Analysis and visualization of biological experiment data in the context of relevant networks

Christian Klukas

Plant Bioinformatics Group
Leibniz Institute of Plant Genetics and Crop Plant Research

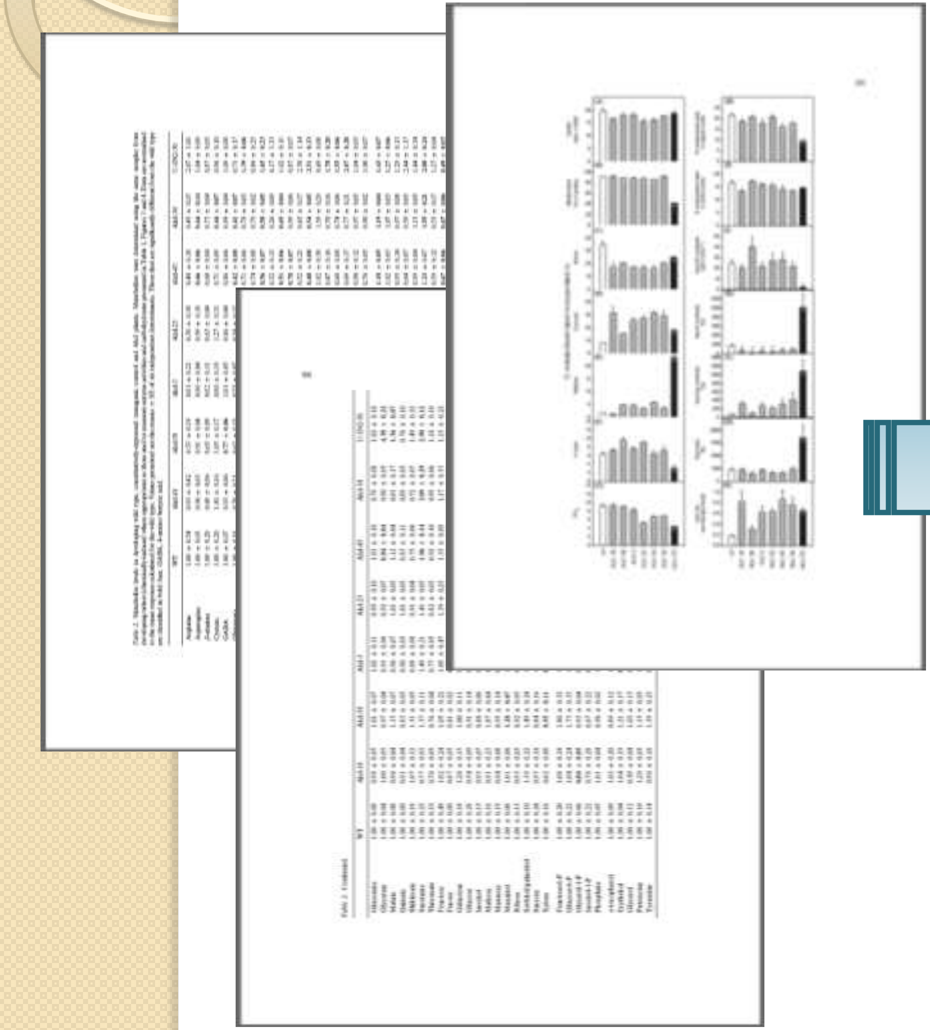
Progress Seminar Department Molecular Genetics

IPK-Gatersleben, January 23rd 2009

Outline

1. Motivation
2. Methods
    - Definition and combination of data models for experiment data and networks
    - Network-integrated visualization of experiment data
    - Interaction-, layout- and navigation-techniques
    - Statistical analysis
3. Implementation
4. Example use-cases (online-demo)

Motivation



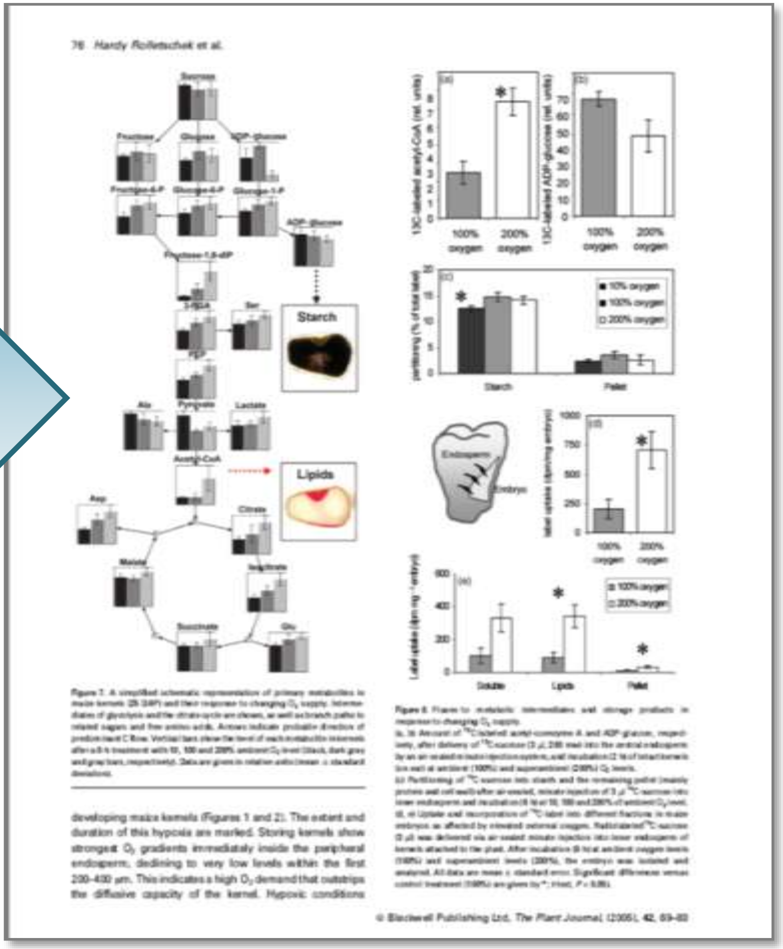


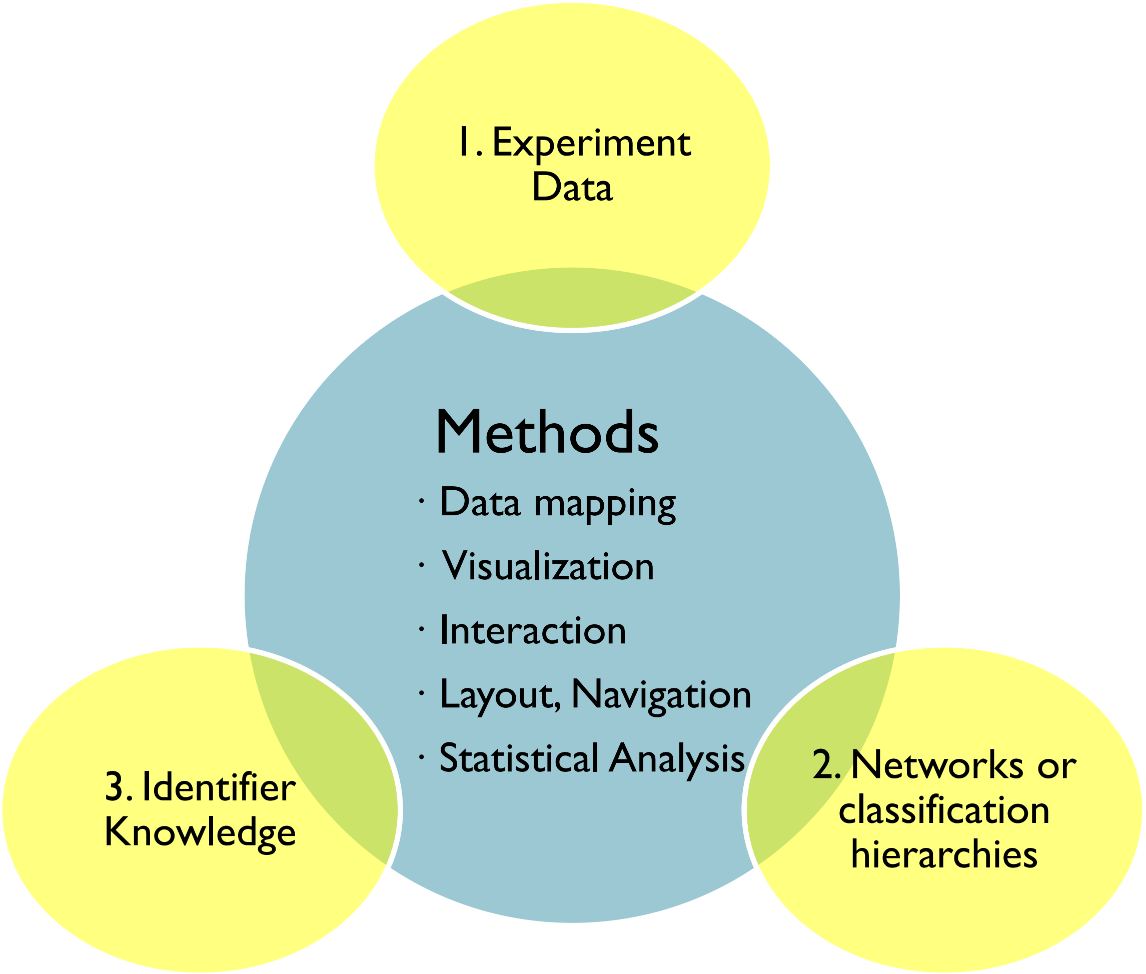
Figure 7. A simplified schematic representation of primary metabolites in maize kernels (25 DAP) and their response to changing O_2 supply. Intermediates of glycolysis and the citrate cycle are shown, as well as branch paths to related sugars and free amino acids. Arrows indicate probable direction of predominant C flow. Vertical bars show the level of each metabolite in kernels after a 8 h treatment with 10, 100 and 200% ambient O_2 level (black, dark grey and gray bars, respectively). Data are given in relative units (mean ± standard deviation).

Figure 8. Fluxes to metabolic intermediates and storage products in response to changing O_2 supply.
(a, b) Amount of ^{13}C-labeled acetyl-coenzyme A and ADP-glucose, respectively, after delivery of ^{13}C-sucrose (3 µl, 200 mM) into the central endosperm by an air-sealed micro-injection system, and incubation (2 h) of intact kernels (on ear) at ambient (100%) and supra-ambient (200%) O_2 levels.
(c) Partitioning of ^{14}C-sucrose into starch and the remaining pellet (mainly proteins and cell walls) after air-sealed, micro-injection of 3 µl ^{14}C-sucrose into lower endosperm and incubation (8 h) at 10, 100 and 200% of ambient O_2 level.
(d, e) Uptake and incorporation of ^{14}C-label into different fractions in maize embryos as affected by elevated external oxygen. Radiolabeled ^{14}C-sucrose (3 µl) was delivered via air-sealed micro injection into lower endosperm of kernels attached to the plant. After incubation (8 h) at ambient oxygen levels (100%) and supra-ambient levels (200%), the embryo was isolated and analyzed. All data are mean ± standard error. Significant differences versus control treatment (100%) are given by *; t-test, $P < 0.05$).

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

Methods' Requirements



1. Experiment Data

Methods

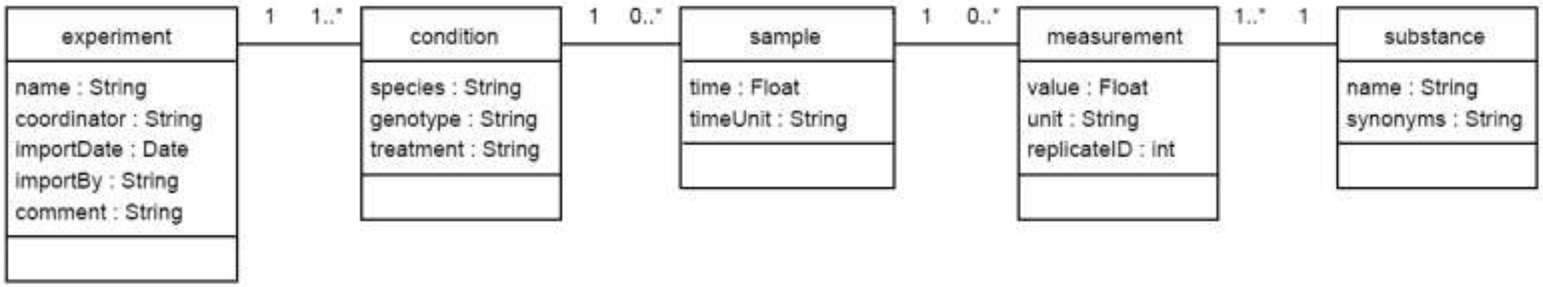
- Data mapping
- Visualization
- Interaction
- Layout, Navigation
- Statistical Analysis

3. Identifier Knowledge

2. Networks or classification hierarchies

Methods: Data model for experiment data

- Different standards for biological domains:
  - Genomics: MIAME / MAGE
  - Proteomics: PEDRo
  - Metabolomics: ArMet
- Existing models differ greatly, contain all kinds of annotations, new model:
  - Flexible model for all of the domains
  - Only data for visualization, analysis and identification



experiment

name : String
coordinator : String
importDate : Date
importBy : String
comment : String

condition

species : String
genotype : String
treatment : String

sample

time : Float
timeUnit : String

measurement

value : Float
unit : String
replicateID : int

substance

name : String
synonyms : String

Methods: Data model for biological networks and classification hierarchies

- (Generic) Mapping-Graph MG=(V, E, l, t_V, $t_{E,}$ z)
  - V – set of nodes
  - E – set of edges (directed or undirected)
  - l – function node/edge → label (L)
  - t_V, t_E – function node/edge → node/edge type (T$_V$, T$_E$)
  - z – function node/edge → mapped experiment data

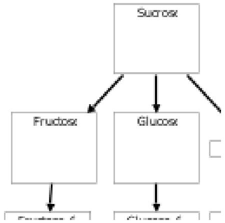
Methods: Data model for biological networks and classification hierarchies

- Protein-Protein Network - MG_{PPI}
  - Nodes represent proteins, T_V={protein}
  - Undirected edges represent interaction between two proteins, T_E={interaction}
- KEGG Pathway – MG_{KEGG}
  - T_V={Orthologe, Enzyme, Gene, Gene-Group, Metabolite, Map-Link}
  - T_E={ECrel, PPrel, GErel, PCrel, rProd, rSub, link} (enzyme-enzyme relation, protein-protein relation, gene-expression, protein-metabolite relation, reaction product, reaction substrate, link to pathway)
- (extended) Pathway-Overview Graph – MG_O, MG_{OE}
- Gene Ontology Hierarchy – MG_{GO}
- KEGG BRITE Hierarchy - MG_{BRITE}

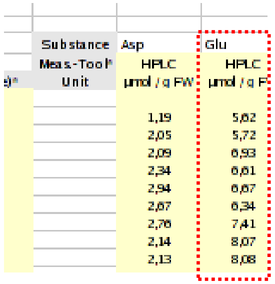
Methods: Data-mapping – combination of network and experiment data

Data

Graph
(Nodes & Edges)



Dataset



		Substance	Asp	Glu
		Meas-Tool	HPLC	HPLC
	e)	Unit	µmol / g FW	µmol / g F
			1,19	5,82
			2,05	5,72
			2,09	6,93
			2,34	6,81
			2,94	6,67
			2,67	6,34
			2,76	7,41
			2,14	8,07
			2,13	8,08

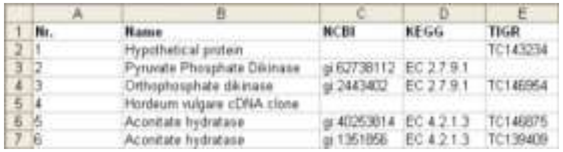
Synonyms

KEGG Compounds

Expasy Enzymes

KEGG KO

Alternative Identifiers

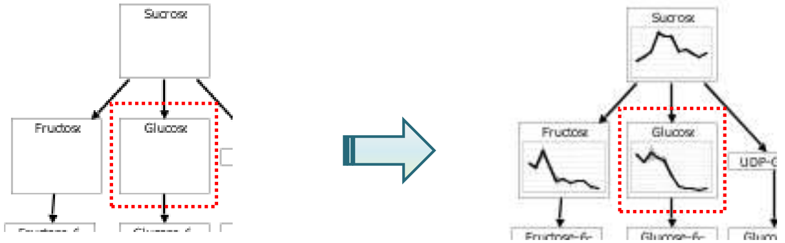


	A	B	C	D	E
1	Nr.	Name	NCBI	KEGG	TIGR
2	1	Hypothetical protein			TC143234
3	2	Pyruvate Phosphate Dikinase	gi 62738112	EC 2.7.9.1	
4	3	Orthophosphate dikinase	gi 2443402	EC 2.7.9.1	TC146954
5	4	Hordeum vulgare cDNA clone			
6	5	Aconitate hydratase	gi 40253814	EC 4.2.1.3	TC146875
7	6	Aconitate hydratase	gi 1351856	EC 4.2.1.3	TC139409

Pseudo code data m.

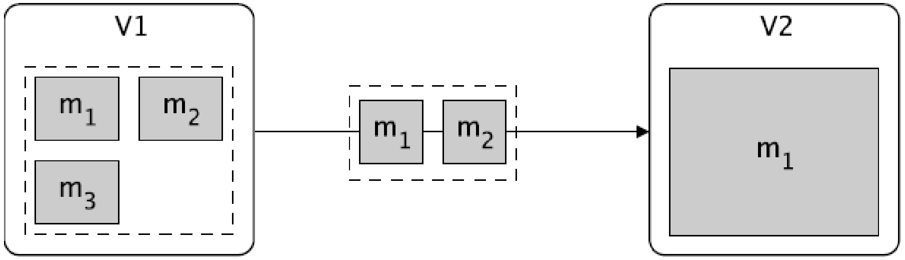
```
for each substance in dataset do
    Set of Substance IDs = expandWithDBids(substance.name,
                                          substance.synonyms)
    for each graphelment in targetGraph do
        Set of Graphelement IDs = expandWithDBids(label)
        if intersection(substance IDs, Graphelement IDs).length > 0 then
            graphelment.assignData(dataset.getSubsetFor(substance))
    end for
end for
```

Result

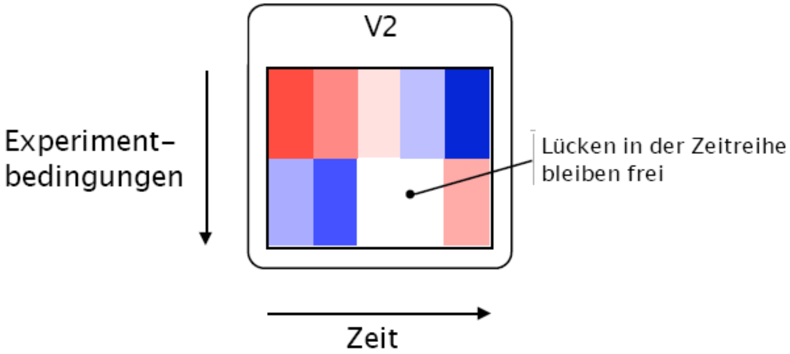


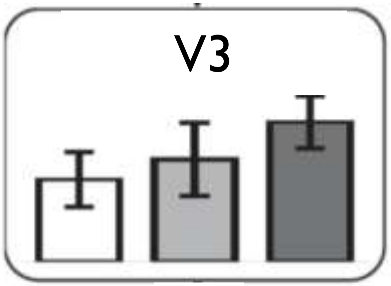
Methods: Network integrated data visualization

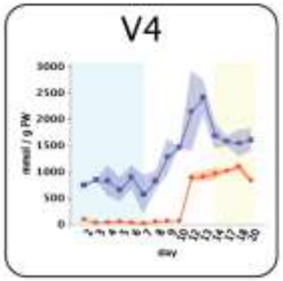
Processing of multiple mappings →

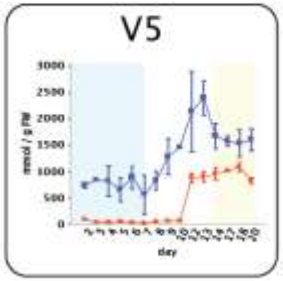


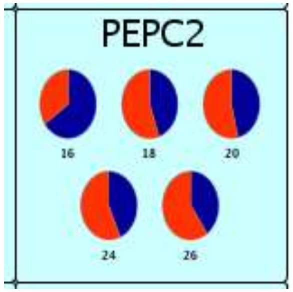
Data visualization for conditions and time →



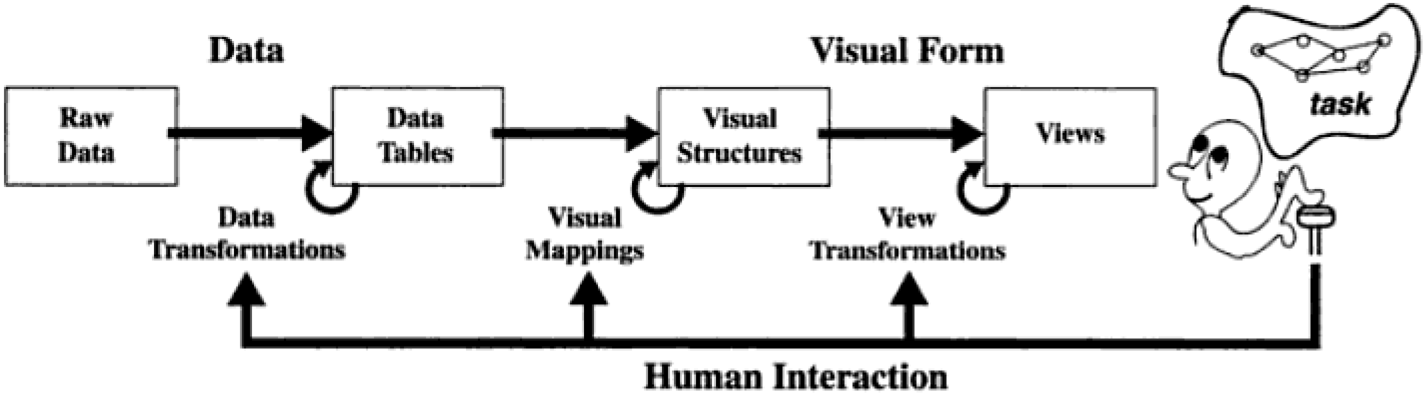








Methods: Interaction techniques

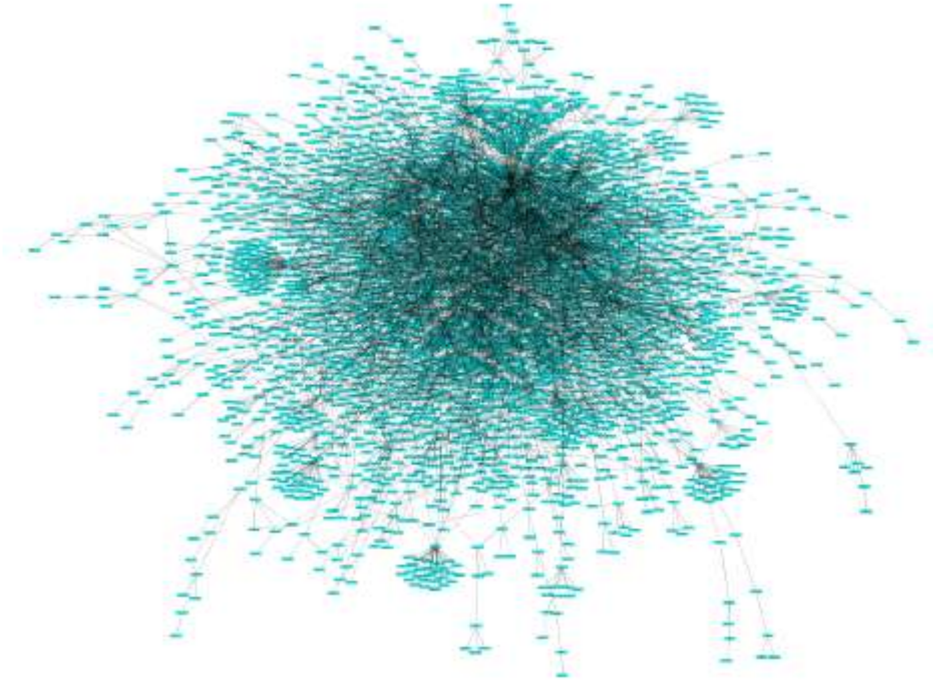


Visualization pipeline (Card, Mackinlay, Shneiderman, 1999)

- Data transformations
  - Dynamic queries (select conditions, search graph elements)
  - Direct walk (pathway navigation)
  - Attribute walk (select similar elements)
  - Details-on-demand (show/hide details)
  - Direct manipulation (node position, label)
- Visual mappings
  - Experiment data display: color-coded, size-/width-coded, diagrams
- View transformations
  - Direct selection (click & select)
  - Overview-and-Detail (multiple views with varying detail)

Methods: Layout of specific mapping-graphs

- MG_{PPI}: Force-Directed (Spring-Embedder)
- MG_{GO}, MG_{BRITE}: Hierarchical (Sugiyama)
- MG_{KEGG}: Manual layout is given

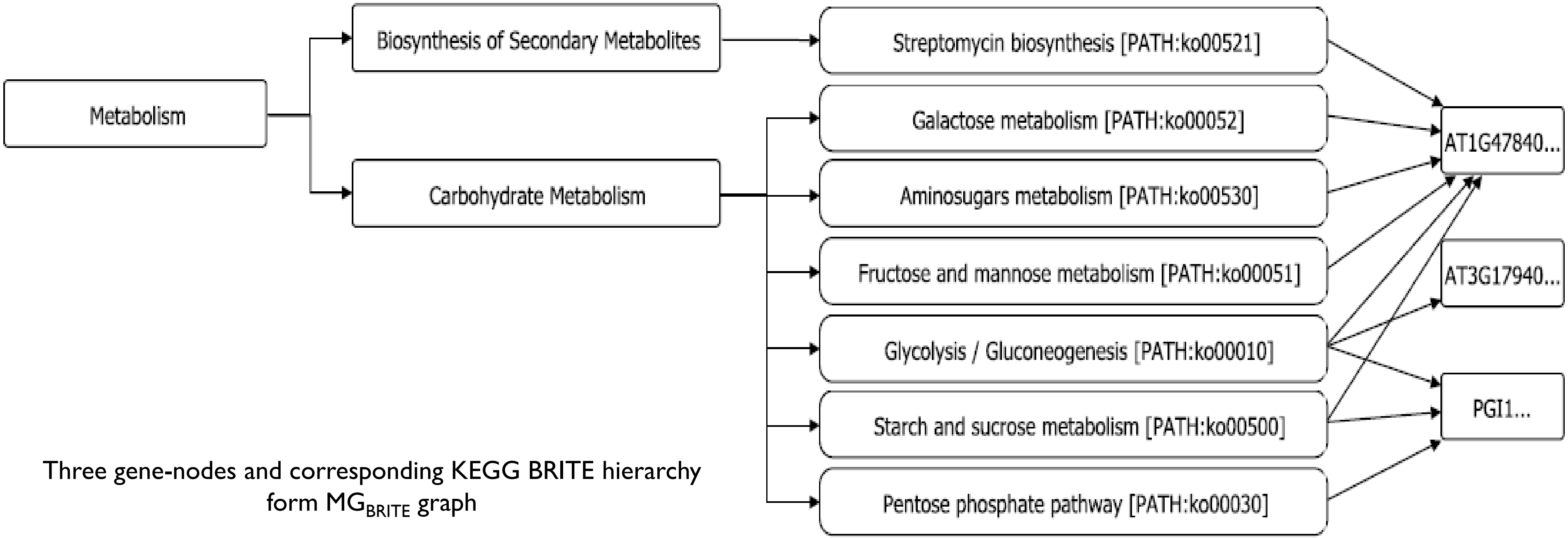


PPI network of worm C.elegans (5418 interactions, 2992 proteins). Datasource: GraphWeb

Methods: Layout of specific mapping-graphs

- MG_{PPI}: Force-Directed (Spring-Embedder)

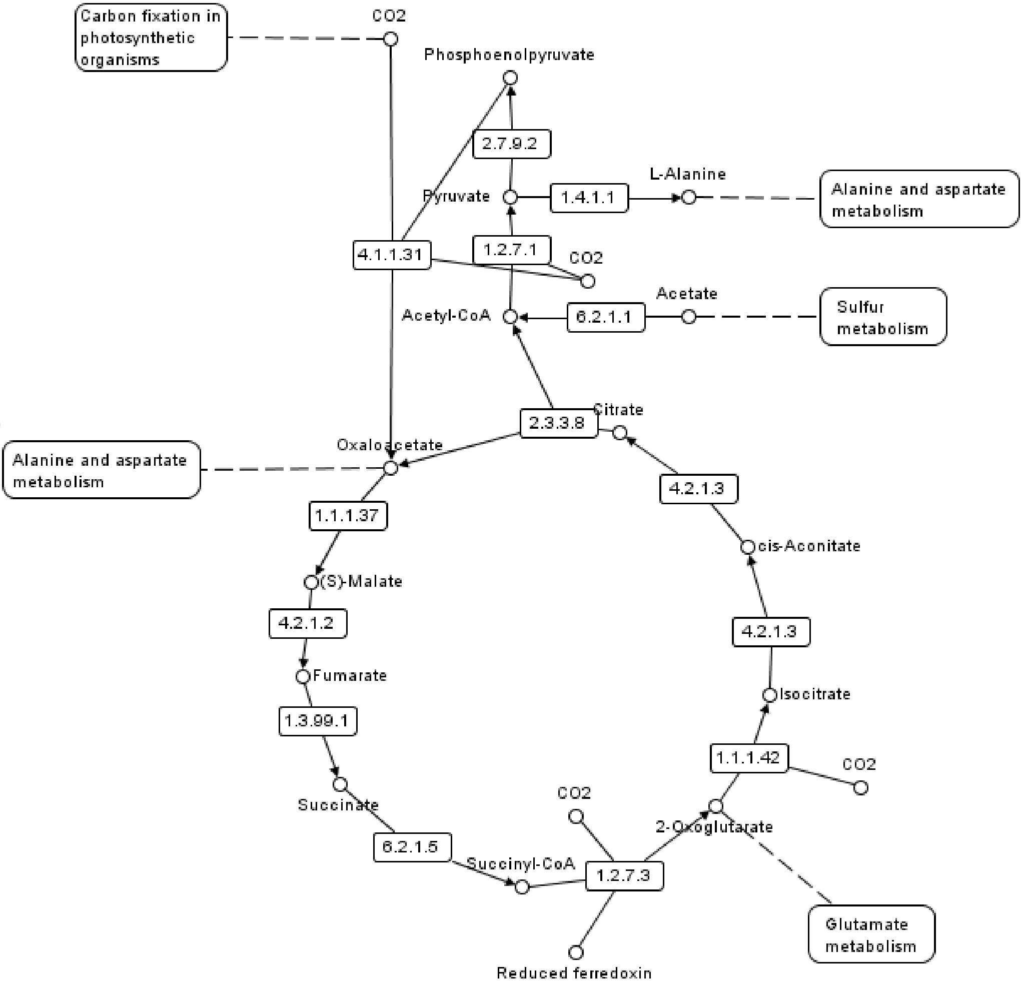
- MG_{GO}, MG_{BRITE}: Hierarchical (Sugiyama)


Metabolism
Biosynthesis of Secondary Metabolites
Carbohydrate Metabolism
Streptomycin biosynthesis [PATH:ko00521]
Galactose metabolism [PATH:ko00052]
Aminosugars metabolism [PATH:ko00530]
Fructose and mannose metabolism [PATH:ko00051]
Glycolysis / Gluconeogenesis [PATH:ko00010]
Starch and sucrose metabolism [PATH:ko00500]
Pentose phosphate pathway [PATH:ko00030]
AT1G47840...
AT3G17940...
PGI1...

Three gene-nodes and corresponding KEGG BRITE hierarchy form MG_{BRITE} graph

Methods: Layout of specific mapping-graphs

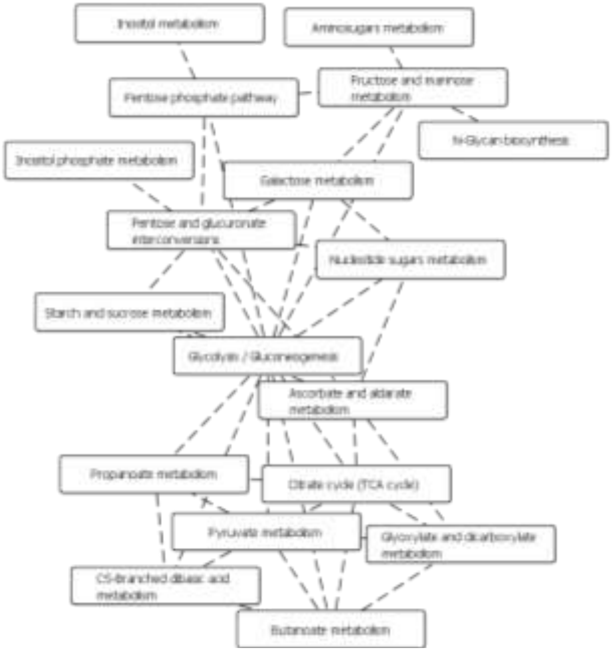
- MG_{PPI}: Force-Directed (Spring-Embedder)
- MG_{GO}, MG_{BRITE}: Hierarchical (Sugiyama)
- MG_{KEGG}: Manual layout is given

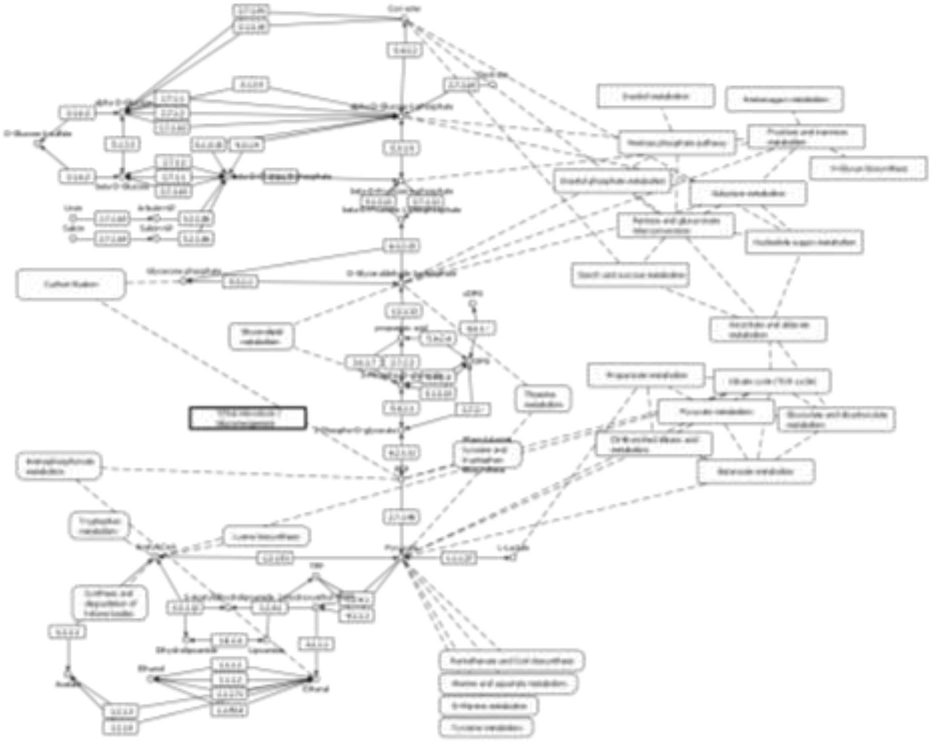


KEGG Pathway „Reductive carboxylate cycle (CO2 fixation)“ , visualization: VANTED

Methods: Pathway-Navigation

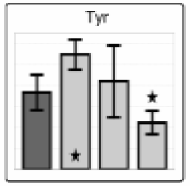
- Extending the overview
- Collapsing pathways
- Arranging pathways
- Stepwise pathway-navigation

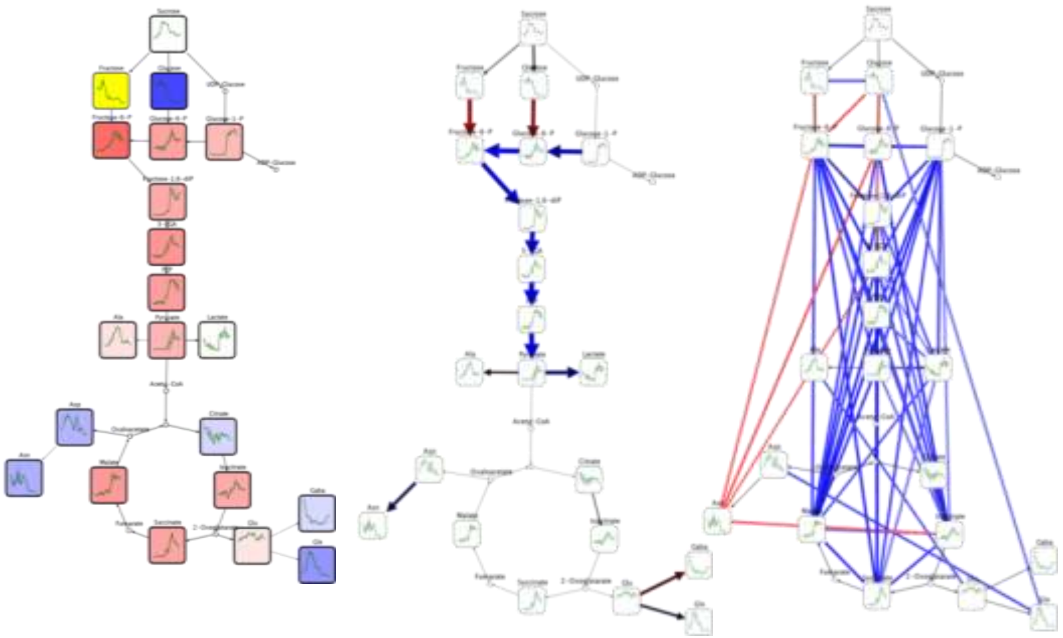

Inositol metabolism
Aminosugars metabolism
Pentose phosphate pathway
Fructose and mannose metabolism
N-Glycan biosynthesis
Inositol phosphate metabolism
Galactose metabolism
Pentose and glucuronate interconversions
Nucleotide sugars metabolism
Starch and sucrose metabolism
Glycolysis / Gluconeogenesis
Ascorbate and aldarate metabolism
Propanoate metabolism
Citrate cycle (TCA cycle)
Pyruvate metabolism
Glyoxylate and dicarboxylate metabolism
C5-Branched dibasic acid metabolism
Butanoate metabolism

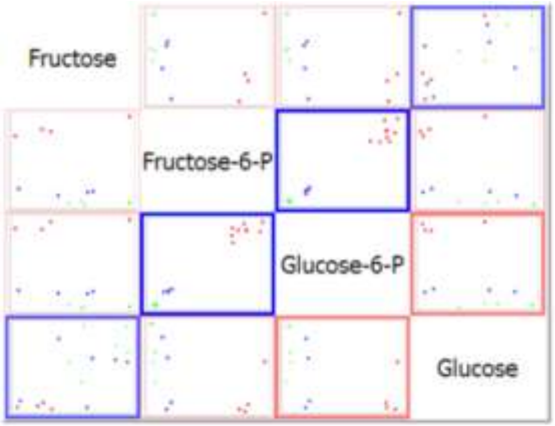


Methods: Statistical analysis

- Statistic tests: t-Test (2 variants), Grubbs-Test, David-Quicktest
- Flexible calculation of correlations
- Significance analysis for MG_{KEGG} and MG_{BRITE} (Fisher's exact test)
- Scatter-Plots

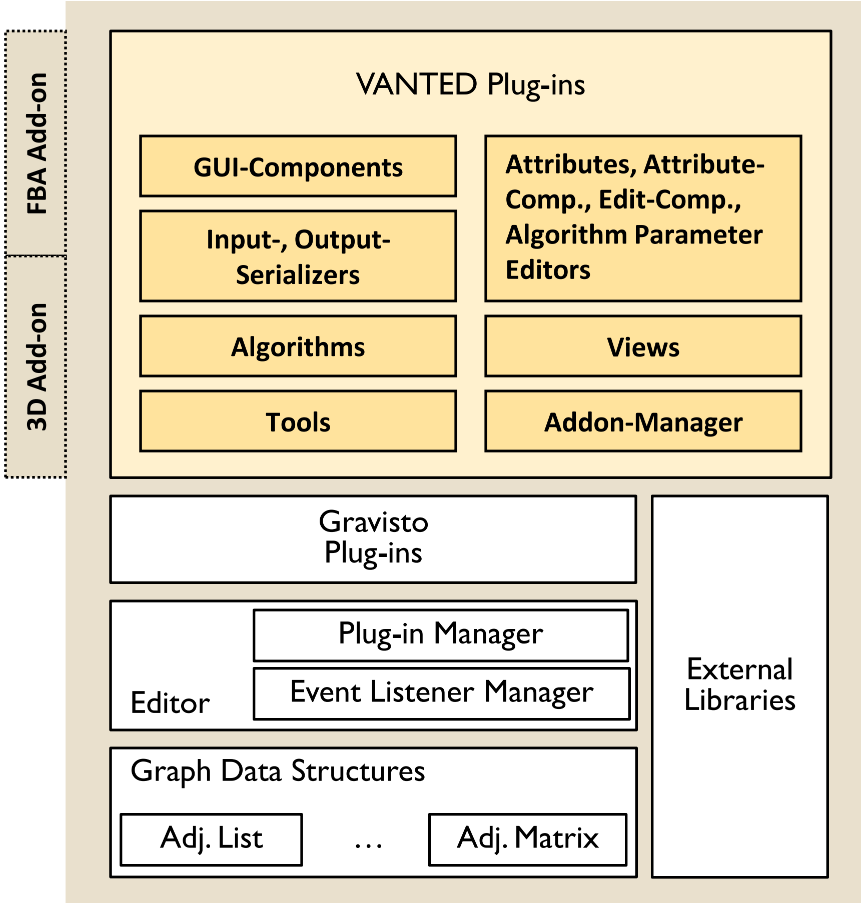






Implementation

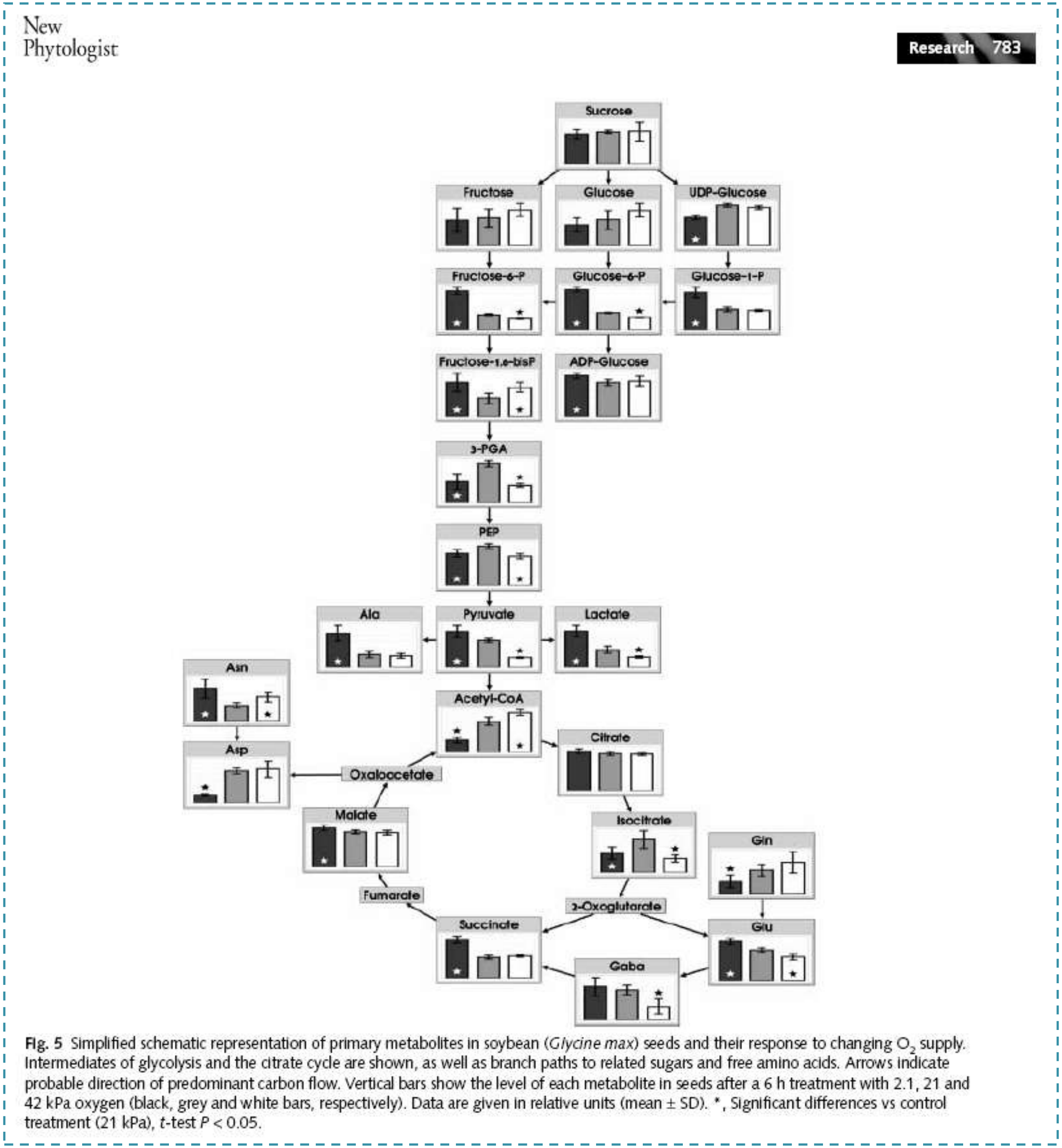
- Based on the extensible, plugin-based graph visualization toolkit **Gravisto** (developed at the University of Passau and at the IPK)
- MVC pattern
- Event management (observer design pattern)
- Java application (Windows exe, Mac OS X image, platform-neutral ZIP, Java WebStart)
- External plug-ins ("Add-ons")



System architecture

Visualization of relative metabolite changes under different stress situations

Rolletschek et al. (2005): Regulation of lipid biosynthesis in soybean seeds: evidence for a key role of photosynthetic oxygen release. **New Phytologist**



New Phytologist

Research 783

Fig. 5 Simplified schematic representation of primary metabolites in soybean (*Glycine max*) seeds and their response to changing O_2 supply. Intermediates of glycolysis and the citrate cycle are shown, as well as branch paths to related sugars and free amino acids. Arrows indicate probable direction of predominant carbon flow. Vertical bars show the level of each metabolite in seeds after a 6 h treatment with 2.1, 21 and 42 kPa oxygen (black, grey and white bars, respectively). Data are given in relative units (mean ± SD). *, Significant differences vs control treatment (21 kPa), *t*-test $P < 0.05$.

Visualization of relative metabolite changes under different stress situations

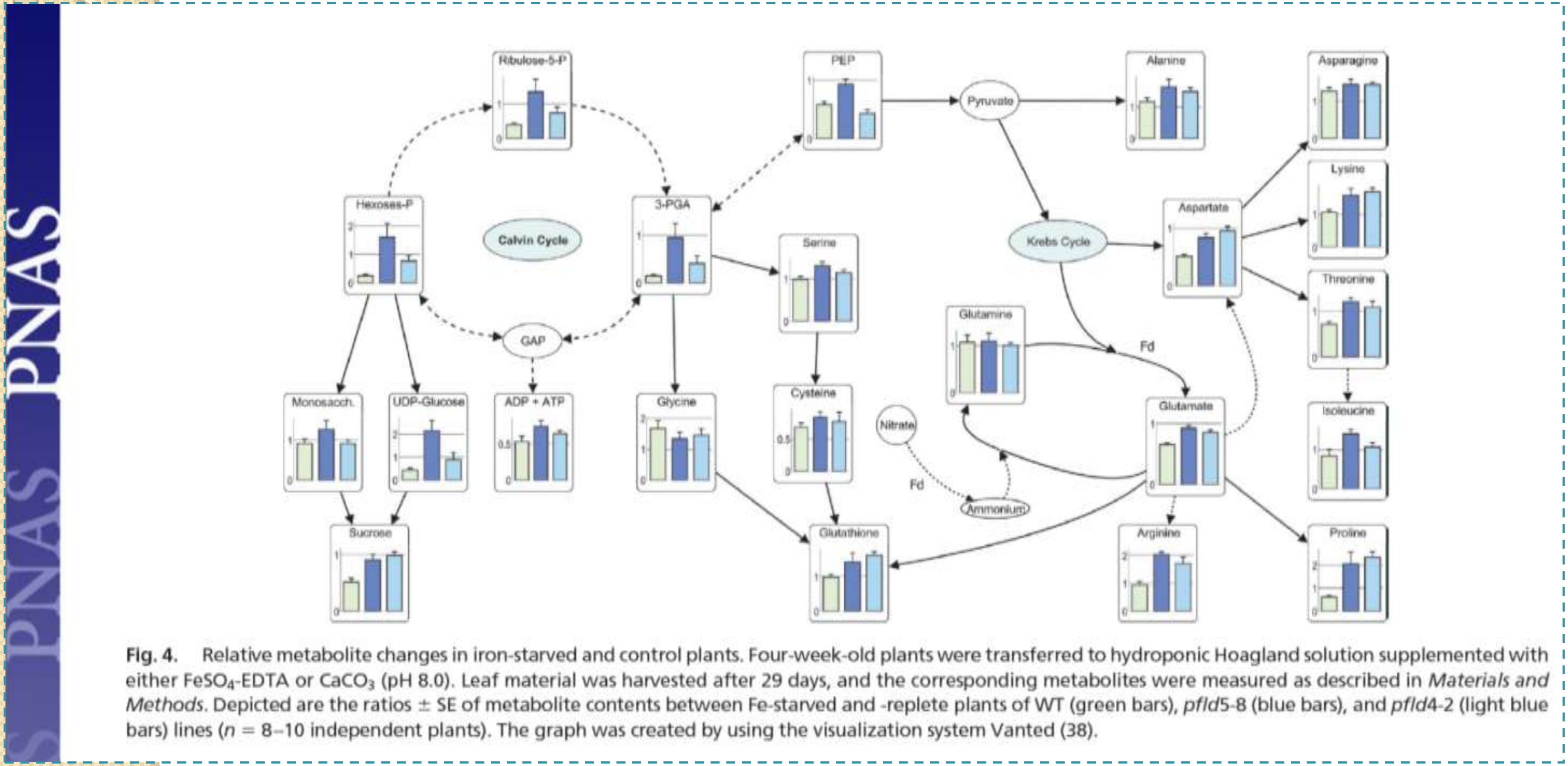
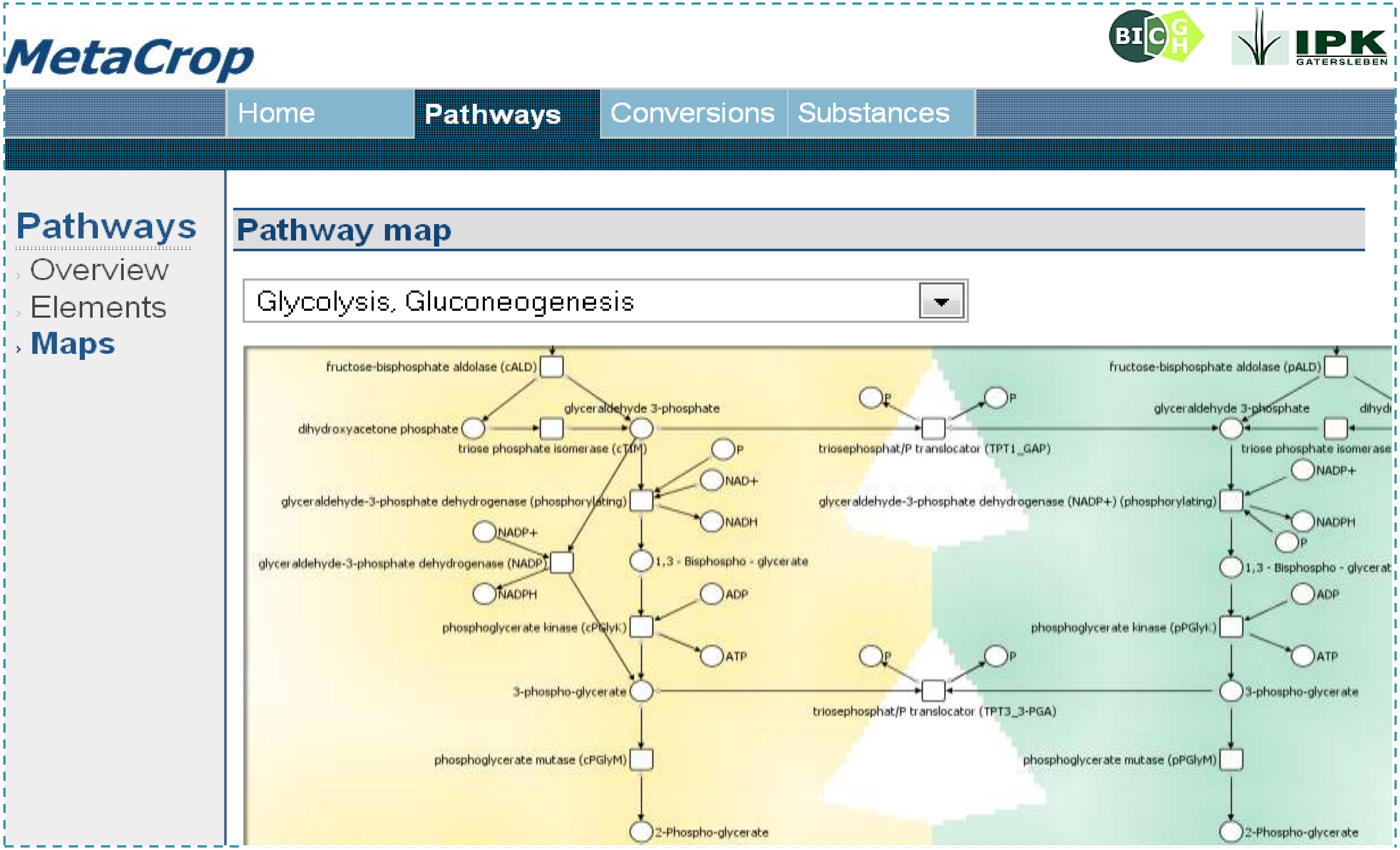


Fig. 4. Relative metabolite changes in iron-starved and control plants. Four-week-old plants were transferred to hydroponic Hoagland solution supplemented with either $FeSO_4$-EDTA or $CaCO_3$ (pH 8.0). Leaf material was harvested after 29 days, and the corresponding metabolites were measured as described 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).

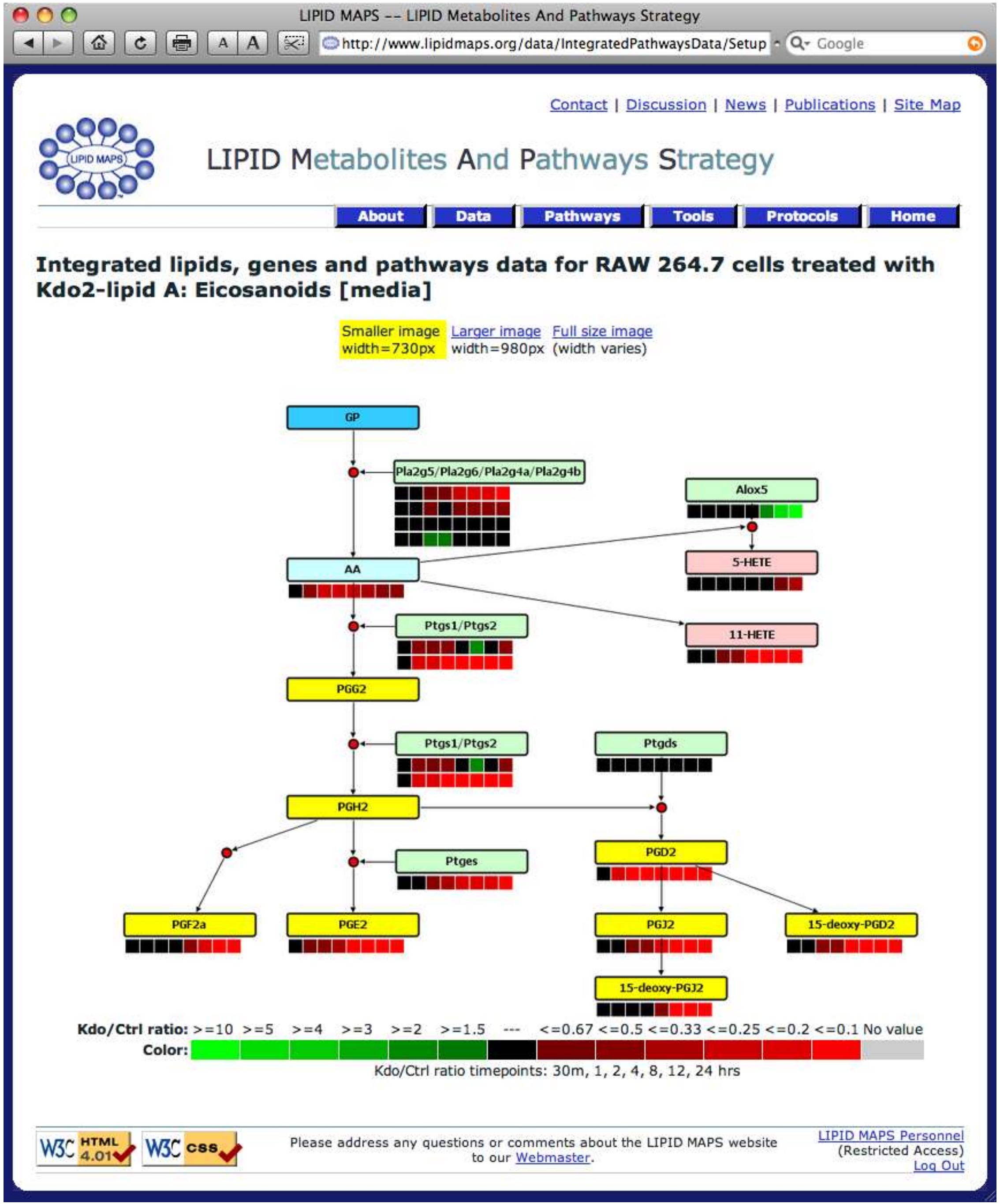
Tognetti et al. (2007): *Enhanced plant tolerance to iron starvation by functional substitution of chloroplast ferredoxin with a bacterial flavodoxin.* **Proc. Natl. Acad. Sci. USA**

Interactive network visualization for internet database systems


MetaCrop
BIC GH
IPK GATERSLEBEN
Home
Pathways
Conversions
Substances
Pathways
Overview
Elements
Maps
Pathway map
Glycolysis, Gluconeogenesis
fructose-bisphosphate aldolase (cALD)
glyceraldehyde 3-phosphate
dihydroxyacetone phosphate
triose phosphate isomerase (cTIM)
P
NAD+
glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)
NADH
NADP+
glyceraldehyde-3-phosphate dehydrogenase (NADP)
1,3 - Bisphospho - glycerate
NADPH
ADP
phosphoglycerate kinase (cPGlyK)
ATP
3-phospho-glycerate
phosphoglycerate mutase (cPGlyM)
2-Phospho-glycerate
triosephosphat/P translocator (TPT1_GAP)
triosephosphat/P translocator (TPT3_3-PGA)
fructose-bisphosphate aldolase (pALD)
glyceraldehyde 3-phosphate
dihydr
triose phosphate isomerase
NADP+
glyceraldehyde-3-phosphate dehydrogenase (NADP+) (phosphorylating)
NADPH
P
1,3 - Bisphospho - glycerat
ADP
phosphoglycerate kinase (pPGlyK)
ATP
3-phospho-glycerate
phosphoglycerate mutase (pPGlyM)
2-Phospho-glycerate

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

Pathway editing, data visualization



LIPID MAPS -- LIPID Metabolites And Pathways Strategy

http://www.lipidmaps.org/data/IntegratedPathwaysData/Setup

Google

Contact | Discussion | News | Publications | Site Map

LIPID Metabolites And Pathways Strategy

Integrated lipids, genes and pathways data for RAW 264.7 cells treated with Kdo2-lipid A: Eicosanoids [media]

Smaller image width=730px	Larger image width=980px	Full size image (width varies)

Kdo/Ctrl ratio:	>=10	>=5	>=4	>=3	>=2	>=1.5	---	<=0.67	<=0.5	<=0.33	<=0.25	<=0.2	<=0.1	No value
Color:	🟩	🟩	🟩	🟩	🟩	🟩	⬛	🟥	🟥	🟥	🟥	🟥	🟥	⬜

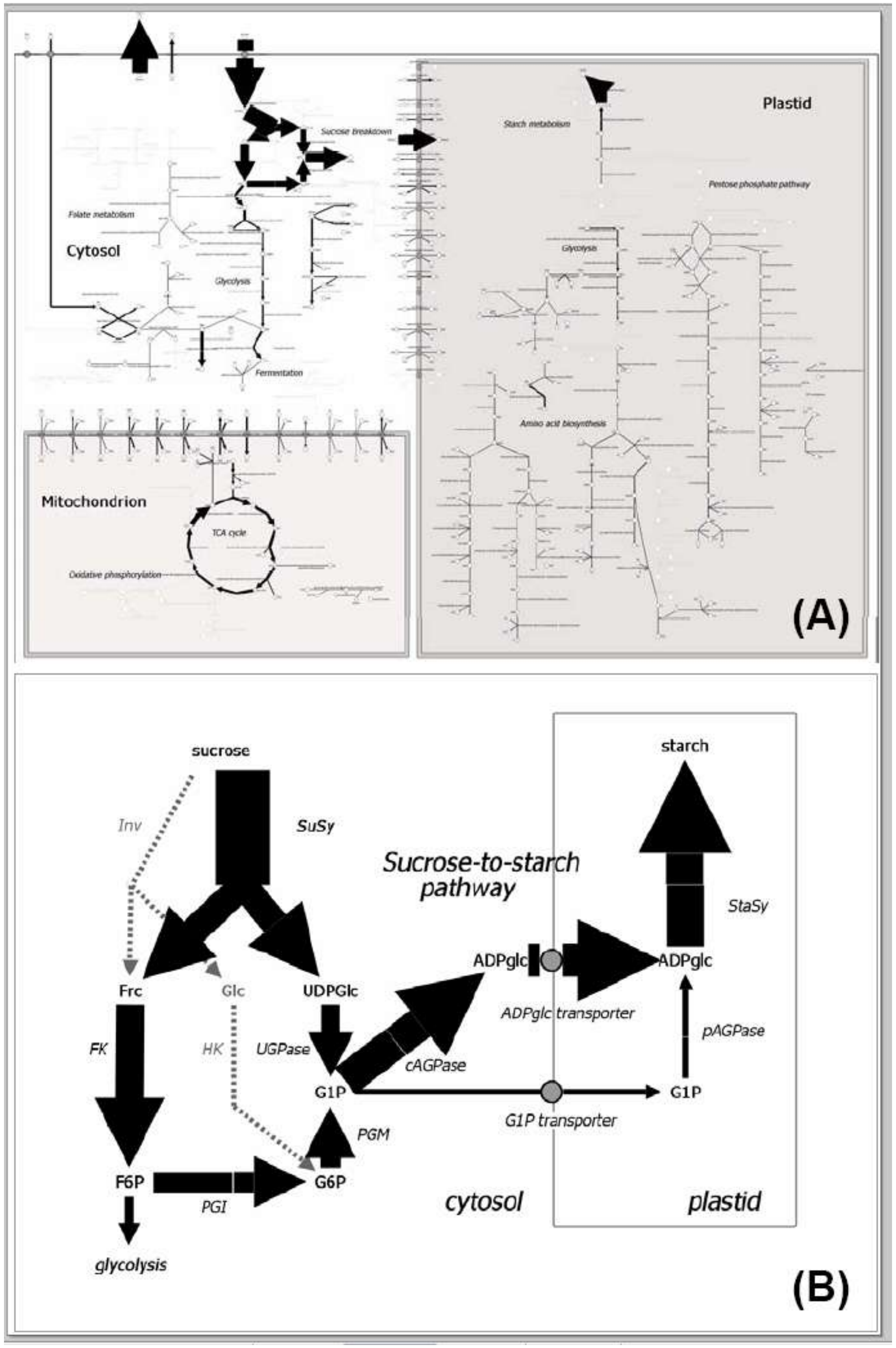
Kdo/Ctrl ratio timepoints: 30m, 1, 2, 4, 8, 12, 24 hrs

Please address any questions or comments about the LIPID MAPS website to our Webmaster.

LIPID MAPS Personnel
(Restricted Access)
Log Out

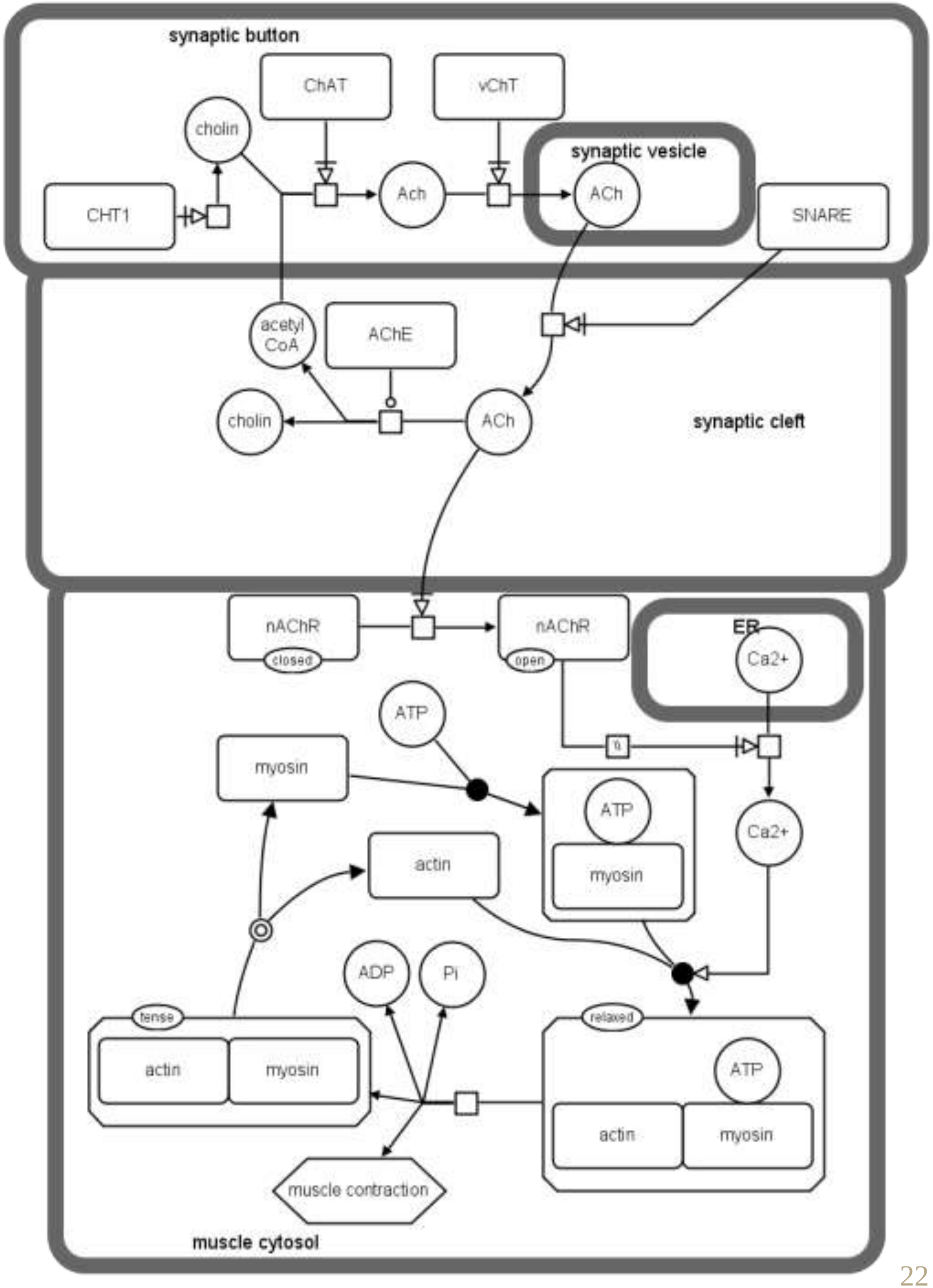
www.lipidmaps.org

Visualization of metabolic flux steady state simulation results

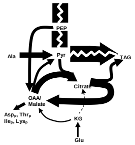


Grafahrend-Belau et al. (2008)

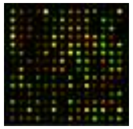
Support for editing of SBGN diagrams



Fluxes

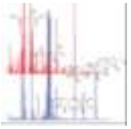


Transcripts



Enzymes

Metabolites

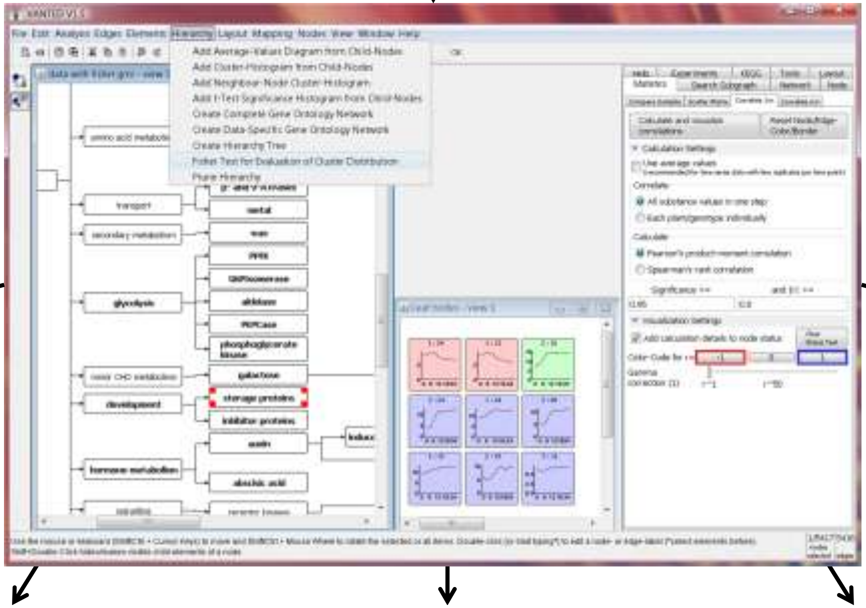


Ontologies

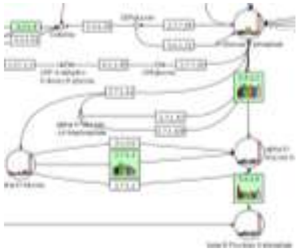
GO, KEGG KO,
user specified

Networks

KEGG, Meta-All,
file formats

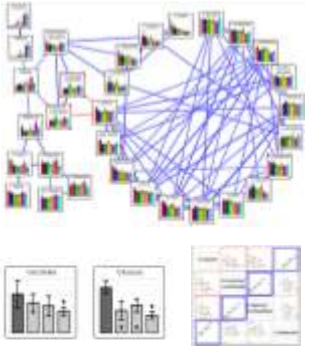

VANTED V1.5
File Edit Analysis Edges Elements Hierarchy Layout Mapping Nodes View Window Help
Add Average-Values Diagram from Child-Nodes
Add Cluster-Histogram from Child-Nodes
Add Neighbour-Node Cluster-Histogram
Add t-Test Significance Histogram from Child-Nodes
Create Complete Gene Ontology Network
Create Data-Specific Gene Ontology Network
Create Hierarchy Tree
Fisher Test for Evaluation of Cluster Distribution
Prune Hierarchy
amino acid metabolism
transport
secondary metabolism
glycolysis
minor CHO metabolism
development
hormone metabolism
metal
PEPCase
phosphoglycerate kinase
galactose
storage proteins
inhibitor proteins
auxin
abscisic acid
Calculation Settings
Use average values
Correlate
All substance values in one step
Each plant/genotype individually
Calculate
Pearson's product-moment correlation
Spearman's rank correlation
Significance
Visualization Settings
Add calculation details to node status

Combination of data from different omics-areas

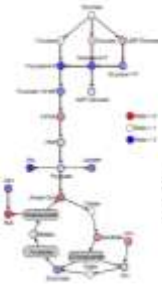


Integration of enzyme and metabolite data into KEGG pathways

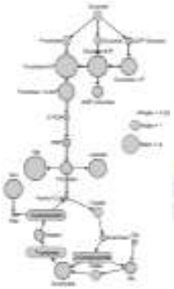
Data analysis, statistics, correlation networks



Flexible data visualization



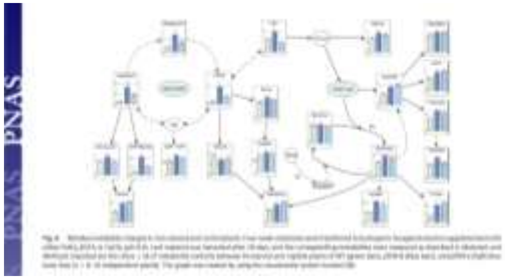
Color-coding



Shape-coding

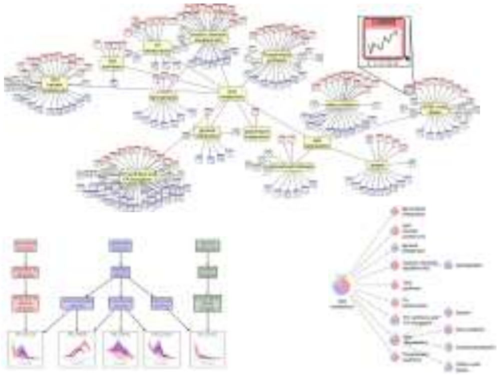
Diagrams

Figures for publications



Tognetti *et al* (2007), PNAS 104: 11495-11500

Processing hierarchical data



http://vanted.ipk-gatersleben.de