# Visualization and analysis of large-scale metabolite & expression data with VANTED

Christian Klukas

Network Analysis Group

Leibniz Institut für Pflanzengenetik
und Kulturpflanzenforschung (IPK)

MPI Golm-Potsdam
(2005-12-05)

**VANTED** - **V**isualization and **A**nalysis of **N**e**t**works containing **E**xperimental **D**ata

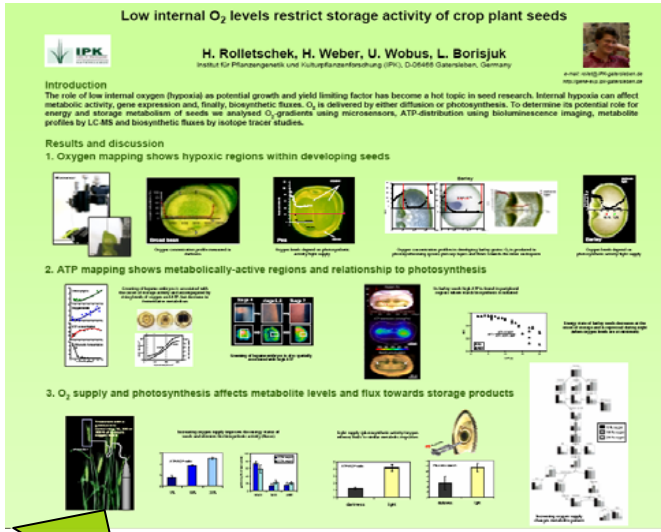
# Overview

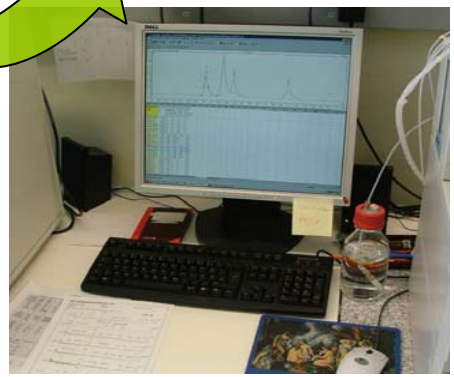
- Motivation
- System overview
- VANTEDs Features
- Demo
- Discussion

# Motivation

- Massively-parallel techniques generate more and more data
  - A top-down view on the biochemistry of a organism is made possible
- The amount of work needed to evaluate the data increases
  - New tools need to be evaluated or developed
- Goals
  - ☑ Show large amounts of data in a readable and understandable form
  - ☑ Consider related networks
  - ☑ Fast data evaluation with the help of statistic functions like t-test or correlation analysis, and clustering algorithms

# Motivation

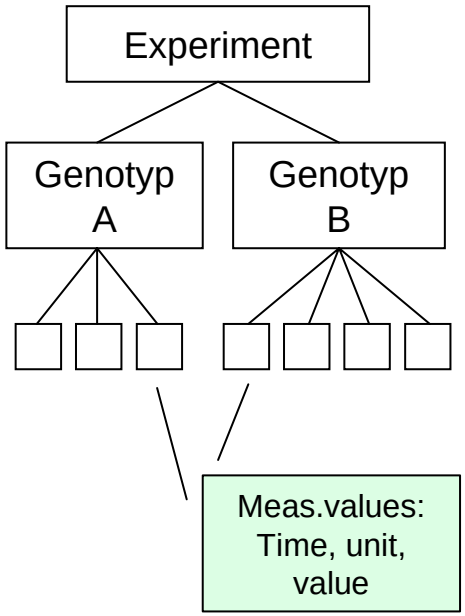

Low internal $O_2$ levels restrict storage activity of crop plant seeds
IPK
H. Rolletschek, H. Weber, U. Wobus, L. Borisjuk
Introduction
Results and discussion
1. Oxygen mapping shows hypoxic regions within developing seeds
2. ATP mapping shows metabolically-active regions and relationship to photosynthesis
3. $O_2$ supply and photosynthesis affects metabolite levels and flux towards storage products



# Key Features

- Data sources
  - Measurement data
    - DBE-Database (→ VANTED-DB)
    - **FLAREX** (Array experiment database at the IPK)
    - Excel Files (VANTED-template)
    - Text Files (**J-Express** format)
  - Pathway data (GML, Pajek-.NET, SBML)
- Data transformation and evaluation
  - t-test, U-test, Pearson- and Spearman correlation, SOM-data clustering, various layout commands, search and filter operations, extensible with script commands
- Data export
  - Image files (JPG, PNG, PDF, SVG)
  - Print out
  - Graph files (GML, Wilmascope-.XWG, DOT)

# Excel Input-Form



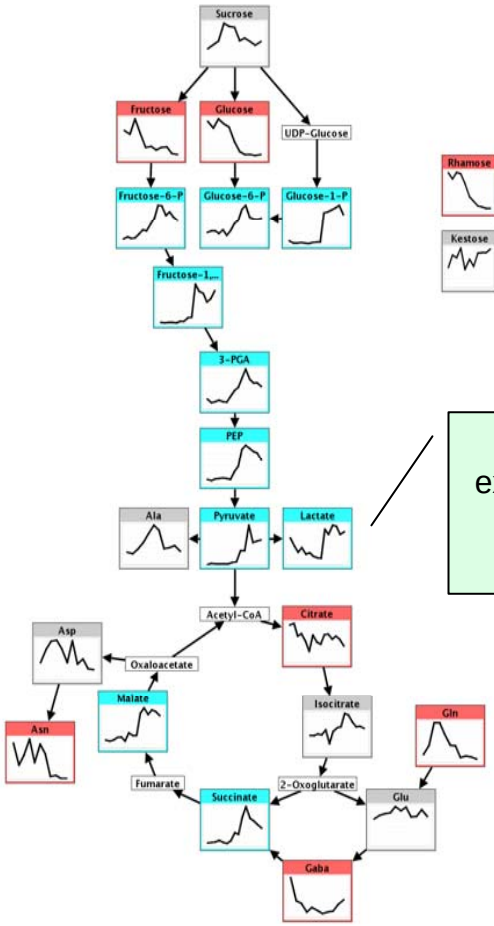

| Experiment | | | | | |
|---|---|---|---|---|---|
| Start of Experiment (Date) | 08.03.2004 | | | | optional |
| Remark* | GPTas-Linien | | | | input |
| Experiment Name (ID) | GPTas-Transgene | | | | contain numbers as 1, 2, 3, . |
| Coordinator | Hardy Rolletschek | | | | *** These cells must correlate to the numbers in * |
| Sequence-Name* | | | | | - The Experiment Name must be unique in the wh |
| | | | | | |
| **Plants/Genotypes**** | 1 | 2 | 3 | 4 | |
| Species | Vicia narbonensis | Vicia narbonensis | Vicia narbonensis | 'icia narbonensis | |
| Variety* | | | | | |
| Genotype | wild type | GPTas9 | GPTas13 | GPTas29 | |
| Growth conditions* | | | | | |
| Treatment* | | | | | |
| | | | | | |
| **Measurements** | | | | **Substance** | Asp / Glu |
| | | | | **Meas.-Tool*** | HPLC / HPLC |
| **Plant/Genotype***** | **Replicate #** | **Time*** | **Unit (Time)*** | **Unit** | Detector response / etector respons |
| 1 | 1 | | | | 4,611704652 / 6,167654385 |
| 1 | 2 | | | | 4,025788159 / 5,447092125 |
| 1 | 3 | | | | 3,805929642 / 4,888978365 |
| 1 | 4 | | | | 3,322600366 / 4,388163141 |
| 1 | 5 | | | | 4,322790612 / 5,194324773 |

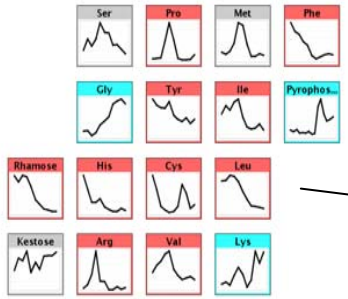
General information about the experiment

List of analysed plants / genotypes

Measurement values

# Data-Visualization

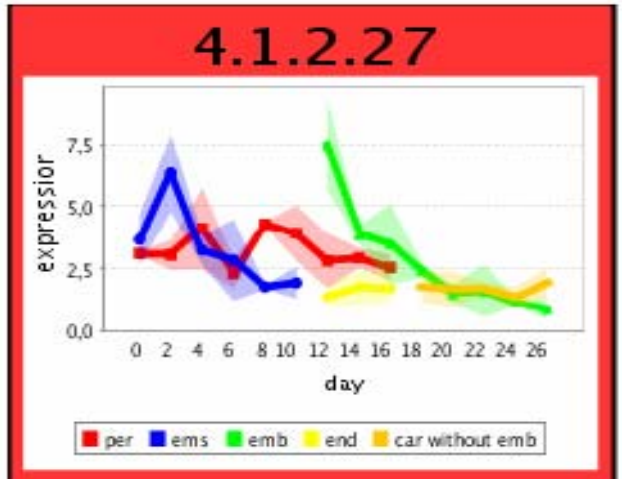



Clustering of data ⟶ Node-colour

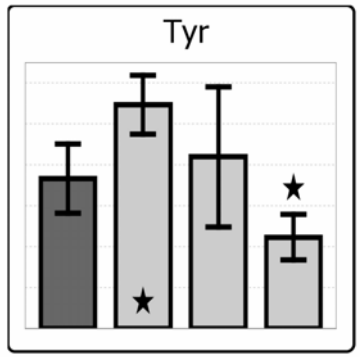
Level of detail

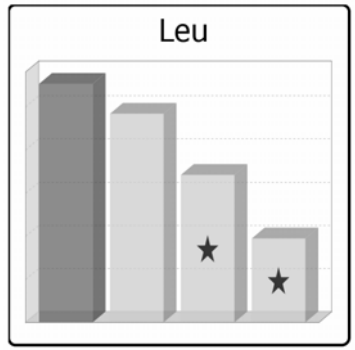
Mapping of experimental data onto pathways ⟶ bar or line charts



# Statistic Tests

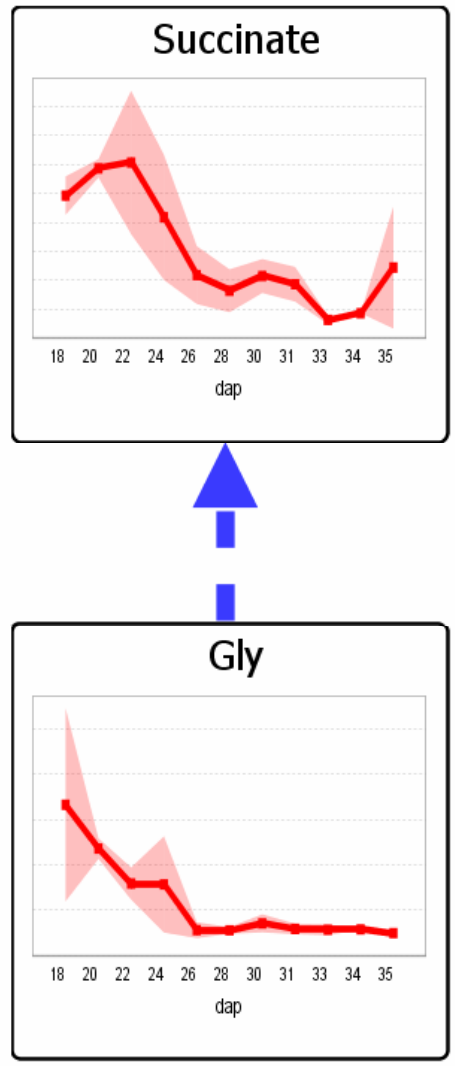
- Analyze data samples…
  - Check for normal distribution
    - ☑ David et al. quick test
    - ☐ Chi-square test
  - Detect/Remove outliers
    - ☑ Grubbs test
  - Detection of significant mean differences with
    - ☑ t-test (2 variants)
    - ☑ U-test (rank-sum test)



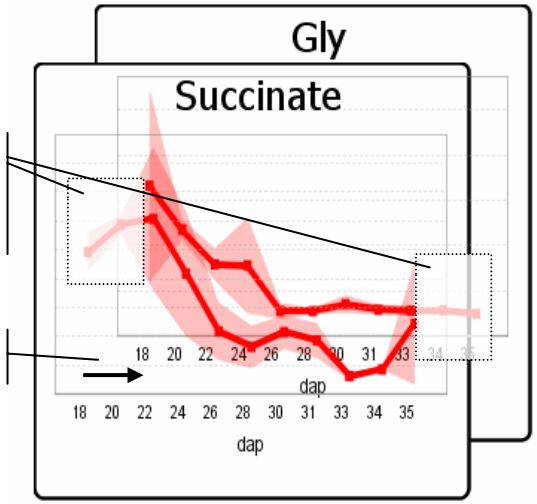
# Correlation Analysis (1/2)

- Calculation of the Pearson (linear) or Spearman (rank-order) correlation
- Detection of correlations, shifted in time:
  - Repeated correlation calculation ($r_i$) for multiple time-offsets ($i=-3...3$, $t_{-3}...t_{+3}$)
  - Using $\max|r_i|$ for data visualization
- Test of significance with approximation to the t-distribution

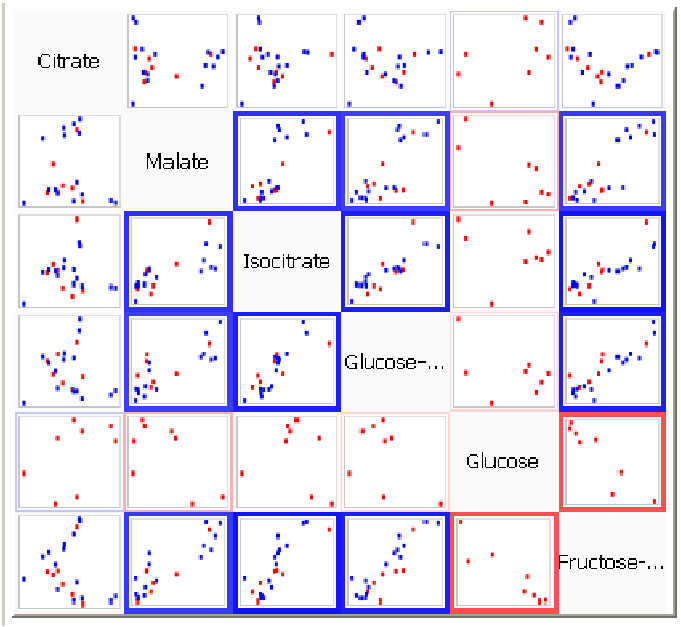



Number of data points decreases with increasing offset (t)

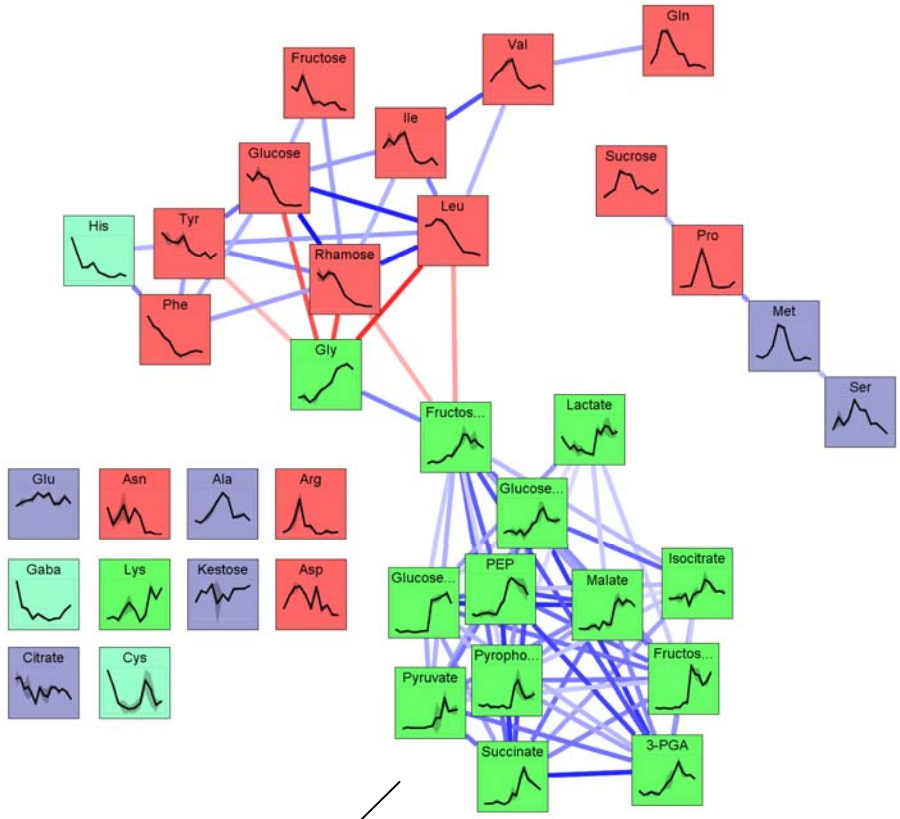
t = 2



# Correlation Analysis (2/2)




Scatter Matrix




Correlation network

# Summary & Outlook

- Website
  - http://vanted.ipk-gatersleben.de
- Publications
  - Borisjuk, Hajirezaei, Klukas, Rolletschek, Schreiber: *Integrating data from biological experiments into metabolic networks with the DBE information system*. In Silico Biology (2004)
  - Rolletschek, Radchuk, Klukas, Schreiber, Borisjuk: *Oil storage in soybean seeds: evidence for a key role of photosynthetic oxygen release*. New Phytologist (2005)
  - BMC Bioinformatics ?
- Outlook
  - Improve analysis and visualization of array data
    - Based on discussions with colleagues and feedback from users of the system
  - Display of simulated experimental data,  generated with SyBME

# Acknowledgements


- Ljudmilla Borisjuk, Mohammad-Reza Hajirezaei, Hardy Rolletschek, Nese Sreenivasulu, Winfriede Weschke, Ruslana Radchuk
  > Discussion of system features and data provision

- Matthias Lange, Uwe Scholz, Andreas Stephanik, Karl Spies
  > Database services and SOAP access to FLAREX

- Dirk Koschützki, Falk Schreiber
  > Discussions