic acid degradation [PATH:ko00641] Metabolism; Xenobiotics Biodegradation and Metabolism; 1- and 2-Methylnaphthalene degradation [PATH:ko00624] Metabolism; Xenobiotics Biodegradation and Metabolism; Metabolism of xenobiotics by cytochrome P450 [PATH:ko00980] BRITE hierarchy

Methods: data mapping

- General data model for experiment data



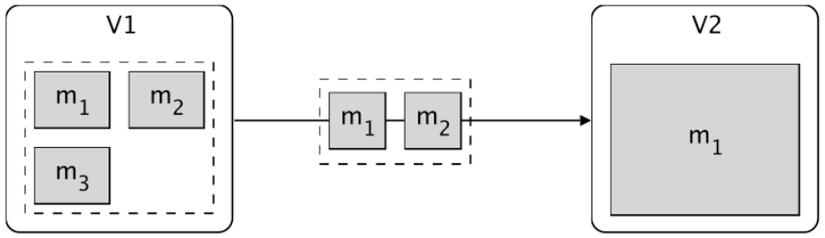
- General data model (mapping-graph) for biological networks and classification hierarchies

- Function for connecting experiment data with mapping-graph

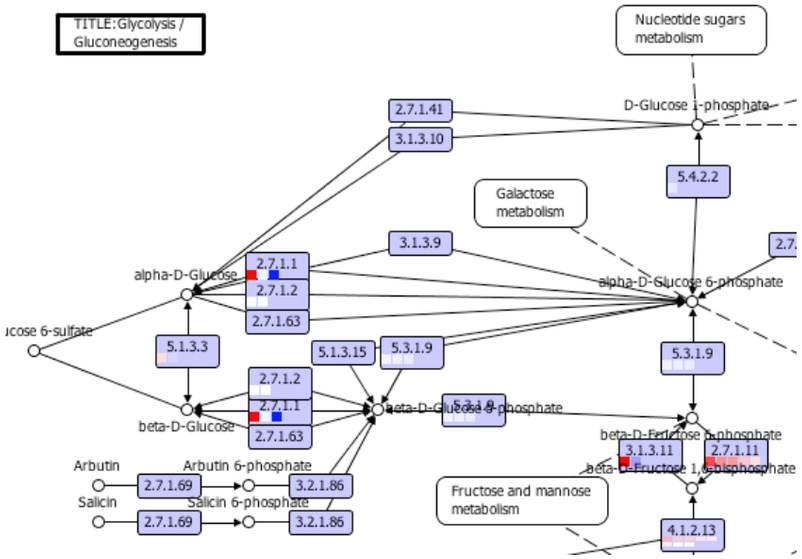
  - Consideration of alternative identifiers for graph *nodes and edges* and alternative identifiers or annotations for experiment data *substances*

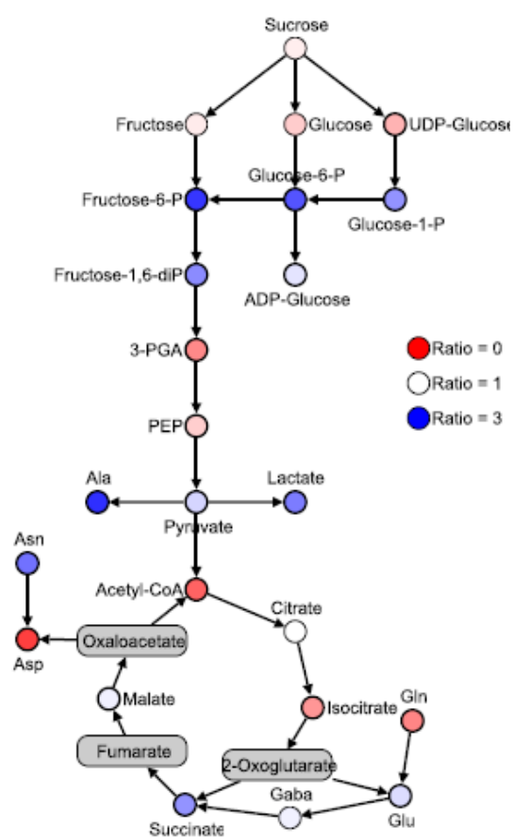
Methods: network integrated data visualization

- Multiple experiment datasets may be mapped to a single graph element, the same dataset may be mapped to multiple graph nodes

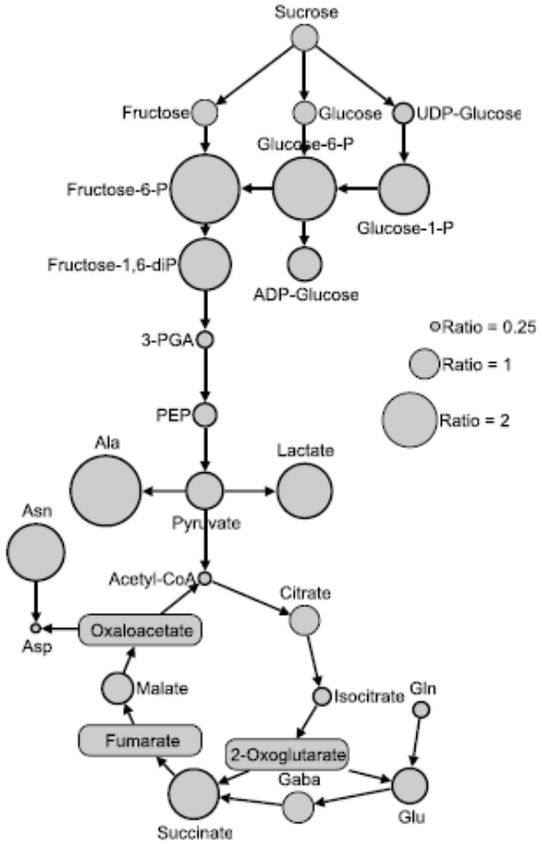


- Mapped data may can be visualized in different ways

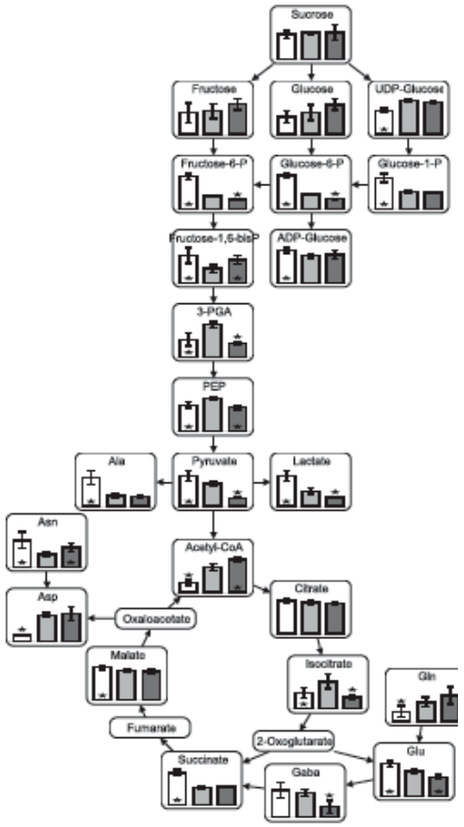




Color-coding
(or "heatmaps")



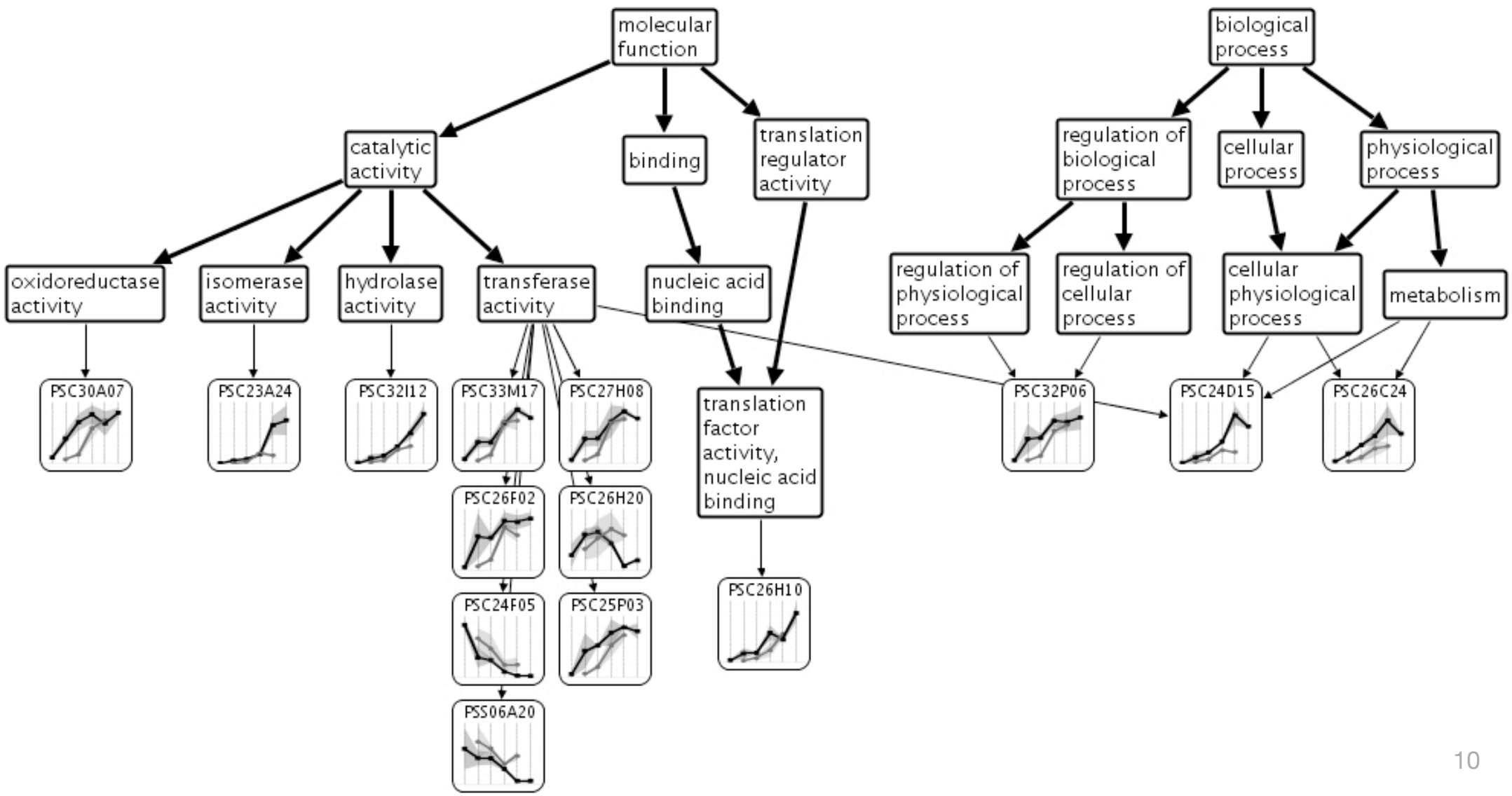
Shape-coding



Diagrams

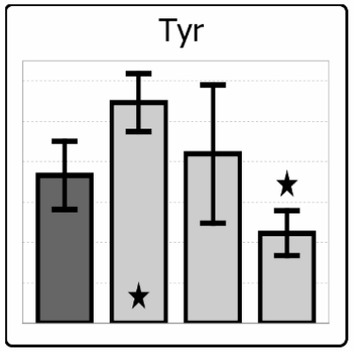
Methods: network integrated data visualization

- For data visualization in context of classification hierarchies, leaf-nodes for experiment data are created and connected to classification nodes

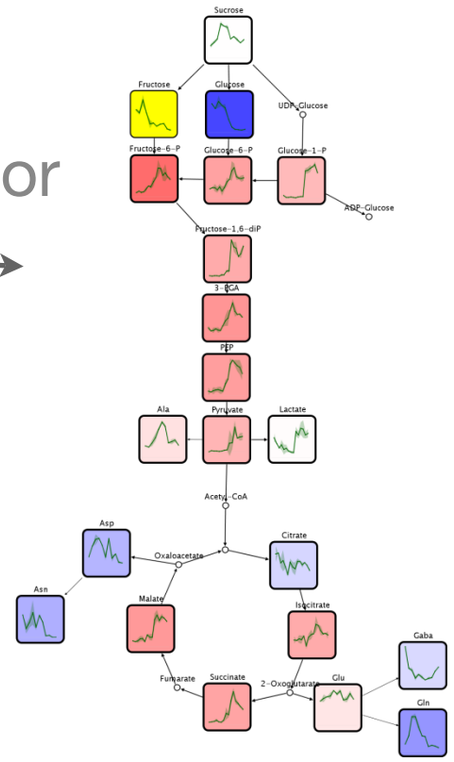


Methods: data analysis (statistics)

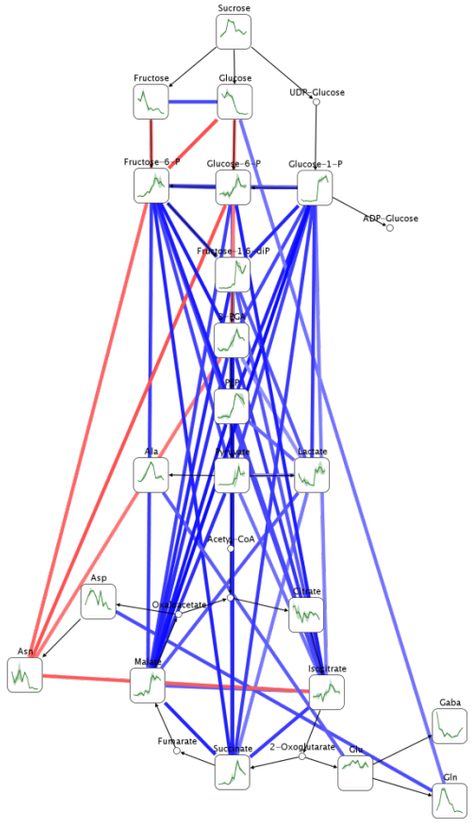
- Statistic tests to compare sample average values (e.g. *t*-Test, U-test, ratio comparison)



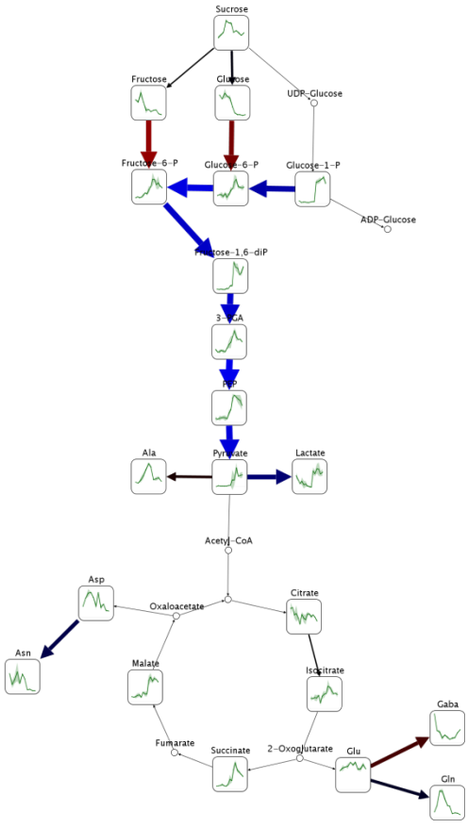
- Detect/remove outliers (Grubbs' test)
- Check for normal distribution (David quick test)
- Correlate time series or replicate data (using Pearson or Spearman correlation)



1:n



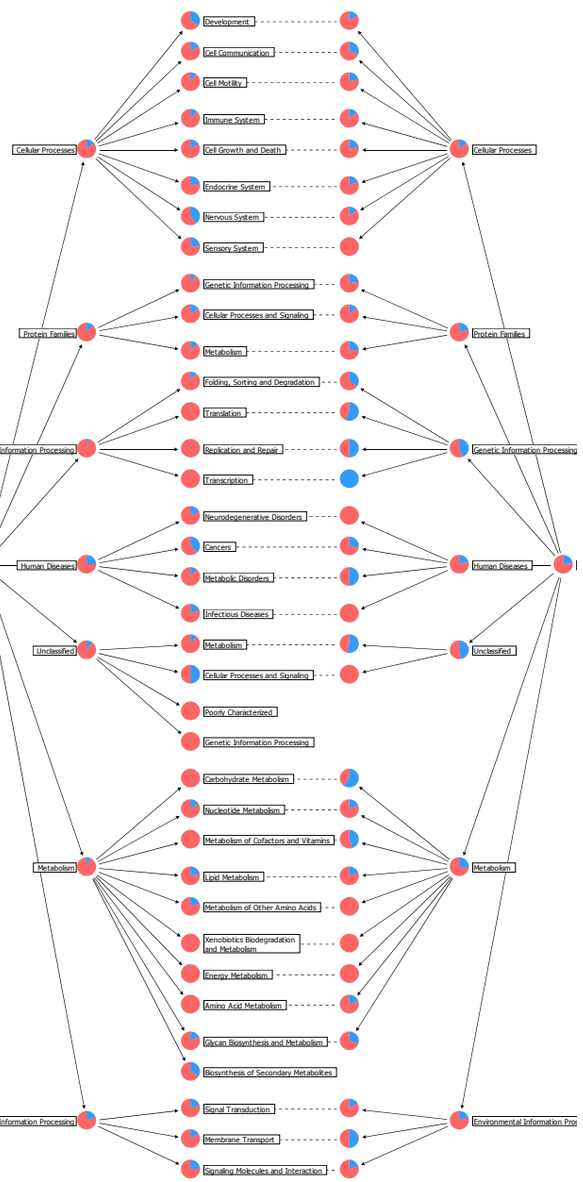
n:n



along path

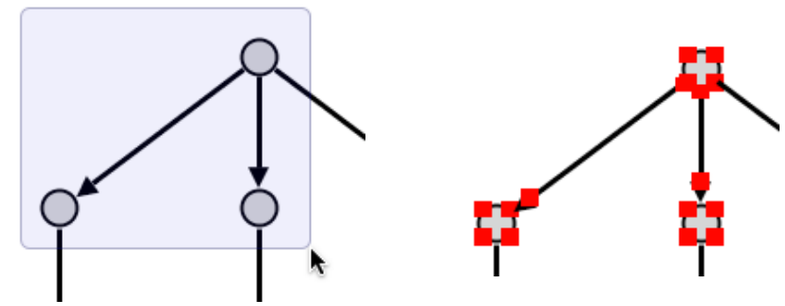
Methods: data analysis (statistics)

- Group data (e.g. up/down regulated genes), calculate histogram

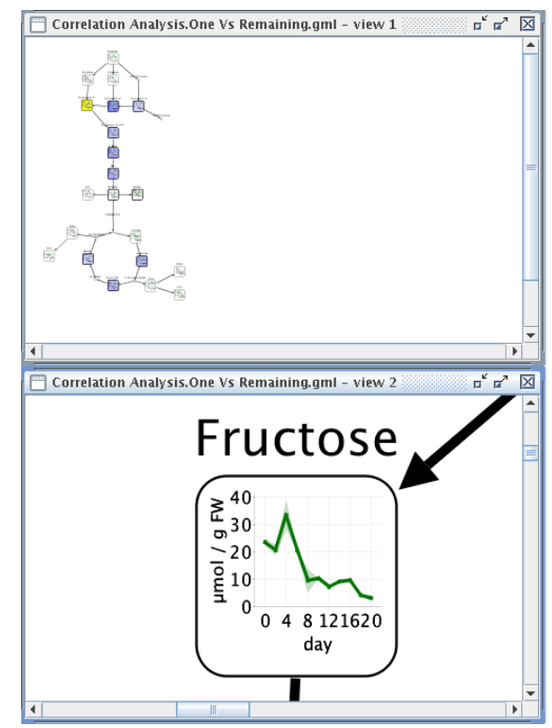


Methods: navigation and interaction

- Support for several interaction techniques
  - Direct selection
    Selection of a set of objects which are highlighted for visual investigation or which are the argument of a subsequent user interaction



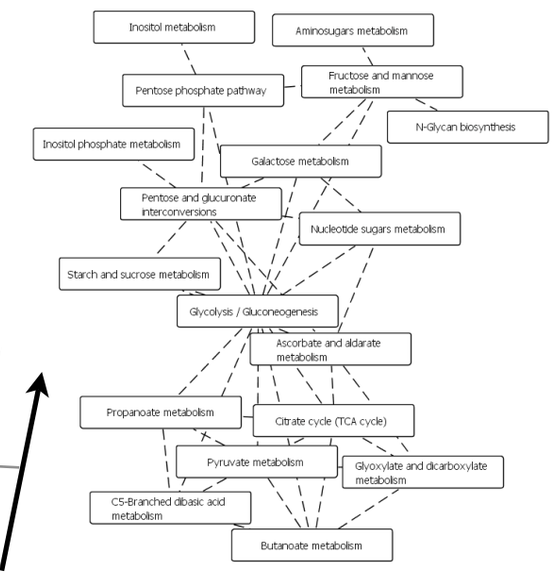
  - Details on demand
    Expansion of the visualization to show more details of a object, show/hide details of a visualization
  - Overview and detail
    A overview of data and a marked region is shown in one view, a second view shows the objects of the highlighted region in more detail

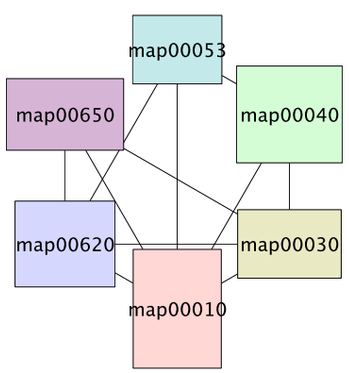


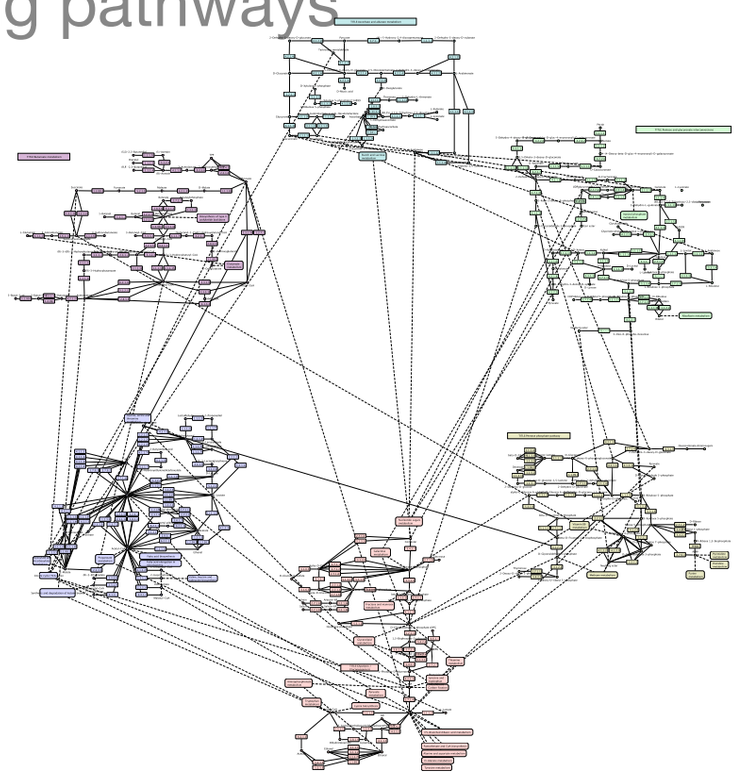
  - Dynamic queries
    Specification and combination of search-criteria for nodes/edges
  - Direct manipulation
    Modification of visualization attributes directly inside the view (node position, edge bends, node size)
  - Attribute walk
    Beginning with the current selection all nodes/edges with the same attribute values are added to the selection

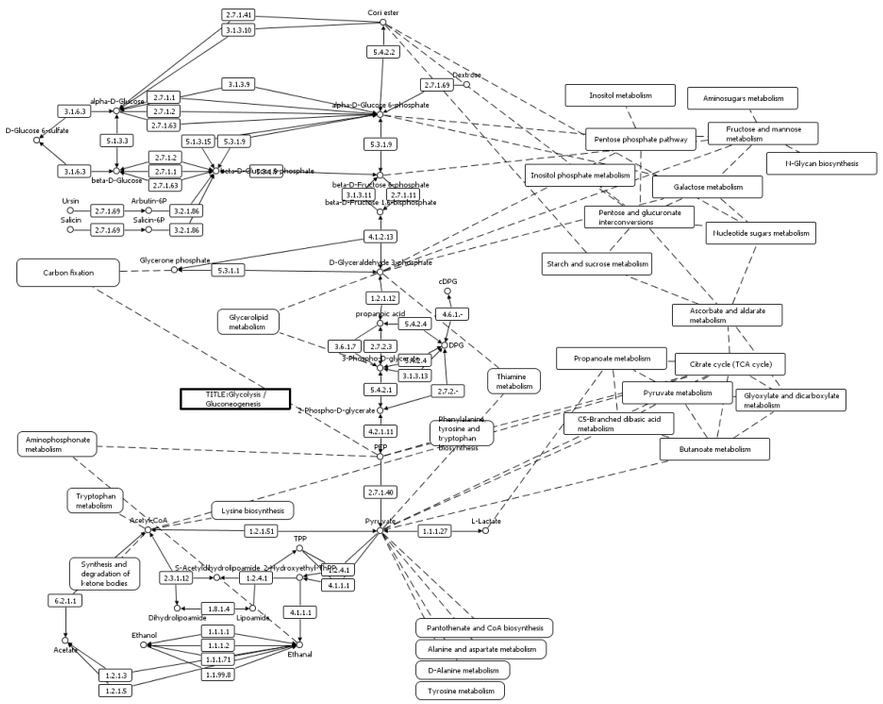
Methods: navigation and interaction

- (KEGG) Pathway navigation techniques
  - Extending the overview
  - Collapsing pathways
  - Arranging pathways



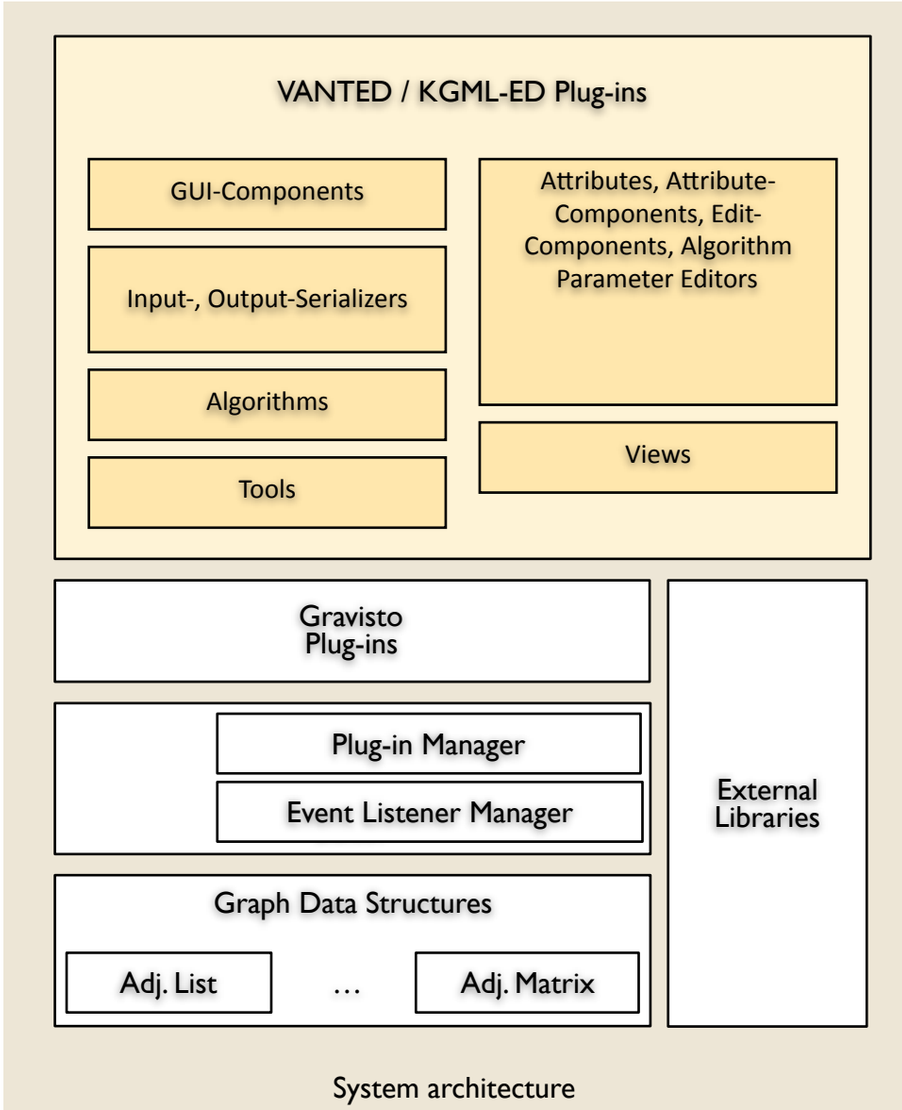






Implementation

- Implemented as open-source Java application
- VANTED is based on the extensible graph library and editor Gravisto, developed at University of Passau, Germany
- Systems architecture
  - Plugin-based extension mechanism
    - New ability to create plugin-jars, which are automatically loaded
    - Information about new/updated plugins via integrated RSS reader
  - Model-View-Controler concept
  - Observer design pattern for event management



System architecture

Implementation

- Development environment
  - Java 1.5 or higher, IDE: Eclipse, Mac/Linux/Windows
- Availability
  - Java Webstart (single-click installation and start)
    - http://vanted.ipk-gatersleben.de
  - Binary download (ZIP file, Mac OS X dmg image
    - http://sourceforge.net/projects/vanted
  - Complete source code
    - CVS Server at SourceForge

Example use cases

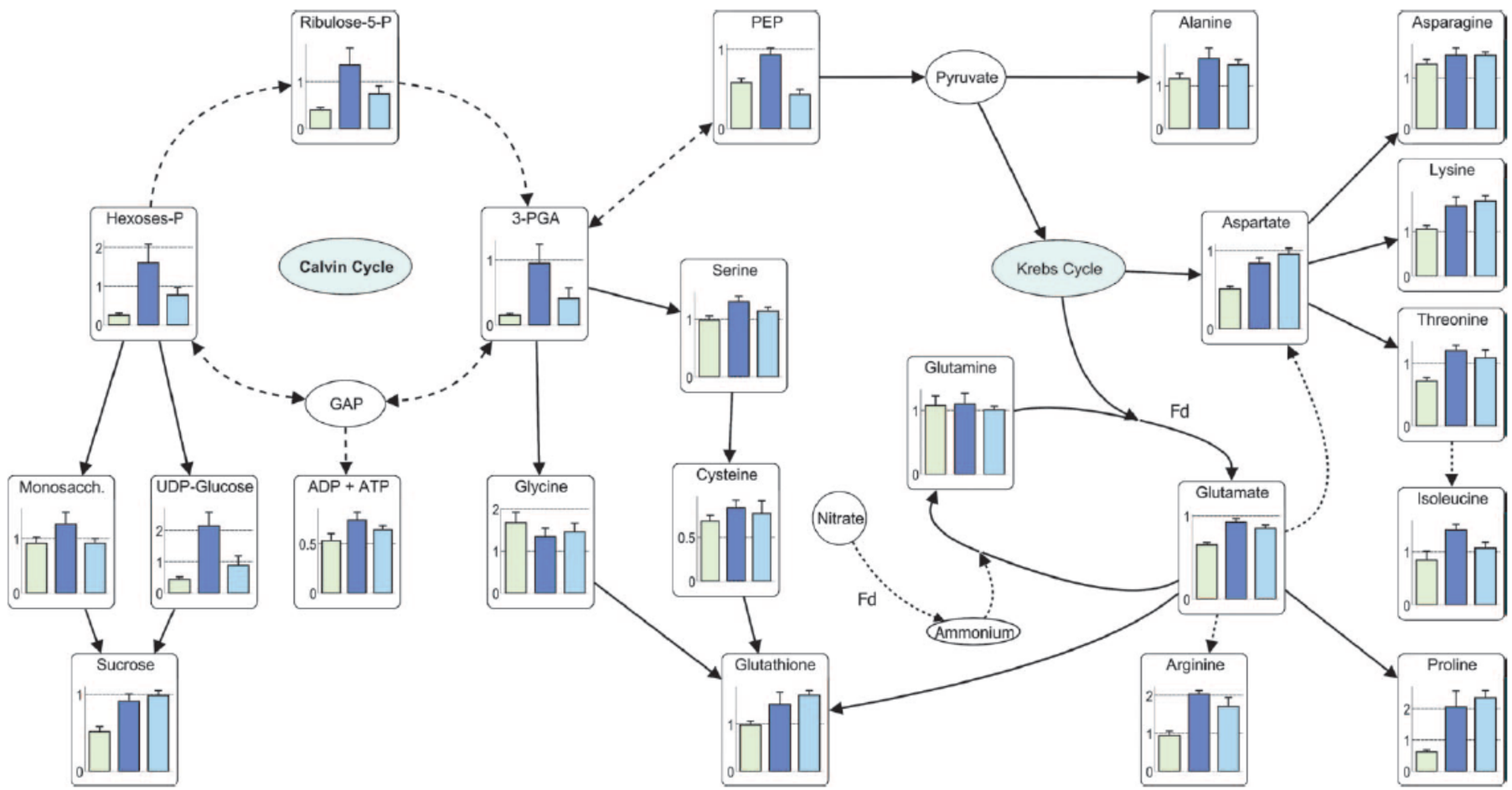
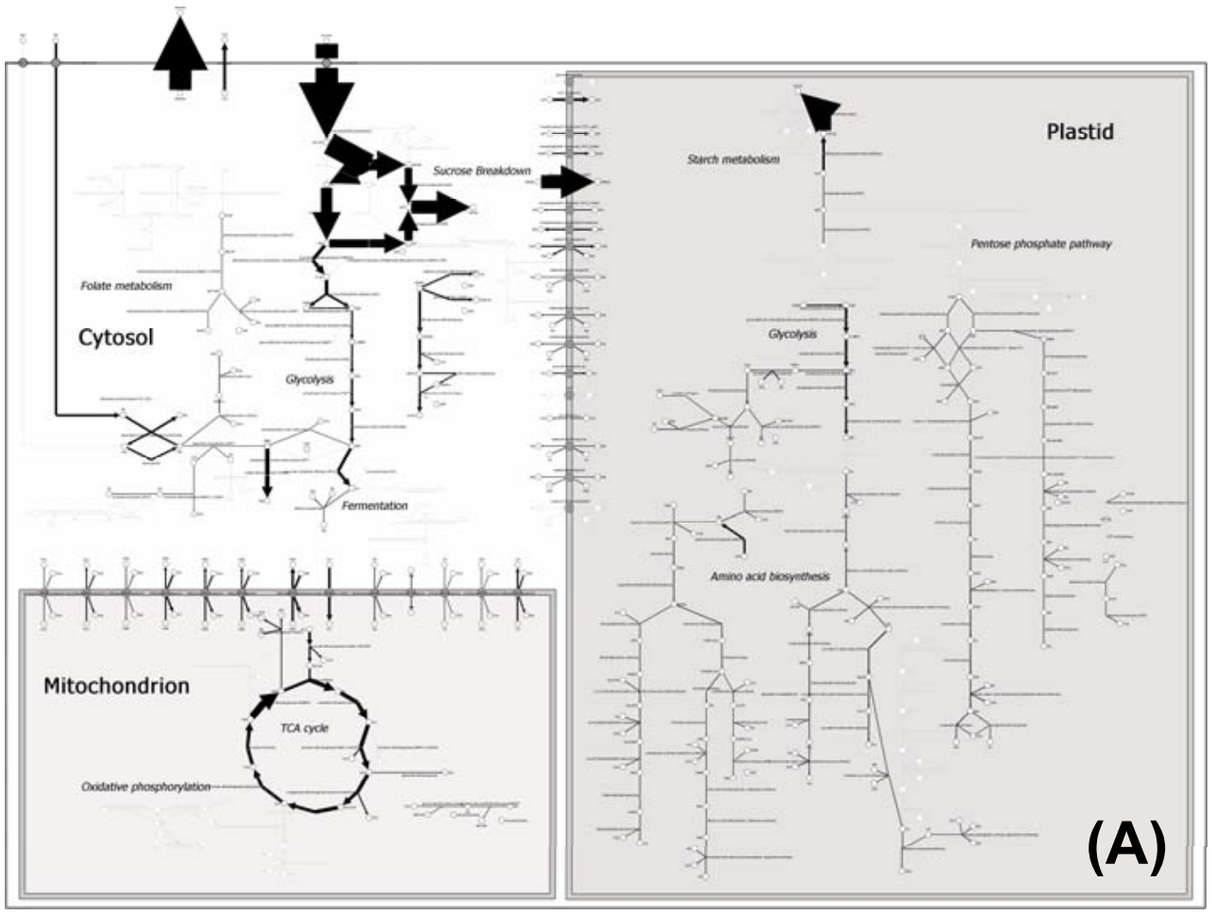
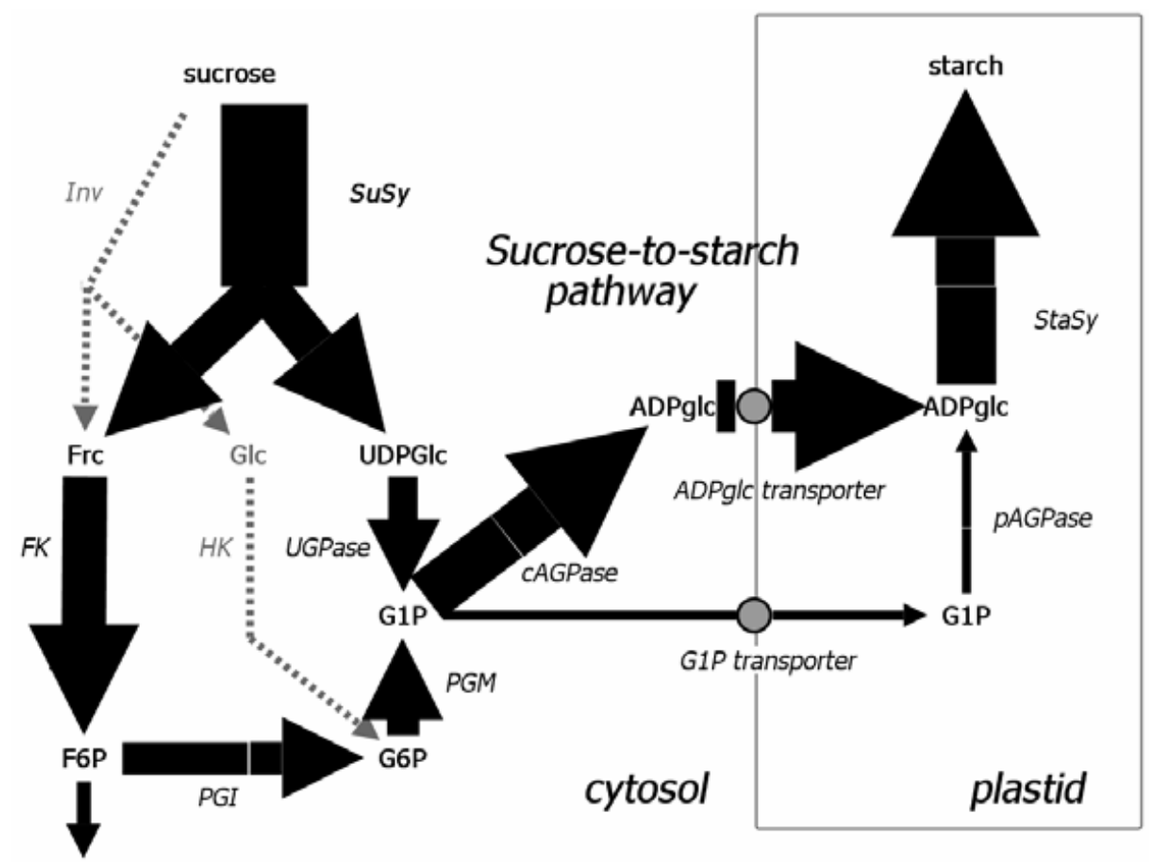


Fig. 4. Relative metabolite changes in iron-starved and control plants. Four-week-old plants were transferred to hydroponic Hoagland solution supplemented with either $FeSO_4$-EDTA or $CaCO_3$ (pH 8.0). Leaf material was harvested after 29 days, and the corresponding metabolites were measured as described in *Materials and Methods*. Depicted are the ratios ± SE of metabolite contents between Fe-starved and -replete plants of WT (green bars), *pfld*5-8 (blue bars), and *pfld*4-2 (light blue bars) lines (n = 8–10 independent plants). The graph was created by using the visualization system Vanted (38).

Tognetti et al. (2007): Enhanced plant tolerance to iron starvation by functional substitution of chloroplast ferredoxin with a bacterial flavodoxin. Proc. Natl. Acad. Sci. USA

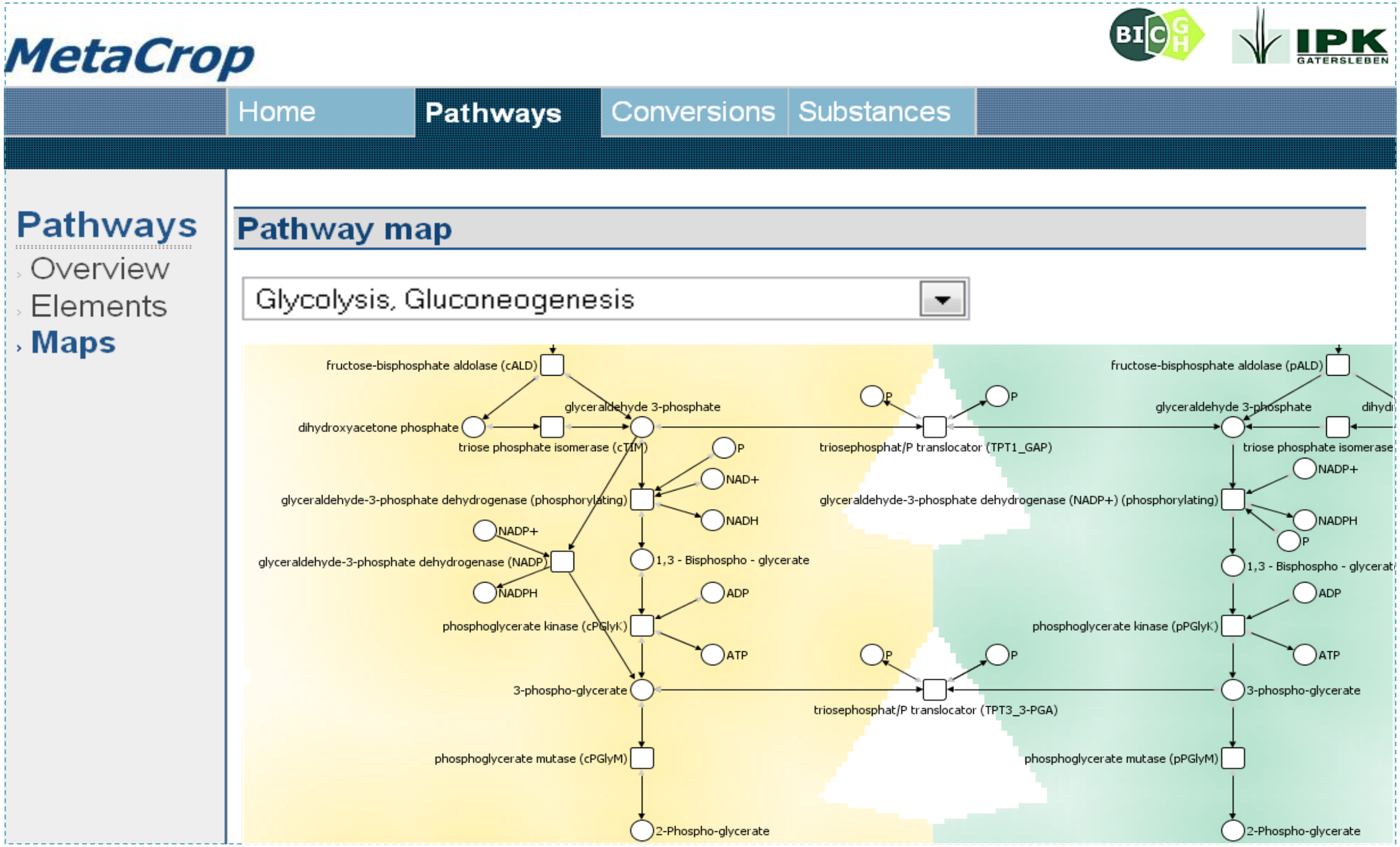
Example use cases





Grafahrend-Belau et al. (2008): Towards Systems Biology of Developing Barley Grains: a Framework for Modeling Metabolism. Proceedings Workshop on Computational Systems Biology (WCSB'08). *TICPS Series* 41: 41-44.

Example use cases



Grafahrend-Belau et al. (2008): MetaCrop – A detailed database for crop plant metabolism. Nucleic Acids Research

Thank you for your attention!

Literature (selection)

DBE information system

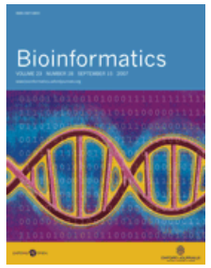
- Borisjuk, Hajirezaei, Klukas, Rolletschek, Schreiber (2004): Integrating data from biological experiments into metabolic networks with the DBE information system. In Silico Biology

VANTED

- Junker, Klukas, Schreiber (2006): VANTED: A system for advanced data analysis and visualization in the context of biological networks.
  BMC Bioinformatics

KEGG pathway navigation

- Klukas and Schreiber (2007): Dynamic exploration and editing of KEGG pathway diagrams.
  Bioinformatics



Internet

- http://vanted.ipk-gatersleben.de
- http://kgml-ed.ipk-gatersleben.de