Visualisierung und Analyse von biochemischen Daten im Kontext relevanter Netzwerke

Christian Klukas

Gruppe Netzwerkanalyse

Leibniz Institute of Plant Genetics and Crop Plant Research
Gatersleben

FOR 666
2006-05-12

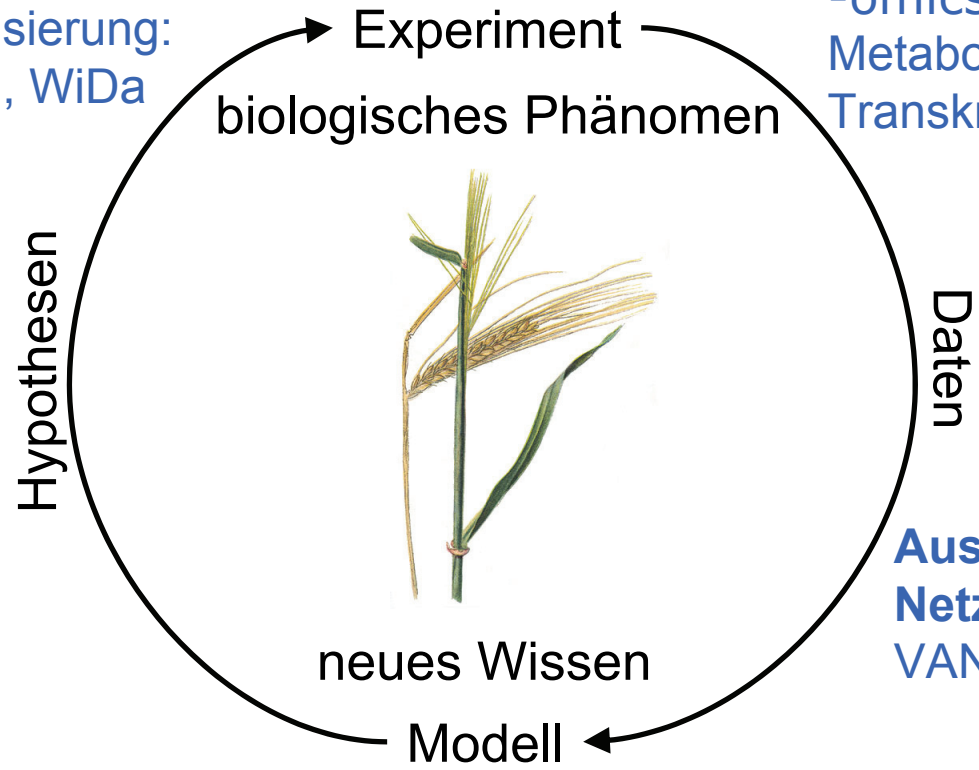
BIC GH

IPK GATERSLEBEN

Fokus der AG NW: Bioinformatorische Unterstützung der Wissensgenerierung in den Lebenswissenschaften

Strukturelle Netzwerkanalyse
Motive, Zentralitäten,
Vergleiche, Visualisierung:
MAVisto, CentiBiN, WiDa

Speicherung Hochdurchsatzdaten
-omics: Files, öffentl. DB
Metabolomics: DBE
Transkriptomics: Flarex (AG BI)



Auswertung Daten im Netzwerkkontext
VANTED (dieser Vortrag)

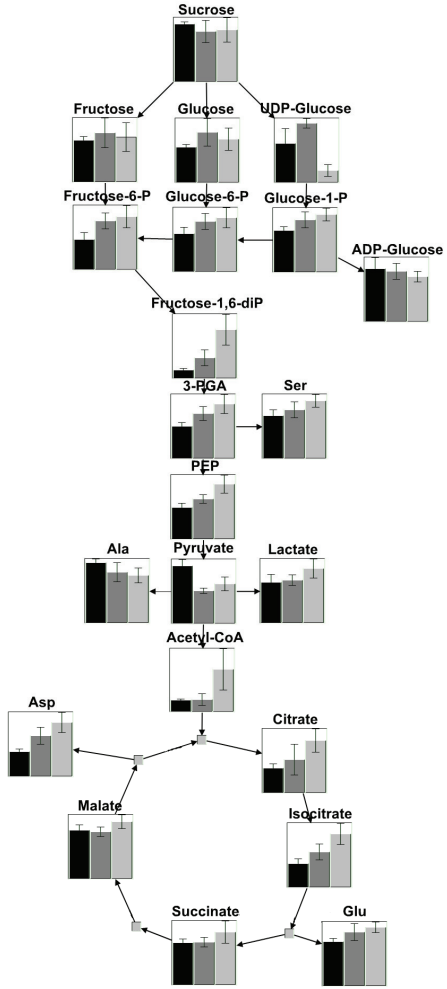
Modellrepräsentation und Simulationsunterstützung
Netzwerke: MetaCrop/MetAll (mit AG BI, PDW)
Simulationsunterstützung: SyBME

Agenda

- Motivation
- Datenquellen
  - Experimentelle Daten
  - Netzwerk-Daten
- Datenanalyse
  - Data-Clustering, Statistik
  - Visualisierungsmethoden
- Datenexport
- Zusammenfassung

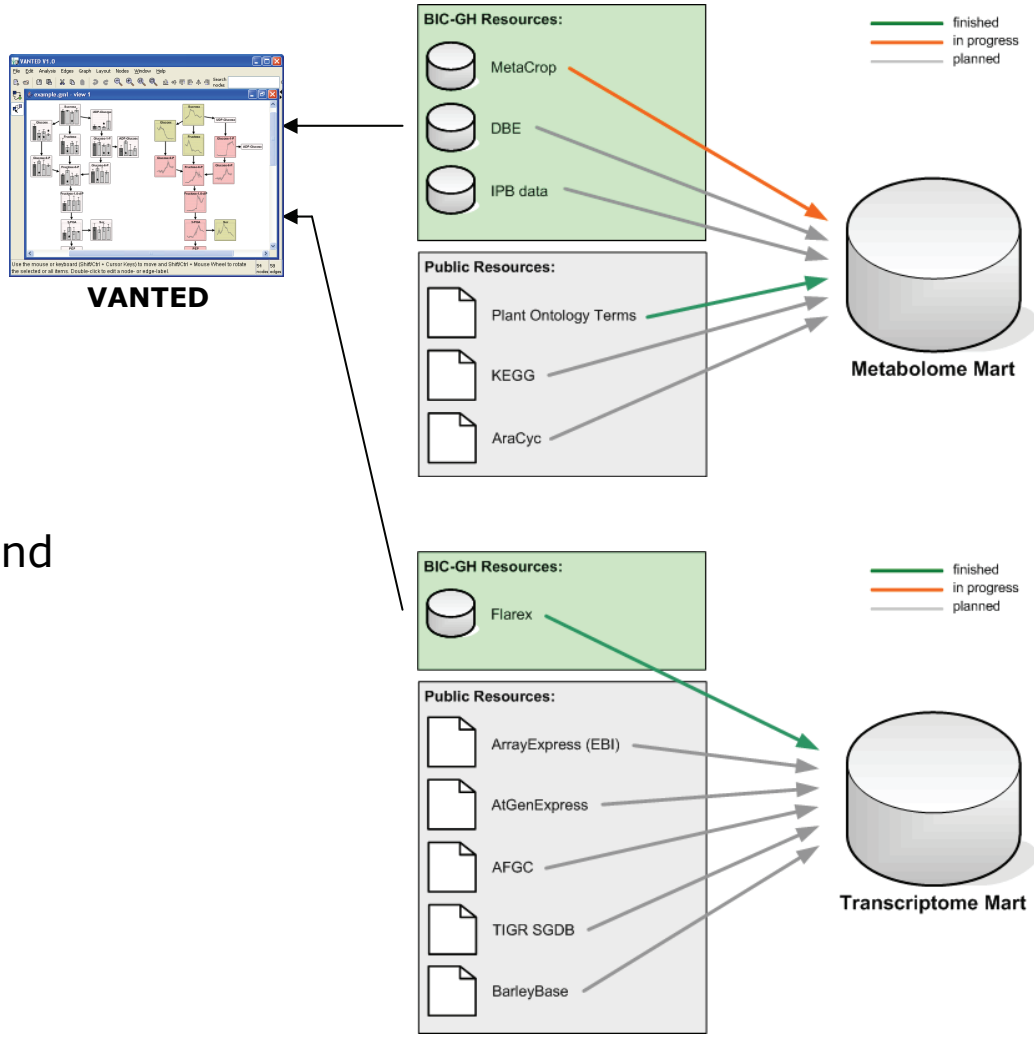
Motivation

- Massiv-parallele Meßtechniken generieren eine steigende Anzahl an Meßwerten
  - Dadurch wird eine Top-Down Sicht auf die Biochemie eines Organismus möglich (→ Systembiologie)
- Damit steigt der Arbeitsaufwand für die Datenanalyse
  - Softwarewerkzeuge müssen evaluiert, angepasst oder neu entwickelt werden
- Ziele
  - ☑ Darstellung von großen Datenmengen in einer leicht verständlichen und übersichtlichen Form
  - ☑ Berücksichtigung von relevanten Netzwerken
  - ☑ Schnelle Datenevaluierung mit Hilfe von statistischen Tests und Daten-Clustering



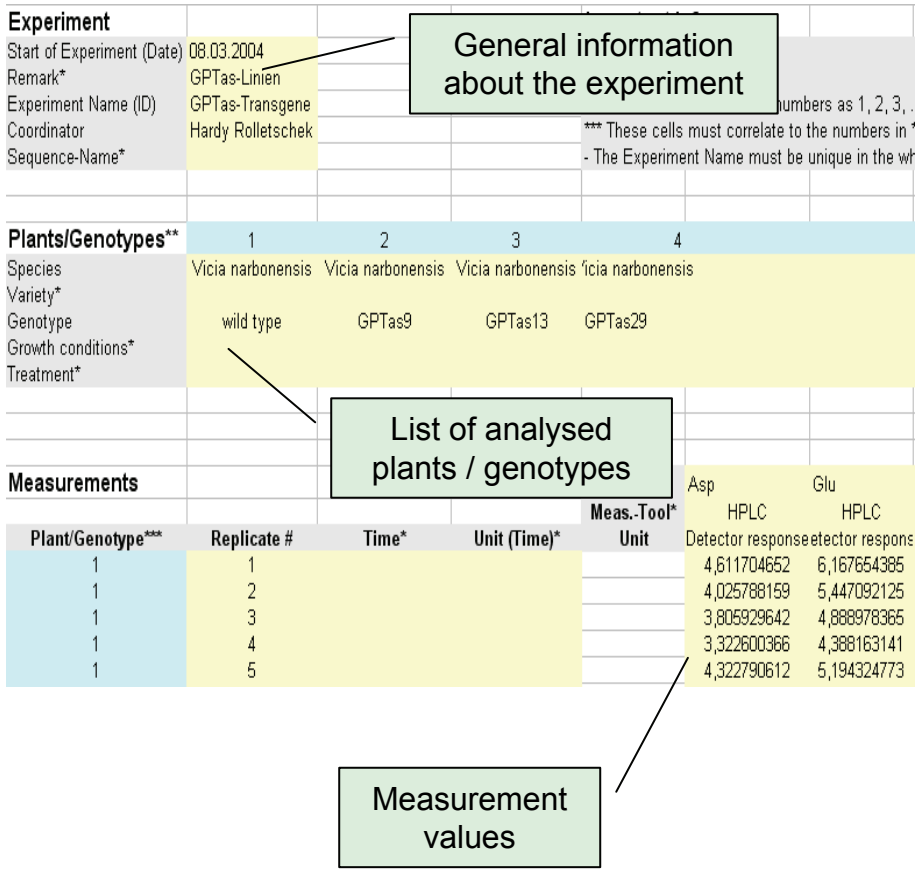
Experimentelle Daten

- **Datenbanken**
  - **DBE**
  - **FLAREX**
  - **...**
- Excel / CSV Dateien
  - VANTED Eingabe-Datei (Metabolit-, Proteomics- und Expressionsdaten)
  - J-Express Dateien (Expressionsdaten)
- Programmerweiterungen
  - Script API: Java/Ruby Quelltext



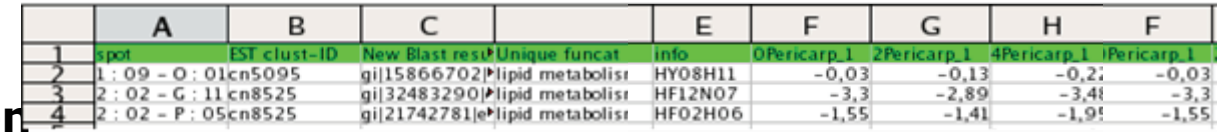
Experimentelle Daten

- Datenbanken
  - DBE
  - FLAREX
  - **...**
- **Excel / CSV Dateien**
  - **VANTED Eingabe-Datei (Metabolit-, Proteomics- und Expressionsdaten)**
  - J-Express Dateien (Expressionsdaten)
- Programmerweiterungen
  - Script API: Java/Ruby Quelltext



Experimentelle Daten

- Datenbanken
  - DBE
  - FLAREX
  - **...**
- Excel / CSV Dateien
  - VANTED Eingabe-Datei (Metabolit-, Proteomics- und Expressionsdaten)
  - **J-Express Dateien (Expressionsdaten)**

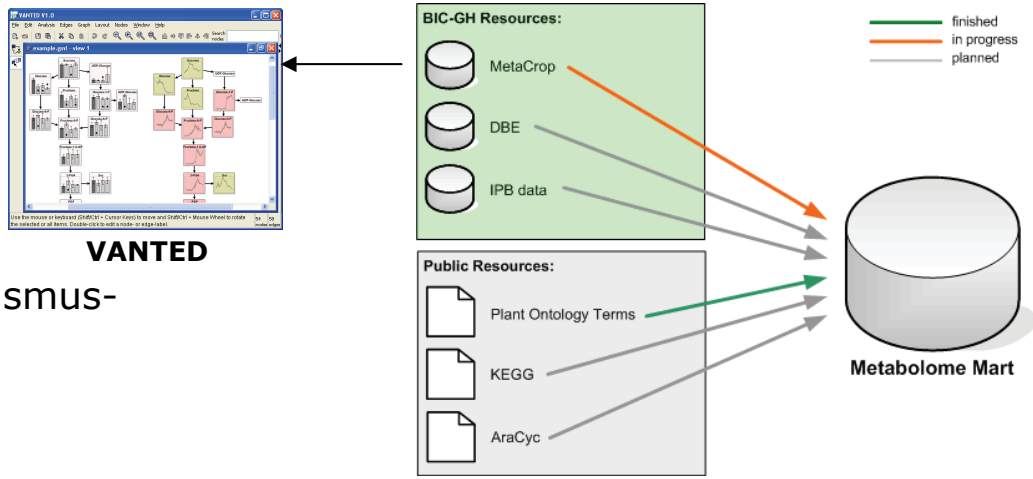


	A	B	C		E	F	G	H	F
1	spot	EST clust-ID	New Blast resu	Unique funcat	info	0Pericarp_1	2Pericarp_1	4Pericarp_1	Pericarp_1
2	1 : 09 – O : 01	cn5095	gi|15866702|	lipid metabolisı	HY08H11	-0,03	-0,13	-0,2	-0,03
3	2 : 02 – G : 11	cn8525	gi|32483290|	lipid metabolisı	HF12N07	-3,3	-2,89	-3,4	-3,3
4	2 : 02 – P : 05	cn8525	gi|21742781|e	lipid metabolisı	HF02H06	-1,55	-1,41	-1,9	-1,55

- Programmerweiterungen
  - Script API: Java/Ruby Quelltext

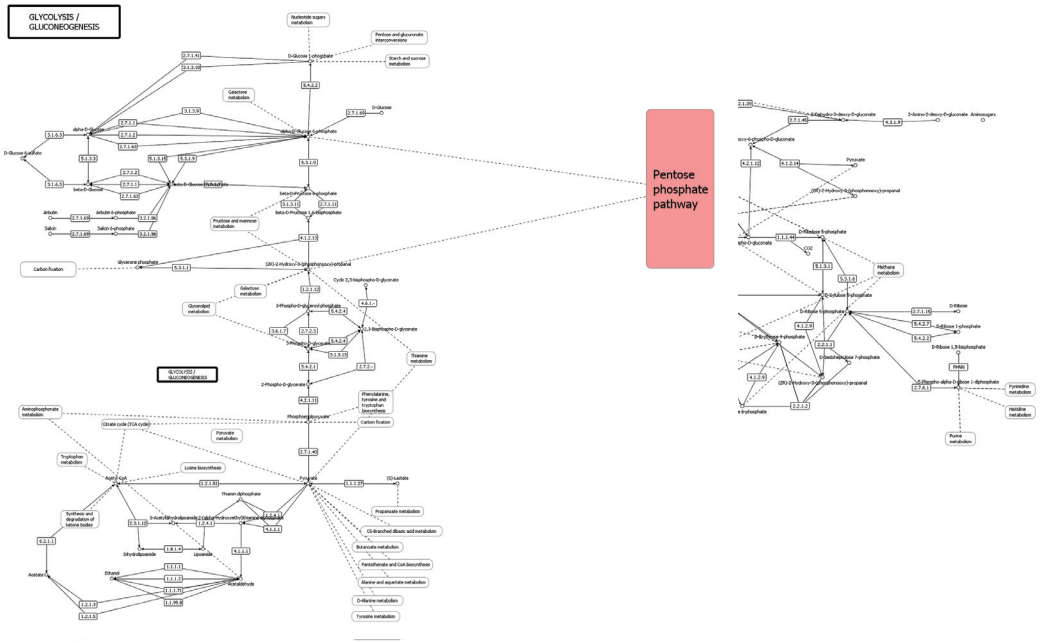
Netzwerk-Daten

- **Datenbanken**
  - **MetaCrop** (SBML Export)
  - KEGG Pathway
    - Referenz-Pathway/Organismus-spezifisch
    - Bottom-Up
    - Top-Down
    - Super-Pathway
- Gene Ontology
  - Vollständige GO Hierarchie
  - Relevanter Teilbereich
- Dateien
  (GML, Pajek-.NET, SBML)



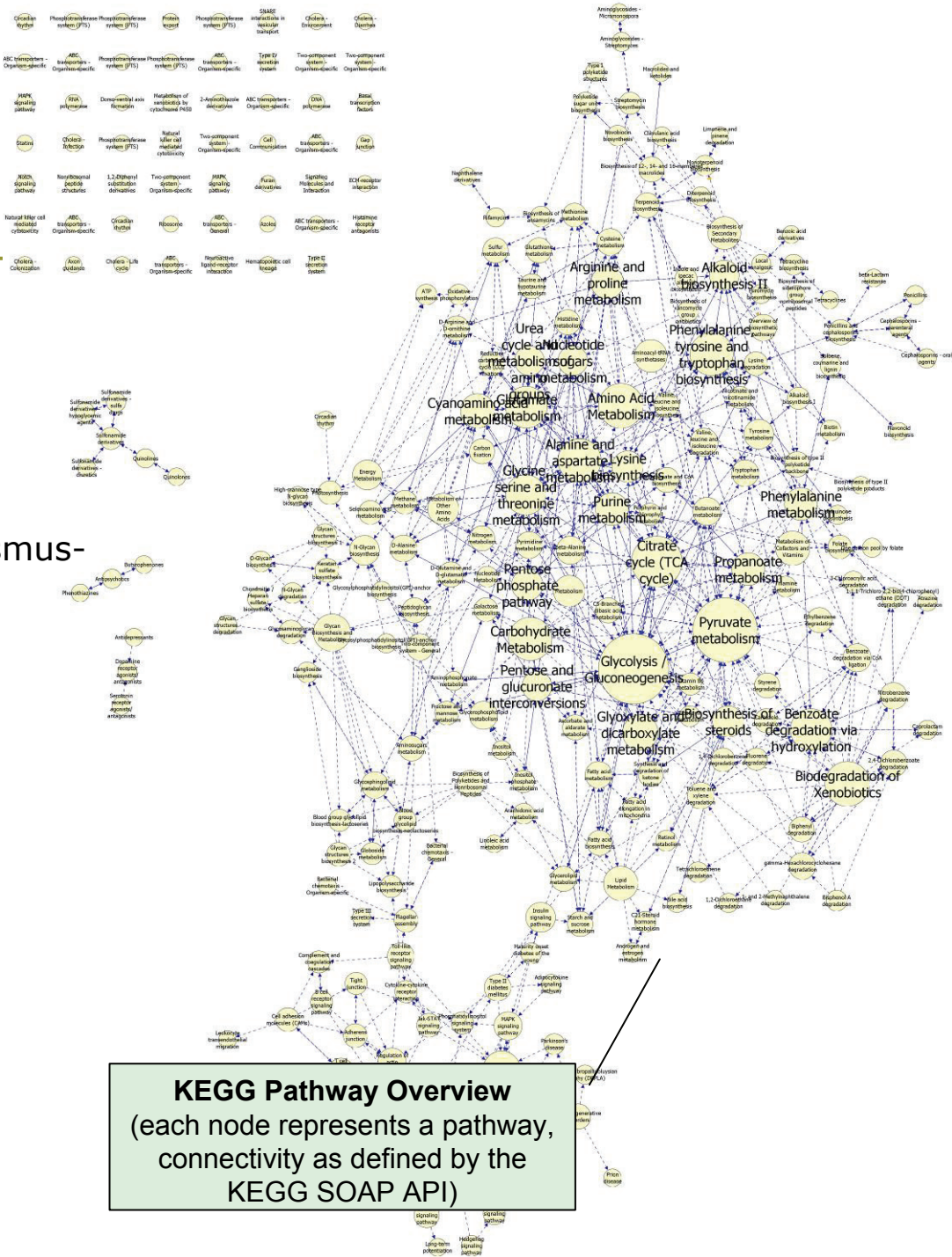
Netzwerk-Daten

- **Datenbanken**
  - MetaCrop (SBML Export)
  - **KEGG Pathway**
    - Referenz-Pathway/Organismus-spezifisch
    - **Bottom-Up**
    - Top-Down
    - Super-Pathway
- Gene Ontology
  - Vollständige GO Hierarchi
  - Relevanter Teilbereich
- Dateien
  (GML, Pajek-.NET, SBML)



Netzwerk-Daten

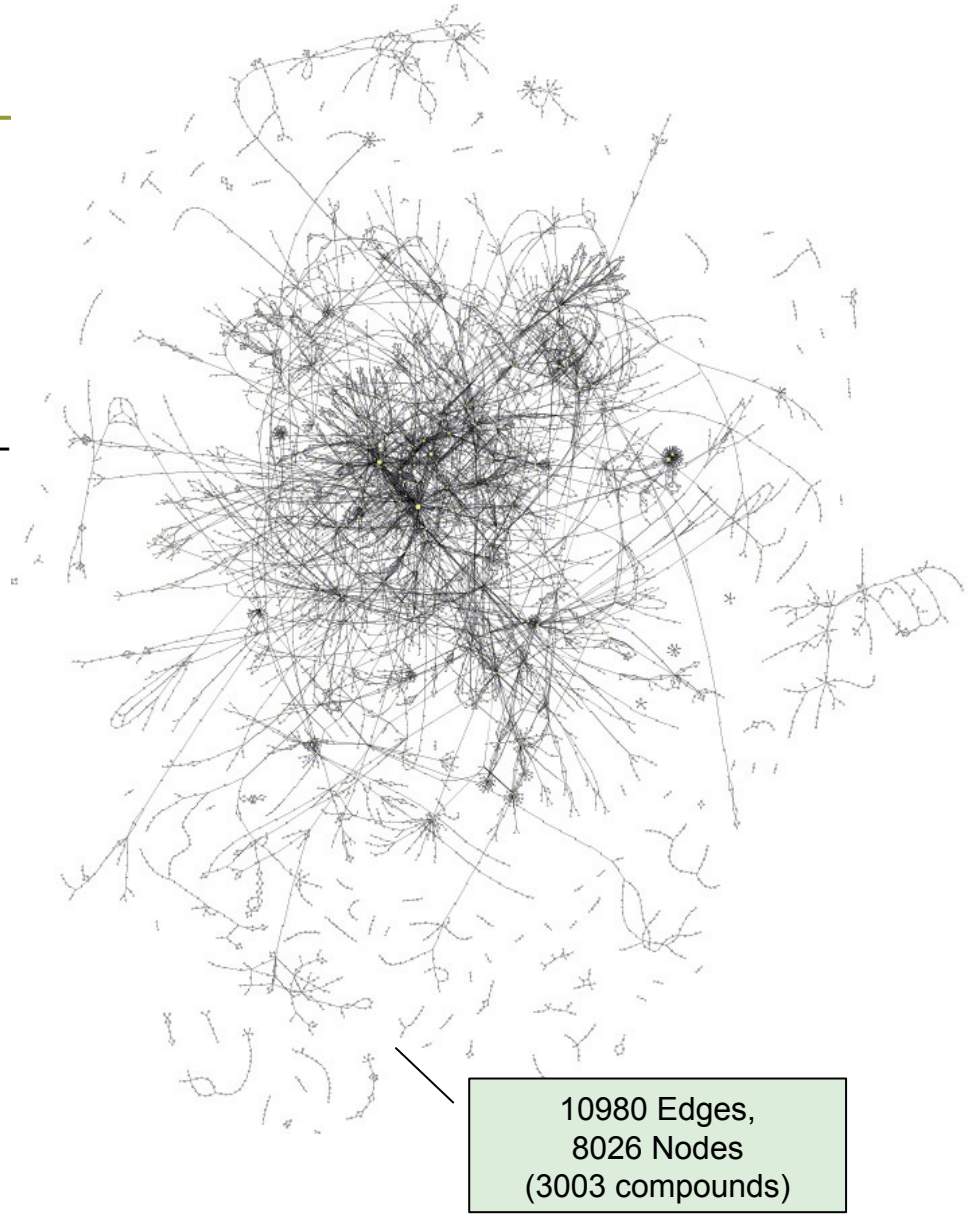
- **Datenbanken**
  - MetaCrop (SBML Export)
  - **KEGG Pathway**
    - Referenz-Pathway/Organismus-spezifisch
    - Bottom-Up
    - **Top-Down**
    - Super-Pathway
- Gene Ontology
  - Vollständige GO Hierarchie
  - Relevanter Teilbereich
- Dateien
  (GML, Pajek-.NET, SBML)



KEGG Pathway Overview
(each node represents a pathway, connectivity as defined by the KEGG SOAP API)

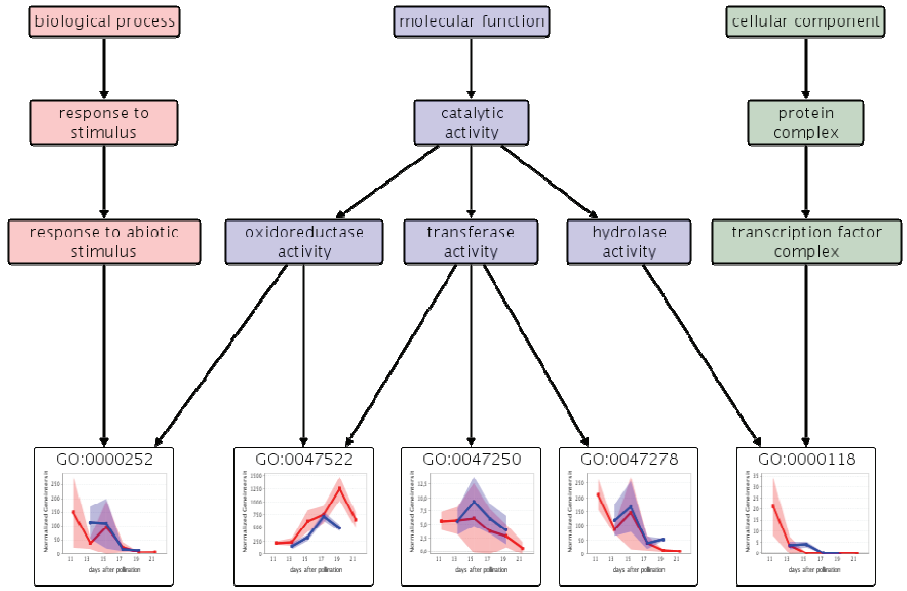
Netzwerk-Daten

- **Datenbanken**
  - MetaCrop (SBML Export)
  - **KEGG Pathway**
    - Reference-Pathway/Organismus-spezifisch
    - Bottom-Up
    - Top-Down
    - **Super-Pathway**
- Gene Ontology
  - Vollständige GO Hierarchie
  - Relevanter Teilbereich
- Dateien
  (GML, Pajek-.NET, SBML)



Netzwerk-Daten

- Datenbanken
  - MetaCrop (SBML Export)
  - KEGG Pathway
    - Reference-Pathway/Organismus-spezifisch
    - Bottom-Up
    - Top-Down
    - Super-Pathway
- **Gene Ontology**
  - Vollständige GO Hierarchie
  - **Relevanter Teilbereich**
- Dateien
  (GML, Pajek-.NET, SBML)



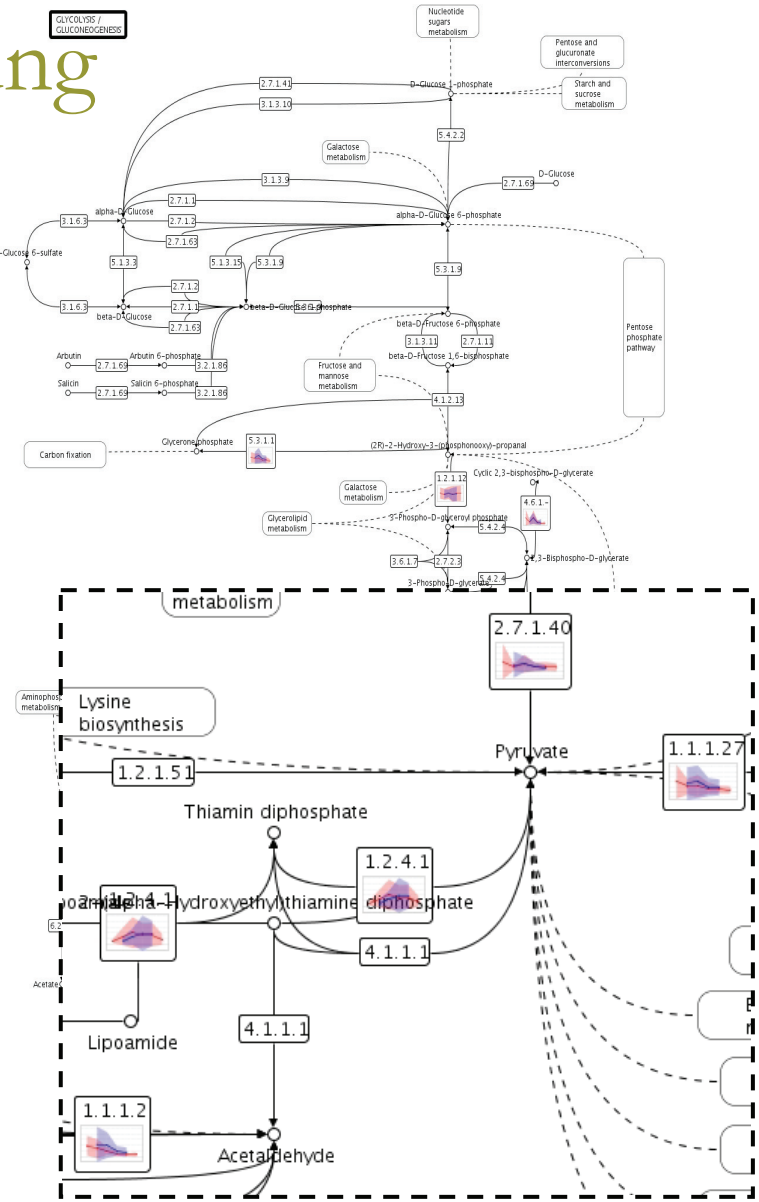
Visualisation of gene expression time series data and corresponding gene ontology data

Datenanalyse: Visualisierung

- Graph layout Algorithmen
- Chart-Techniken
  - Daten-Mapping, Netzwerk-integrierte Sicht von experimentellen Daten
  - Zeit-Serien Daten
    - Linien Diagramme (1)
  - Daten ohne Zeitbezug
    - Balken Diagramme (2)
    - Kreis Diagramme (3)
    - Verhältnis Sicht (4)
  - Filter Operationen
    - Darstellung/Analyse eines Teils der Daten (bestimmte Zeitpunkte oder Linien)
- Kombination von Daten verschiedener "–omics" Bereiche

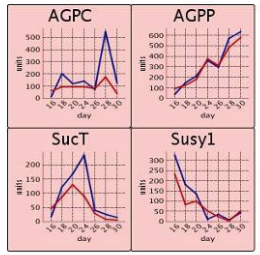
Datenanalyse: Visualisierung

- Graph layout Algorithmen
- Chart-Techniken
  - **Daten-Mapping, Netzwerk-integrierte Sicht von experimentellen Daten**
  - Zeit-Serien Daten
    - Linien Diagramme (1)
  - Daten ohne Zeitbezug
    - Balken Diagramme (2)
    - Kreis Diagramme (3)
    - Verhältnis Sicht (4)
  - Filter Operationen
    - Darstellung/Analyse eines Teils der Daten (bestimmte Zeitpunkte oder Linien)
- Kombination von Daten verschiedener "–omics" Bereiche

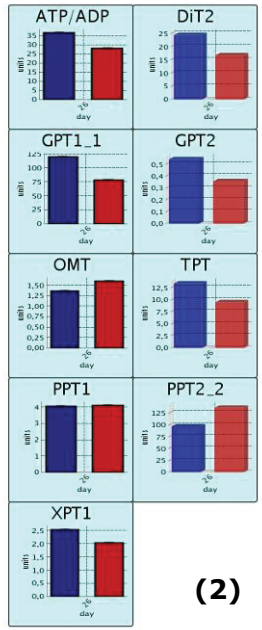


Datenanalyse: Visualisierung

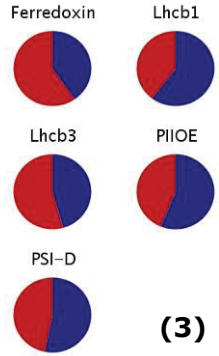
- Graph layout Algorithmen
- Chart-Techniken
  - Daten-Mapping, Netzwerk-integrierte Sicht von experimentellen Daten
  - **Zeit-Serien Daten**
    - **Linien Diagramme (1)**
  - **Daten ohne Zeitbezug**
    - **Balken Diagramme (2)**
    - **Kreis Diagramme (3)**
    - **Verhältnis Sicht (4)**
  - Filter Operationen
    - Darstellung/Analyse eines Teils der Daten (bestimmte Zeitpunkte oder Linien)
- Kombination von Daten verschiedener "–omics" Bereiche



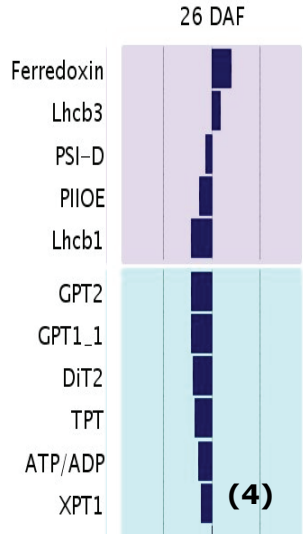
(1)



(2)



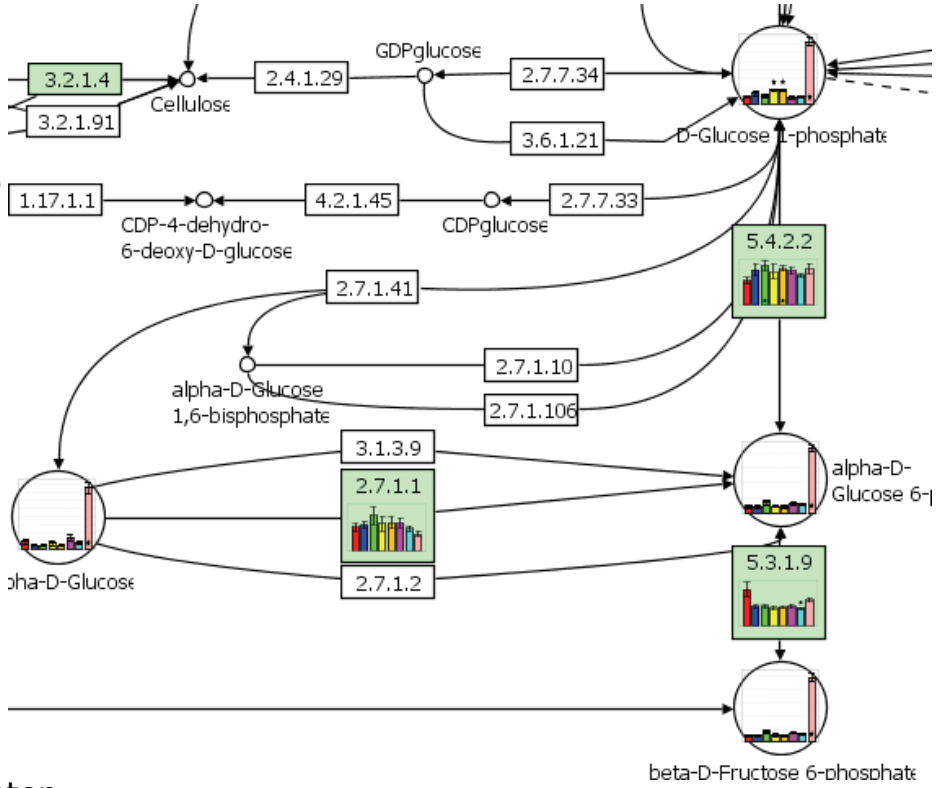
(3)



(4)

Datenanalyse: Visualisierung

- Graph layout Algorithmen
- Chart-Techniken
  - Daten-Mapping, Netzwerk-integrierte Sicht von experimentellen Daten
  - Zeit-Serien Daten
    - Linien Diagramme (1)
  - Daten ohne Zeitbezug
    - Balken Diagramme (2)
    - Kreis Diagramme (3)
    - Verhältnis Sicht (4)
  - Filter Operationen
    - Darstellung/Analyse eines Teils der Daten (bestimmte Zeitpunkte oder Linien)
- **Kombination von Daten verschiedener "–omics" Bereiche**



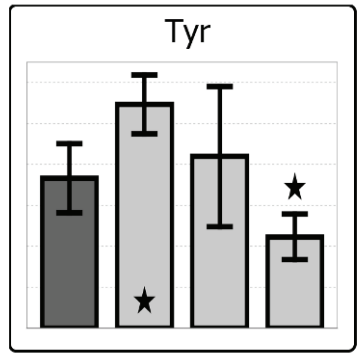
Compound / Enzyme information, mapped onto a KEGG pathway

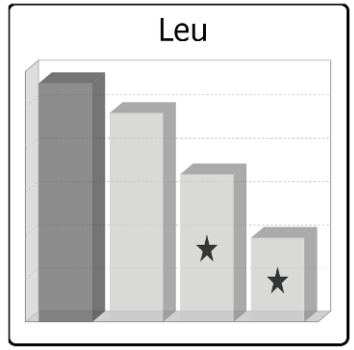
Datenanalyse: Statistik

- Prüfung auf Normalverteilung
  - ☑ David-Schnelltest
  - ☐ Chi-Quadrat Test
- Erkennung/Entfernung von Ausreißern
  - ☑ Grubbs Test
- Erkennung von signifikanten Mittelwert-Unterschieden
  - ☑ t-Test (2 Varianten)
  - ☑ U-Test (Rang-Summen-Test)
- Korrelationsanalyse
  - ☑ Korrelation von Zeitserien-Profilen
  - ☑ Korrelation von Proben (Replikate)

Datenanalyse: Statistik

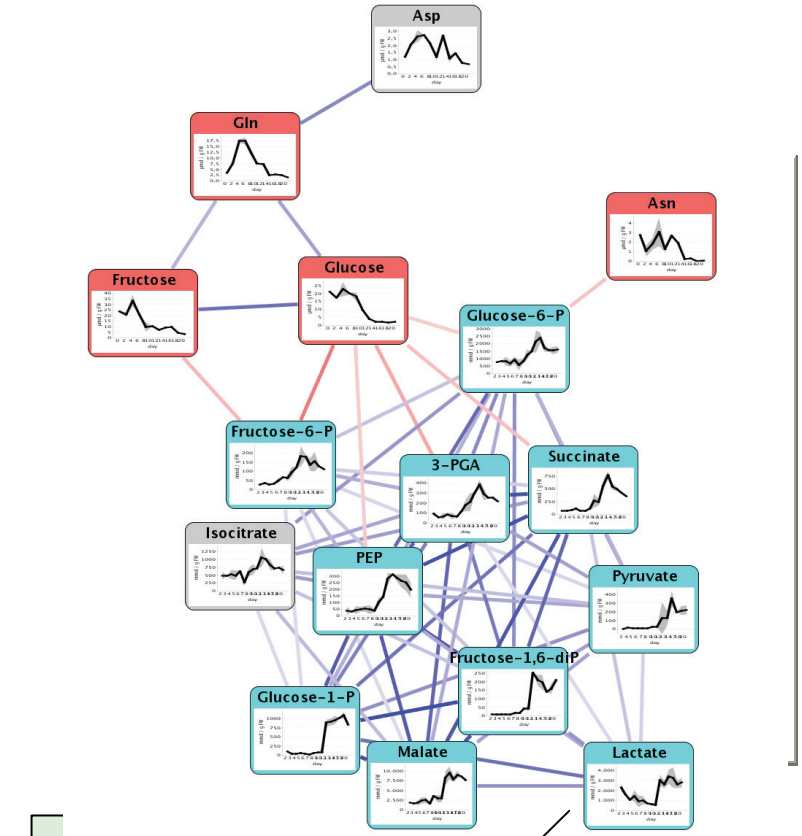
- Prüfung auf Normalverteilung
  - ☑ David-Schnelltest
  - ☐ Chi-Quadrat Test
- Erkennung/Entfernung von Ausreißern
  - ☑ Grubbs Test
- **Erkennung von signifikanten Mittelwert-Unterschieden**
  - ☑ **t-Test (2 Varianten)**
  - ☑ **U-Test (Rang-Summen-Test)**
- Korrelationsanalyse
  - ☑ Korrelation von Zeitserien-Profilen
  - ☑ Korrelation von Proben (Replikate)





Datenanalyse: Statistik

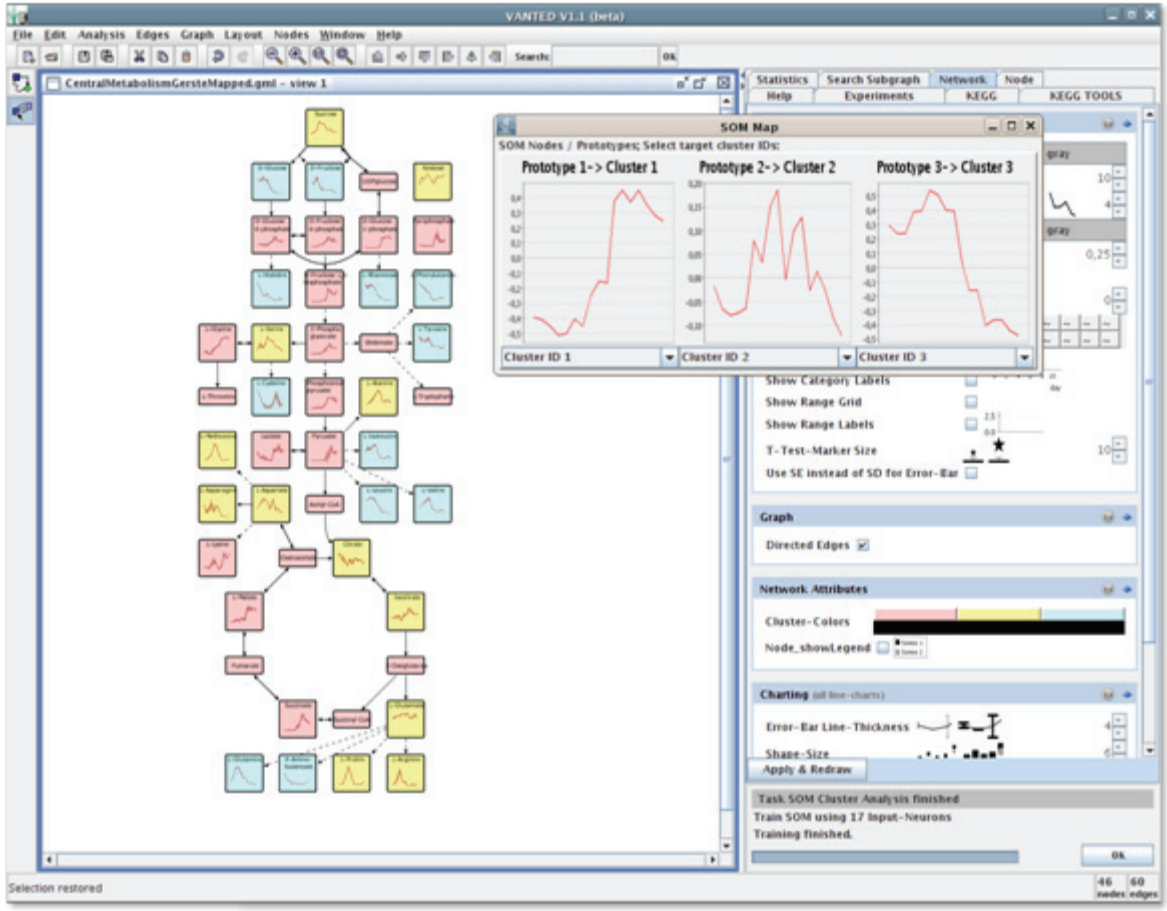
- Prüfung auf Normalverteilung
  - ☑ David-Schnelltest
  - ☐ Chi-Quadrat Test
- Erkennung/Entfernung von Ausreißern
  - ☑ Grubbs Test
- Erkennung von signifikanten Mittelwert-Unterschieden
  - ☑ t-Test (2 Varianten)
  - ☑ U-Test (Rang-Summen-Test)
- **Korrelationsanalyse**
  - ☑ **Korrelation von Zeitserien-Profilen**
  - ☑ Korrelation von Proben (Replikate)



Korrelationsnetzwerk

Datenanalyse: Daten-Clustering

- Integierter Self-Organizing Map (SOM) Algorithmus
  - Erkennung typischer Datenprofile, z.B. up-Regulation
- Verbindung zu externen Daten-Gruppierungs-ansätzen



Datenexport

- Bild-Dateien
  - JPG, PNG, PDF, SVG
- Druck
- Netzwerk-Dateien
  - GML, Wilmascope-.XWG, DOT

Zusammenfassung

- VANTED hilft bei Analyse und Visualisierung von Metabolit-, Proteomics- und Transcriptomics-Daten im Kontext relevanter Netzwerke
- VANTED wird innerhalb und außerhalb des IPK verwandt
- Weitere Informationen sind unter http://vanted.ipk-gatersleben.de verfügbar
- Bei Interesse Demo möglich

Danksagungen

- Ljudmilla Borisjuk, Mohammad-Reza Hajirezaei, Björn H. Junker, Dirk Koschützki, Rainer Lemke, Ruslana Radchuk, Hardy Rolletschek, Falk Schreiber, Nese Sreenivasulu, Winfriede Weschke
  > Discussion of system features and data provision

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