Visualization and Analysis of FLAREX Gene Expression Data with VANTED

Christian Klukas

Network Analysis Group

Leibniz Institut für Pflanzengenetik
und Kulturpflanzenforschung (IPK)

PGRC Progress Seminar
(2005-11-24)

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

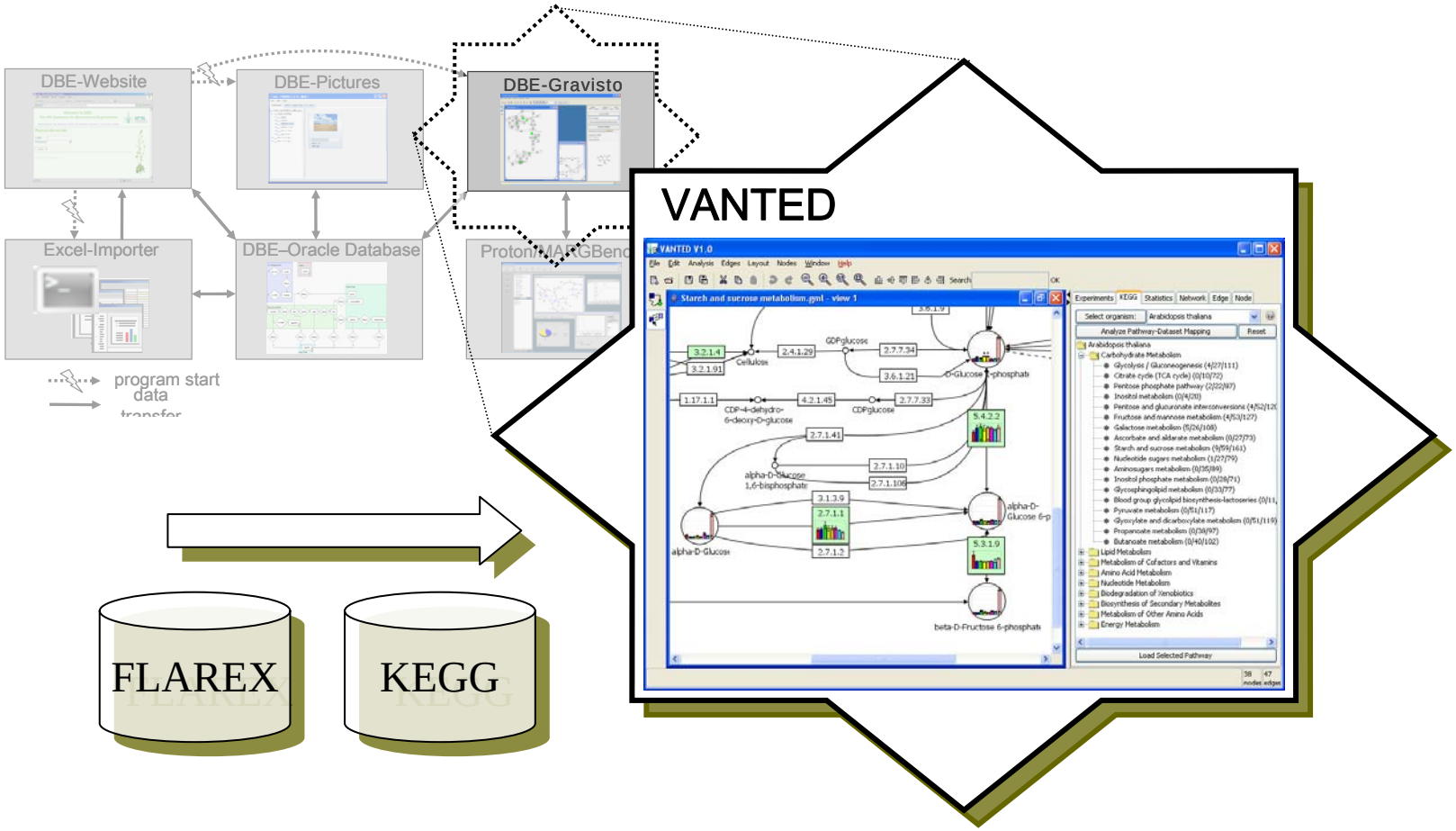
Overview

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

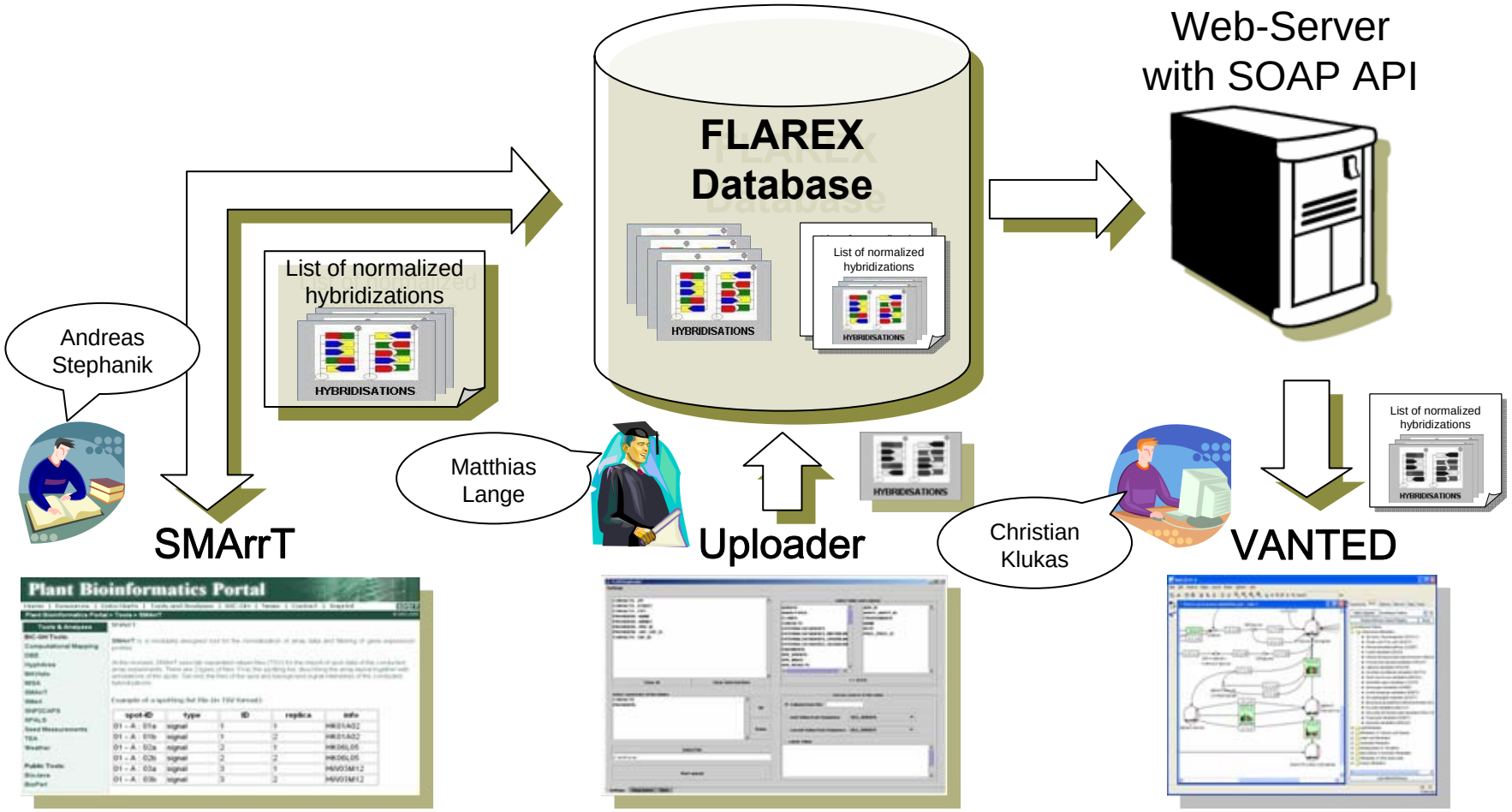
Motivation

- 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

System overview
Roots of VANTED

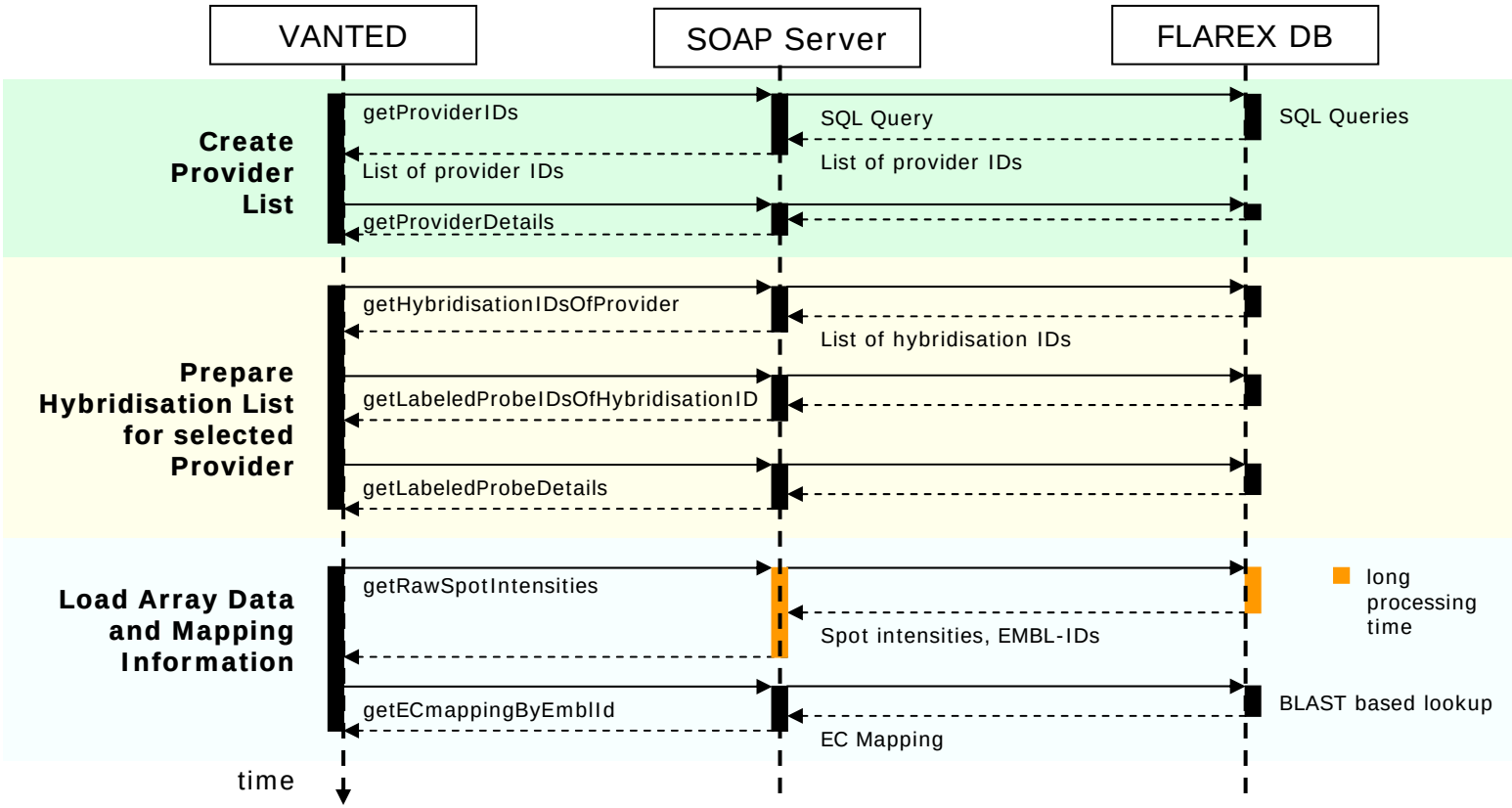


System overview
Related Tools and Devices



SOAP Access to FLAREX

- API designed together with Andreas Stephanik and Matthias Lange, implemented by Karl Spies
- Contains 19 methods
- Sequence – Diagram (Example for access to not-normalized spot-intensities):

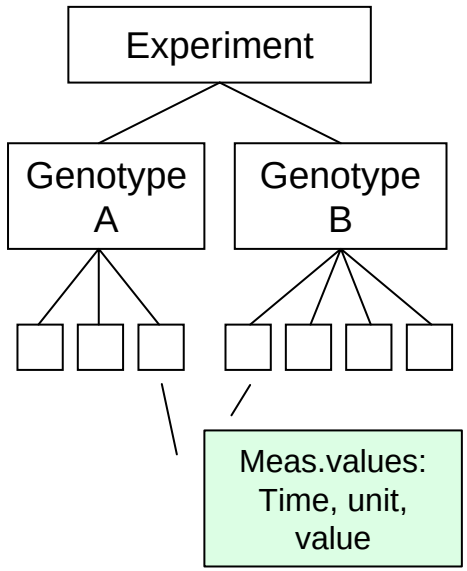


Key Features

- 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)

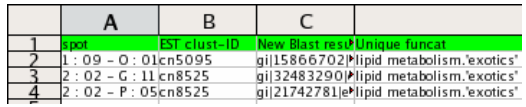
Loading of experimental data

Data hierarchy



VANTED Excel Template

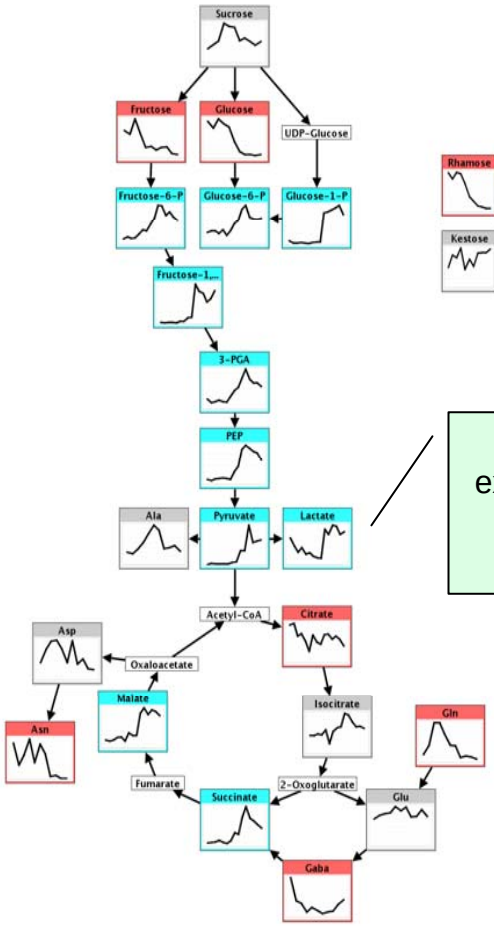
J-Express Text/Excel Files

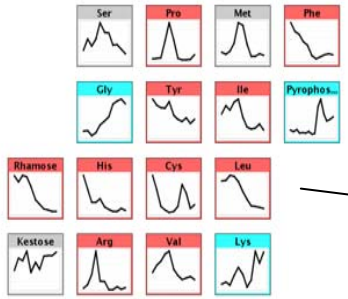


	A	B	C	
1	spot	EST clust-ID	New Blast resu	Unique funcat
2	1 : 09 - O : 01	cn5095	gi|15866702|	lipid metabolism.'exotics'
3	2 : 02 - G : 11	cn8525	gi|32483290|	lipid metabolism.'exotics'
4	2 : 02 - P : 05	cn8525	gi|21742781|e	lipid metabolism.'exotics'

FLAREX Hybridizations

Data-Visualization

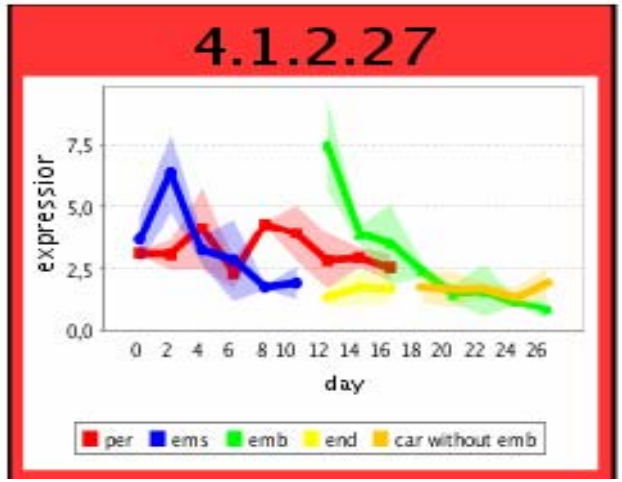




Clustering of data ⟶ Node-colour

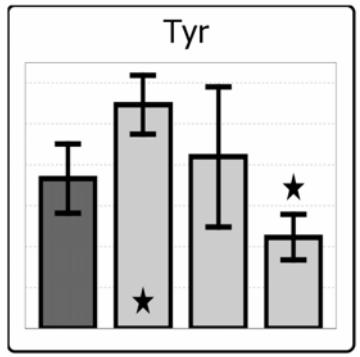
Level of detail

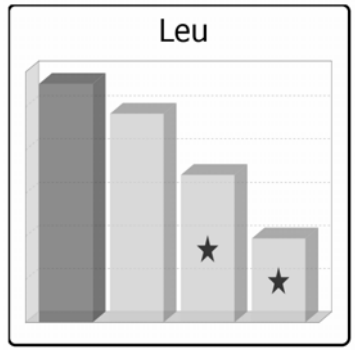
Mapping of experimental data onto pathways ⟶ bar or line charts



Statistic Tests

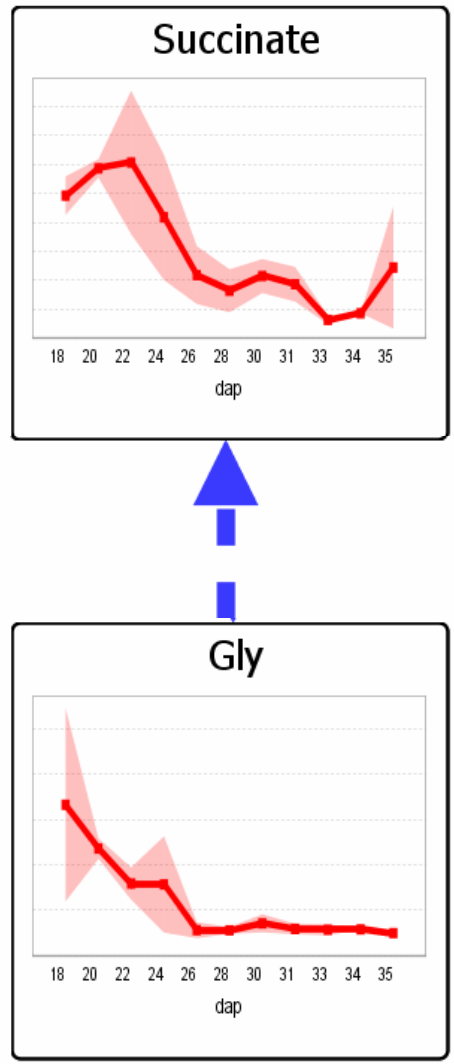
- Analyze data samples…
  - Check for normal distribution
    - ☑ David et al. quick test
    - ☐ Chi-square test
  - Detect/Remove outliers
    - ☑ Grubbs test
  - Detection of significant mean differences with
    - ☑ t-test (2 variants)
    - ☑ U-test (rank-sum test)





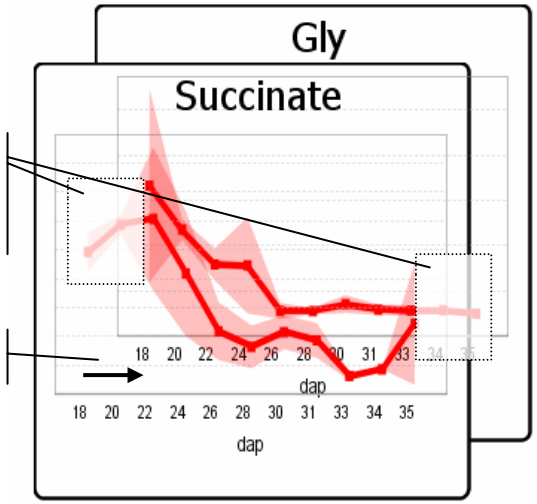
Correlation Analysis (1/2)

- 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

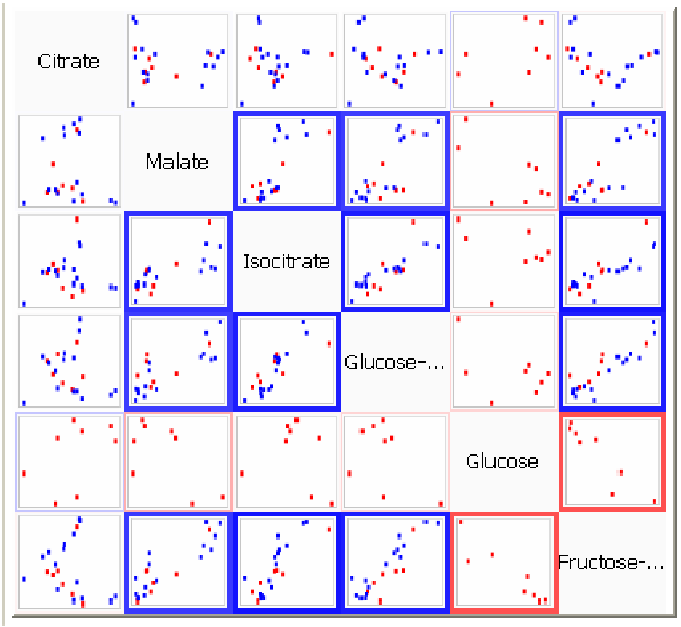


Number of data points decreases with increasing offset (t)

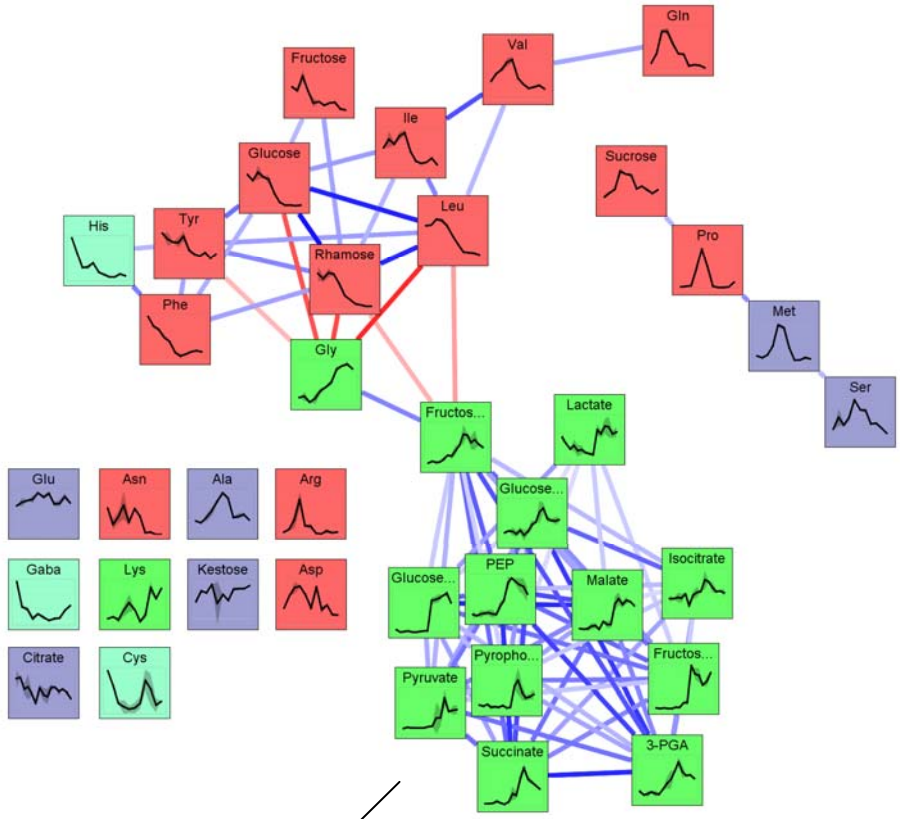
t = 2



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

Thanks to Colleagues

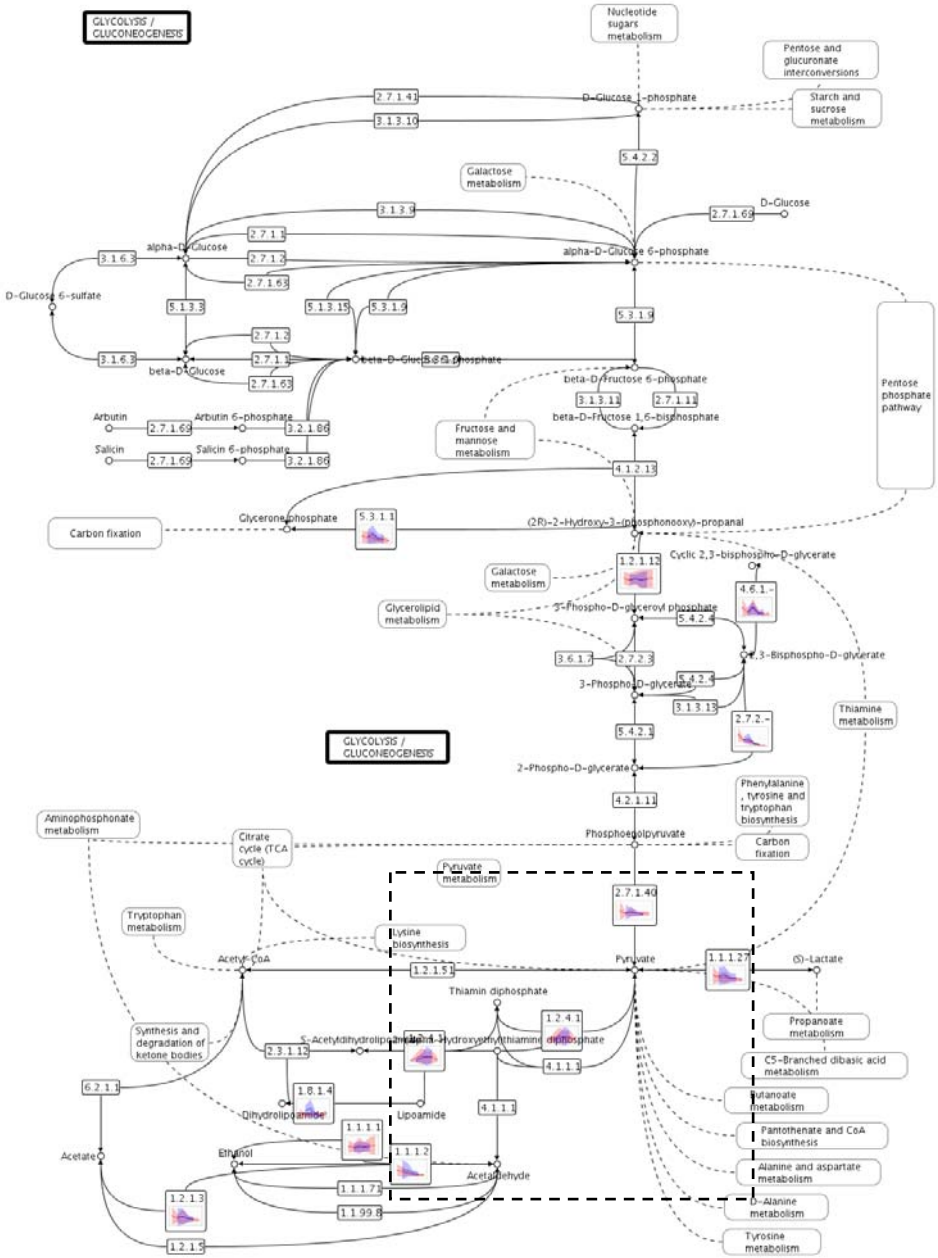
- Ljudmilla Borisjuk, Mohammad-Reza Hajirezaei, Björn Junker, Hardy Rolletschek, Nese Sreenivasulu, Winfriede Weschke, Ruslana Radchuk
  > Discussion of system features and data provision

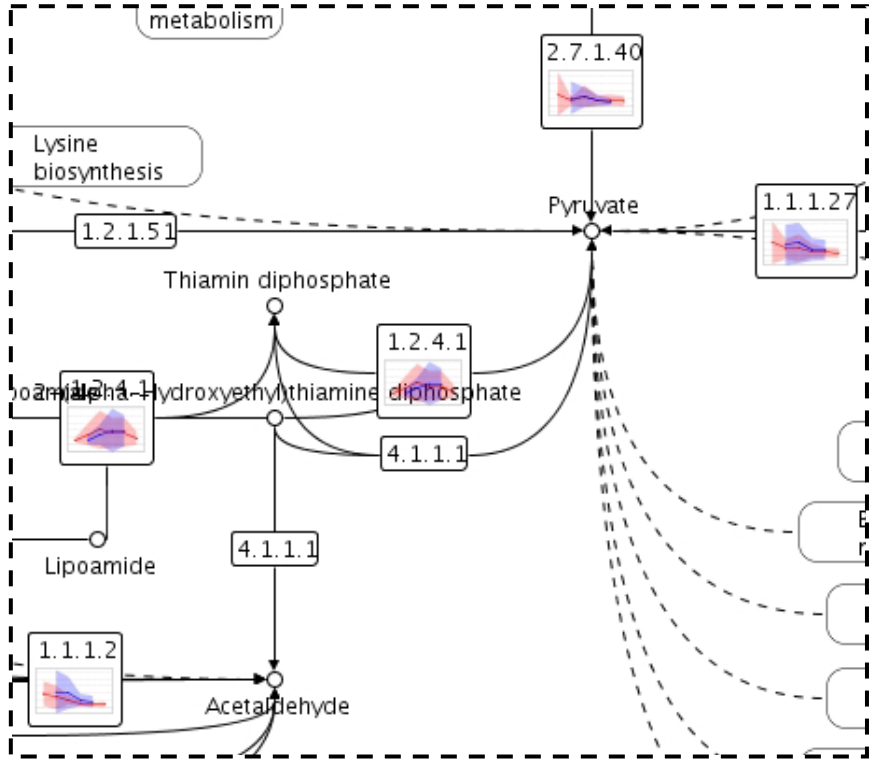
- Matthias Lange, Uwe Scholz, Andreas Stephanik, Karl Spies
  > Database services and SOAP access to FLAREX

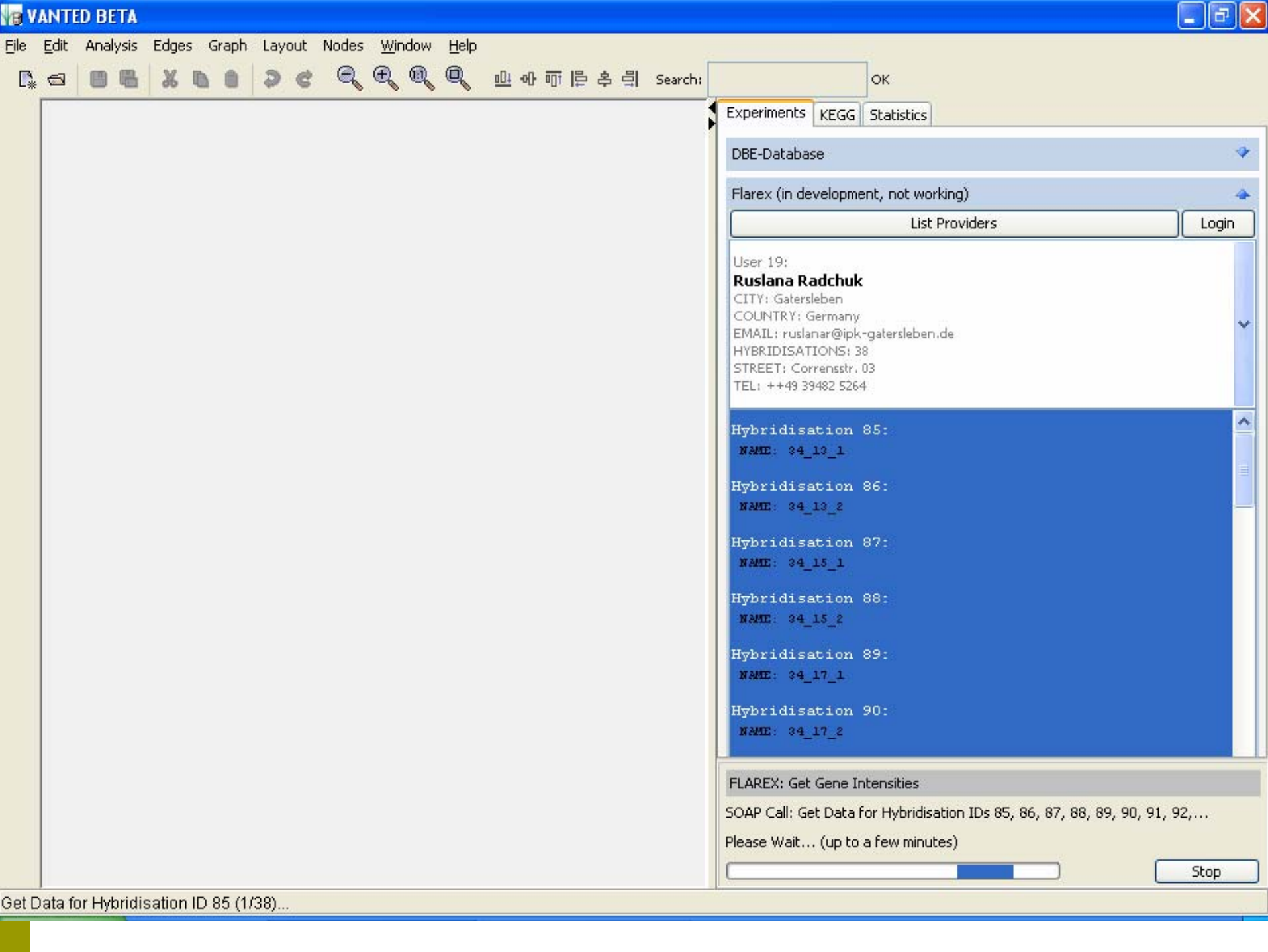
- Dirk Koschützki, Falk Schreiber
  > Support and Inspiration

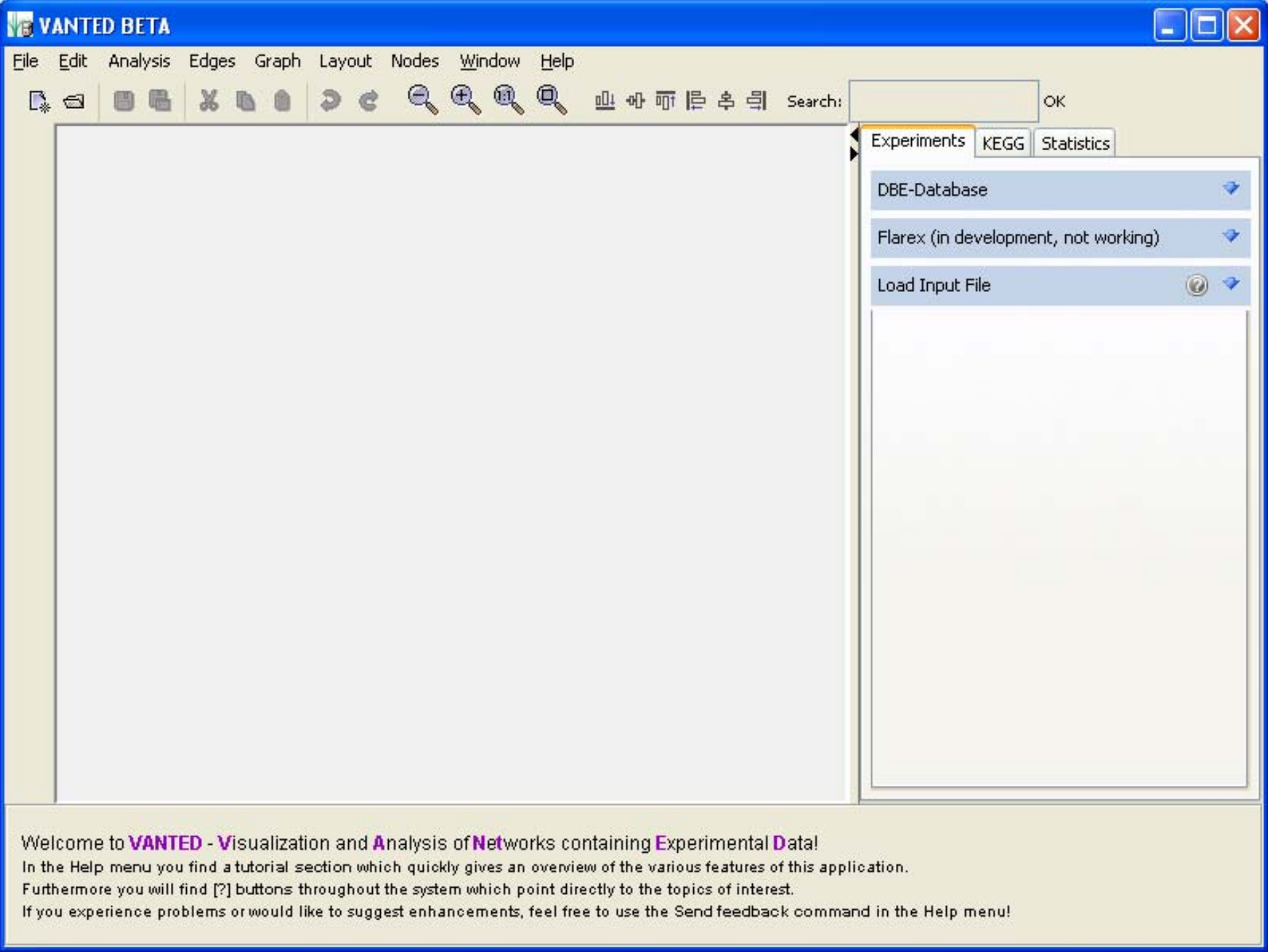
Software Demo

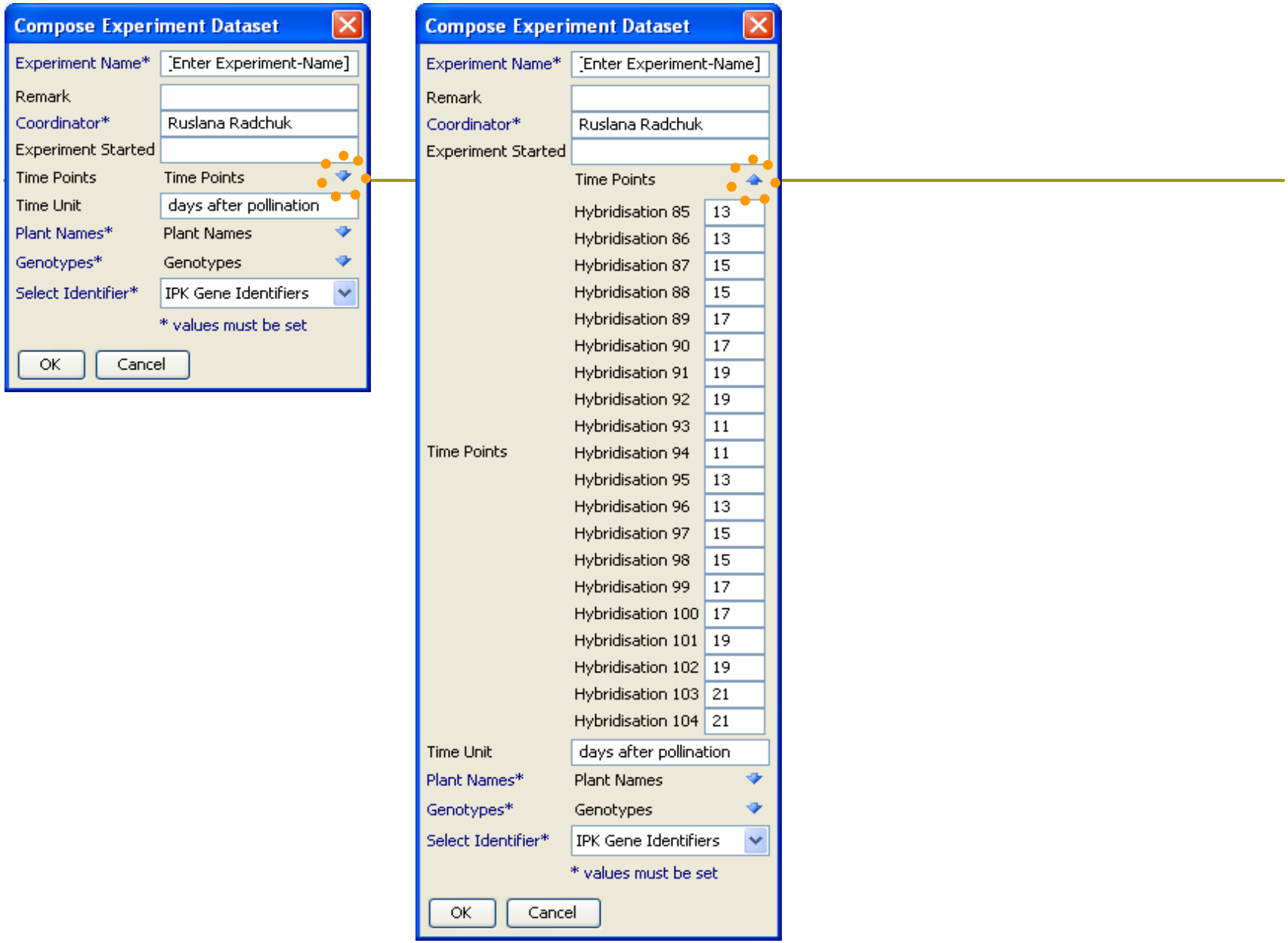
1. Loading of Array Data
   - 23 Hybridizations:
     - Wild type / trangenic line
     - 6 time points
     - ~ 146000 measurement values
     - ~ 3800 clones
2. Filter Dataset
   - Remove 50% of the data with highest average standard deviations
   - Remove 90% of remaining nodes with low differences between wild type and transgenic line
3. Detect Common Developmental Patterns (SOM-Clustering)

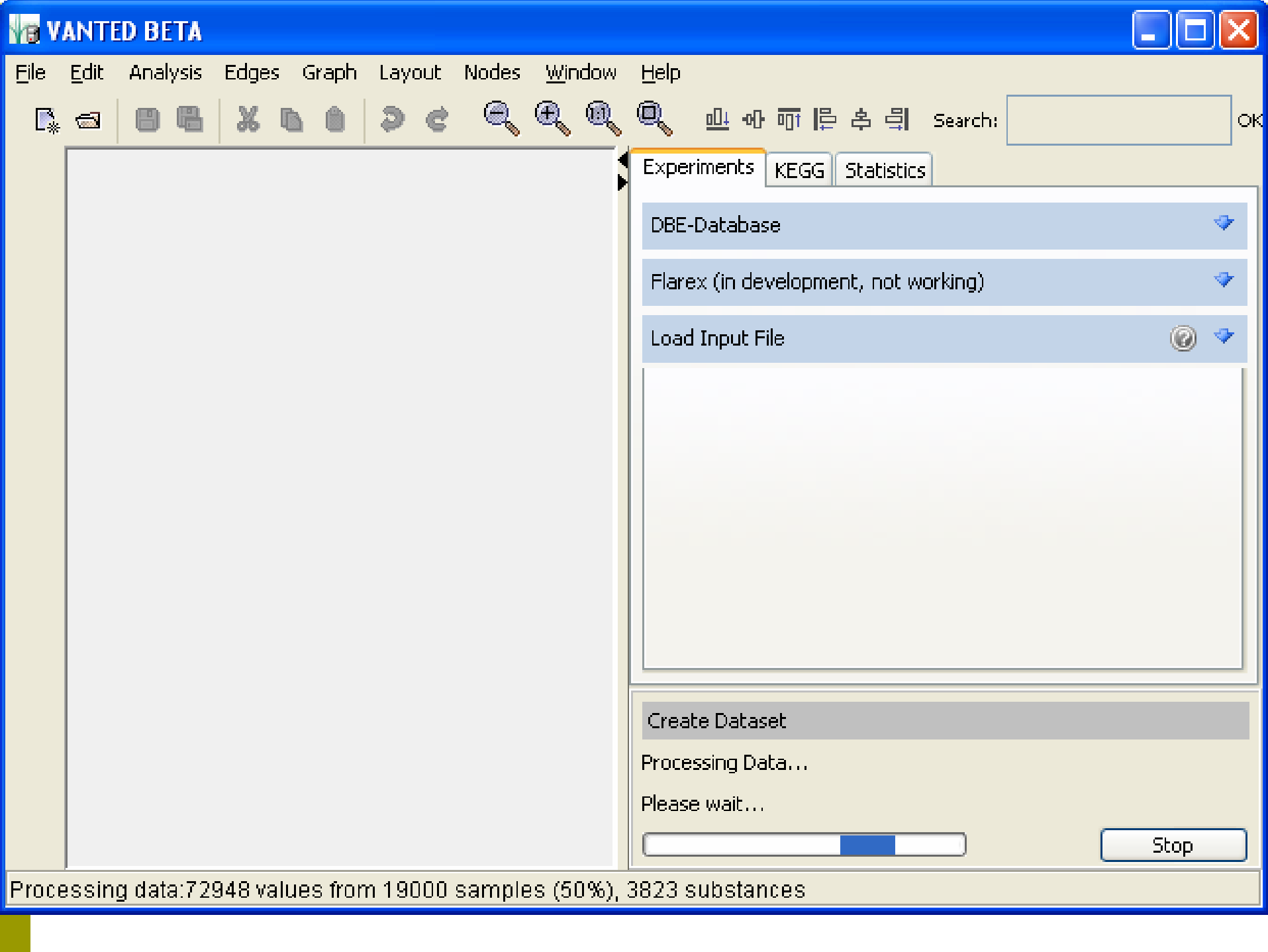

GLYCOLYSIS / GLUCONEOGENESIS
Nucleotide sugars metabolism
Pentose and glucuronate interconversions
Starch and sucrose metabolism
D-Glucose 1-phosphate
2.7.1.41
3.1.3.10
5.4.2.2
Galactose metabolism
3.1.3.9
2.7.1.69
D-Glucose
alpha-D-Glucose
alpha-D-Glucose 6-phosphate
3.1.6.3
2.7.1.1
2.7.1.2
2.7.1.63
D-Glucose 6-sulfate
5.1.3.3
5.1.3.15
5.3.1.9
beta-D-Glucose
beta-D-Glucose 6-phosphate
beta-D-Fructose 6-phosphate
3.1.3.11
2.7.1.11
beta-D-Fructose 1,6-bisphosphate
Pentose phosphate pathway
Arbutin
Arbutin 6-phosphate
Salicin
Salicin 6-phosphate
3.2.1.86
Fructose and mannose metabolism
4.1.2.13
Carbon fixation
Glycerone phosphate
5.3.1.1
(2R)-2-Hydroxy-3-(phosphonooxy)-propanal
Cyclic 2,3-bisphospho-D-glycerate
1.2.1.12
Galactose metabolism
Glycerolipid metabolism
3-Phospho-D-glyceroyl phosphate
5.4.2.4
4.6.1.-
3.6.1.7
2.7.2.3
2,3-Bisphospho-D-glycerate
3-Phospho-D-glycerate
5.4.2.1
3.1.3.13
2.7.2.-
Thiamine metabolism
GLYCOLYSIS / GLUCONEOGENESIS
2-Phospho-D-glycerate
4.2.1.11
Phenylalanine, tyrosine and tryptophan biosynthesis
Aminophosphonate metabolism
Phosphoenolpyruvate
Citrate cycle (TCA cycle)
Carbon fixation
Pyruvate metabolism
2.7.1.40
Tryptophan metabolism
Lysine biosynthesis
Acetyl-CoA
1.2.1.51
Pyruvate
1.1.1.27
(S)-Lactate
Thiamin diphosphate
1.2.4.1
Propanoate metabolism
Synthesis and degradation of ketone bodies
2.3.1.12
S-Acetyldihydrolipoamide-E
alpha-Hydroxyethylthiamine diphosphate
4.1.1.1
C5-Branched dibasic acid metabolism
6.2.1.1
1.8.1.4
Dihydrolipoamide
Lipoamide
4.1.1.1
Butanoate metabolism
1.1.1.1
Pantothenate and CoA biosynthesis
Acetate
Ethanol
1.1.1.2
Acetaldehyde
Alanine and aspartate metabolism
1.2.1.3
1.1.1.71
1.1.99.8
D-Alanine metabolism
1.2.1.5
Tyrosine metabolism

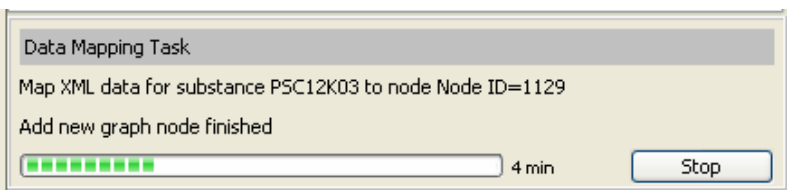

metabolism
2.7.1.40
Lysine biosynthesis
Pyruvate
1.1.1.27
1.2.1.51
Thiamin diphosphate
1.2.4.1
alpha-Hydroxyethylthiamine diphosphate
4.1.1.1
Lipoamide
4.1.1.1
1.1.1.2
Acetaldehyde


VANTED BETA
File Edit Analysis Edges Graph Layout Nodes Window Help
Search:
OK
Experiments
KEGG
Statistics
DBE-Database
Flarex (in development, not working)
List Providers
Login
User 19:
Ruslana Radchuk
CITY: Gatersleben
COUNTRY: Germany
EMAIL: ruslanar@ipk-gatersleben.de
HYBRIDISATIONS: 38
STREET: Corrensstr. 03
TEL: ++49 39482 5264
Hybridisation 85:
NAME: 34_13_1
Hybridisation 86:
NAME: 34_13_2
Hybridisation 87:
NAME: 34_15_1
Hybridisation 88:
NAME: 34_15_2
Hybridisation 89:
NAME: 34_17_1
Hybridisation 90:
NAME: 34_17_2
FLAREX: Get Gene Intensities
SOAP Call: Get Data for Hybridisation IDs 85, 86, 87, 88, 89, 90, 91, 92,...
Please Wait... (up to a few minutes)
Stop
Get Data for Hybridisation ID 85 (1/38)...


VANTED BETA
File
Edit
Analysis
Edges
Graph
Layout
Nodes
Window
Help
Search:
OK
Experiments
KEGG
Statistics
DBE-Database
Flarex (in development, not working)
Load Input File
Welcome to VANTED - Visualization and Analysis of Networks containing Experimental Data!
In the Help menu you find a tutorial section which quickly gives an overview of the various features of this application.
Furthermore you will find [?] buttons throughout the system which point directly to the topics of interest.
If you experience problems or would like to suggest enhancements, feel free to use the Send feedback command in the Help menu!


Compose Experiment Dataset
Experiment Name*
[Enter Experiment-Name]
Remark
Coordinator*
Ruslana Radchuk
Experiment Started
Time Points
Time Points
Time Unit
days after pollination
Plant Names*
Plant Names
Genotypes*
Genotypes
Select Identifier*
IPK Gene Identifiers
* values must be set
OK
Cancel
Compose Experiment Dataset
Experiment Name*
[Enter Experiment-Name]
Remark
Coordinator*
Ruslana Radchuk
Experiment Started
Time Points
Time Points
Hybridisation 85 13
Hybridisation 86 13
Hybridisation 87 15
Hybridisation 88 15
Hybridisation 89 17
Hybridisation 90 17
Hybridisation 91 19
Hybridisation 92 19
Hybridisation 93 11
Hybridisation 94 11
Hybridisation 95 13
Hybridisation 96 13
Hybridisation 97 15
Hybridisation 98 15
Hybridisation 99 17
Hybridisation 100 17
Hybridisation 101 19
Hybridisation 102 19
Hybridisation 103 21
Hybridisation 104 21
Time Unit
days after pollination
Plant Names*
Plant Names
Genotypes*
Genotypes
Select Identifier*
IPK Gene Identifiers
* values must be set
OK
Cancel


VANTED BETA
File
Edit
Analysis
Edges
Graph
Layout
Nodes
Window
Help
Search:
OK
Experiments
KEGG
Statistics
DBE-Database
Flarex (in development, not working)
Load Input File
Create Dataset
Processing Data...
Please wait...
Stop
Processing data:72948 values from 19000 samples (50%), 3823 substances



[Enter Experiment-Name]

Close this Tab

Perform Data Mapping

(if no graph window is open, a node grid will be created in a new editor window)

Experiment Info

Experiment-Name: [Enter Experiment-Name]
Remark:
Coordinator: Ruslana Radchuk
Measurement values: 146092
Import time: Tue Nov 22 11:19:57 CET 2005
Experiment started: Tue Nov 22 11:19:43 CET 2005

Specify Mapping-Data

Time Points

- days after pollination 11
- days after pollination 13
- days after pollination 15
- days after pollination 17
- days after pollination 19

Plants/Genotypes

- 34 ([Enter Genotype]) id=1
- WT ([Enter Genotype]) id=2

Specify Mapping-Options

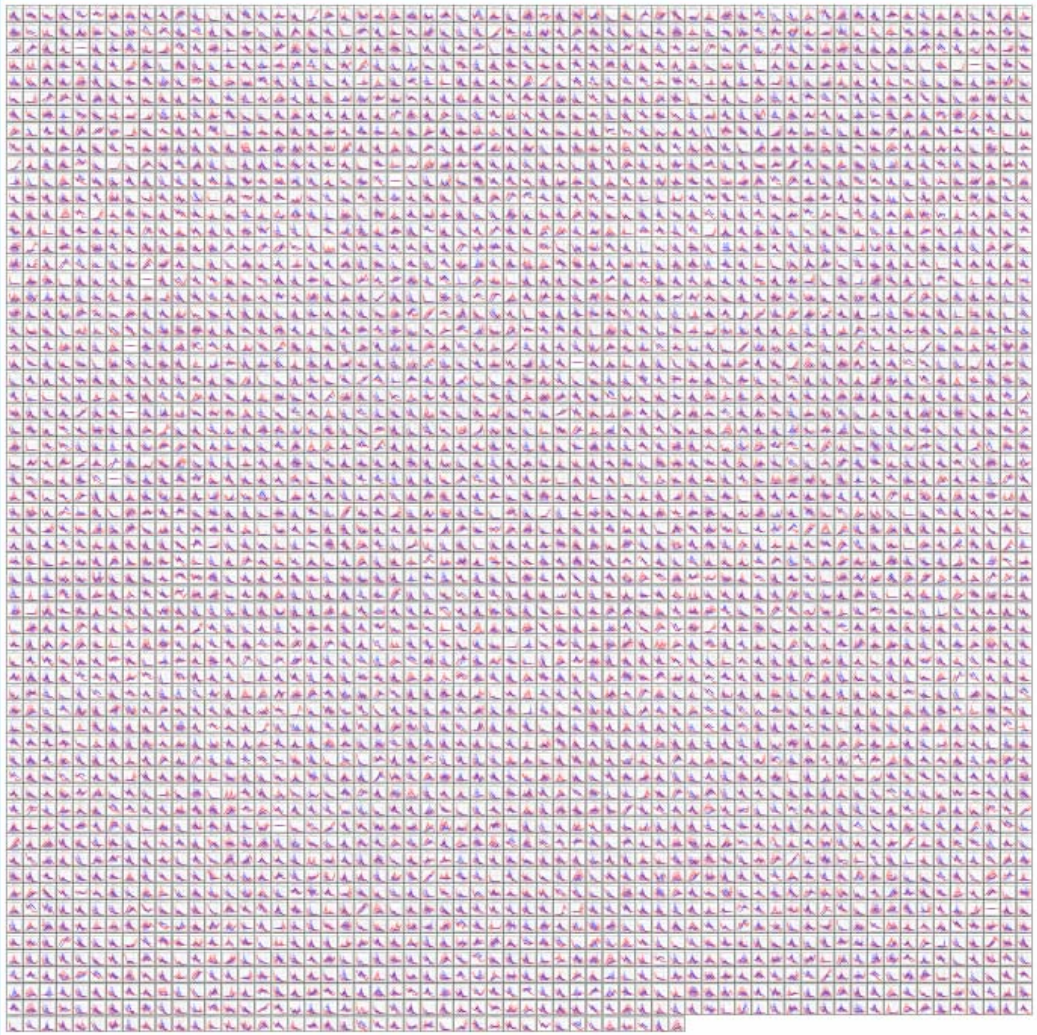
- [ ] Map empty datasets
  (Useful if more than one (incomplete) dataset will be mapped onto the nodes)
- [x] Create new nodes for datasets that can otherwise not be mapped onto the active graph/nodes
- [ ] Ask for user-given mapping if no automatic mapping is possible

VANTED BETA

File Edit Analysis Edges Graph Layout Nodes Window Help

Search: OK

[not saved] * - view 1



3823 nodes no edges

BSH Script
Calculate Average Standard Deviation of Samples

- Goal:
  - Remove 50% of the data with highest average standard deviations

1. Calculate average standard deviation for all clones (graph nodes)

```
//@Nodes:Calculate Average Sample StdDev§
series = node.getMappedSeriesData();
stddevs = new ArrayList();
for (SeriesData sd : series)
    stddevs.addAll(sd.getStdDevValues());
double sum = 0;
int i=0;
for (Double stddev : stddevs)
    sum += stddev;
node.setAttributeValue("script", "avg_stddev",
    new Double(sum/stddevs.size()));
```

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

The standard deviation value represents the average distance of a set of scores from the mean.

BSH Script
Calculate Average Ratio of Sample-Mean Differences

- Goal:
  - Identify clones with high differences between wild type / transgenic line

1. Calculate ratio-difference for all clones (graph nodes)

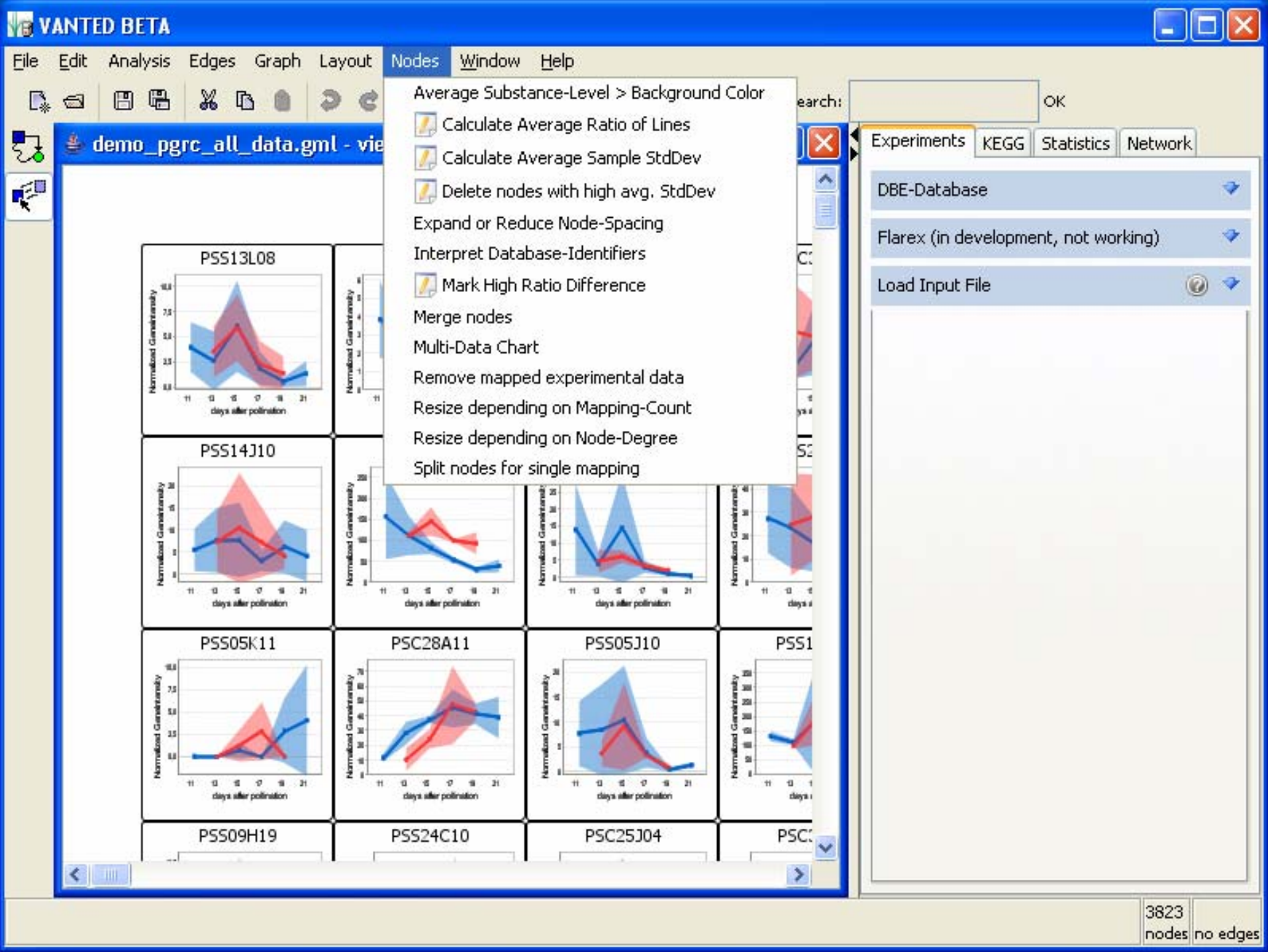
```
//@Nodes:Calculate Average Ratio of Lines§
Set timePoints = node.getMappedTimePointsCoveredByAllLines();
double sum = 0;
int i = 0;
for (int tp : timePoints) {
    ArrayList meanValues = node.getMappedMeanValuesForTimePoint(tp);
    if (meanValues.size()==2) {
          double a = meanValues.get(0);
          double b = meanValues.get(1);
          if (a>b)
                    sum += Math.abs(a/b);
          else
                    sum += Math.abs(b/a);
          i++;
    }
}
node.setAttributeValue("script", "avg_ratio", new Double(sum/i));
```

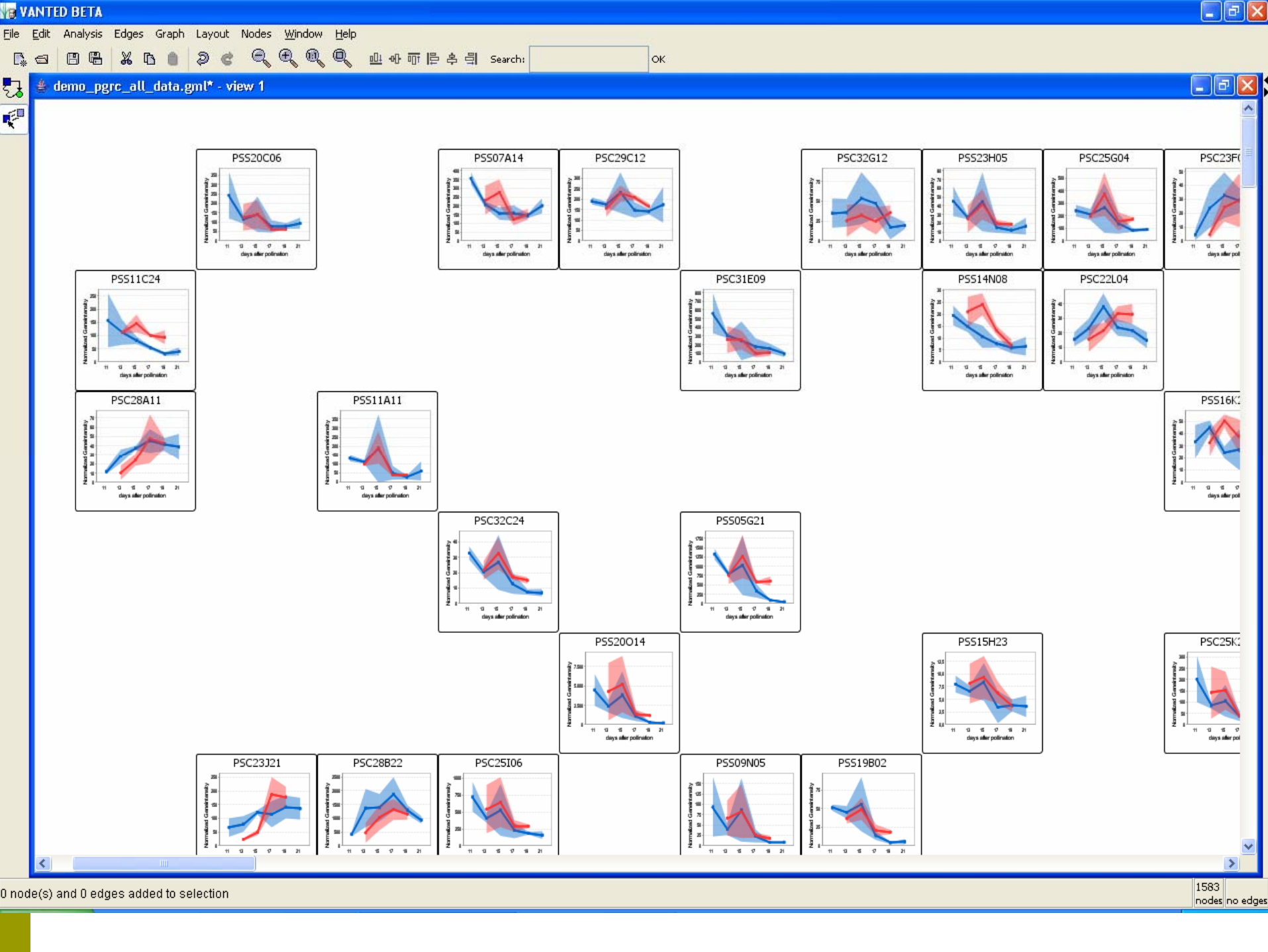
BSH Script
Remove 50% of clones with highest average standard deviation

```
//@Nodes:Delete nodes with high avg. StdDev
TreeMap m = new TreeMap();
for (NodeHelper node : worknodes) {
    Double avgStdDev = node.getAttributeValue("script", "avg_stddev", null,
          new Double(0d));
    if (avgStdDev!=null)
          m.put(avgStdDev, node);
}
int resultCount = (int)(m.size()*0.5d);
int i = 0;
for (node : m.values()) {
    i++;
    if (i>=resultCount)
          graph.deleteNode(node.getGraphNode());
}
```

BSH Script
Remove 90% of clones with low average ratio between WT / transgenic line

```
//@Nodes:Delete Nodes with Low Ratio Difference
TreeMap m = new TreeMap();
for (NodeHelper node : worknodes) {
    Double avgRatio = node.getAttributeValue("script", "avg_ratio", null,
          new Double(0d));
    if (avgRatio!=null)
          m.put(avgRatio, node);
}
int resultCount = (int)(m.size()*0.9d);
int i = 0;
for (node : m.values()) {
    i++;
    if (i<resultCount)
          graph.deleteNode(node.getGraphNode());
}
```


VANTED BETA
File
Edit
Analysis
Edges
Graph
Layout
Nodes
Window
Help
Average Substance-Level > Background Color
Calculate Average Ratio of Lines
Calculate Average Sample StdDev
Delete nodes with high avg. StdDev
Expand or Reduce Node-Spacing
Interpret Database-Identifiers
Mark High Ratio Difference
Merge nodes
Multi-Data Chart
Remove mapped experimental data
Resize depending on Mapping-Count
Resize depending on Node-Degree
Split nodes for single mapping
earch:
OK
demo_pgrc_all_data.gml - vie
PSS13L08
PSS14J10
PSS05K11
PSC28A11
PSS05J10
PSS09H19
PSS24C10
PSC25J04
Experiments
KEGG
Statistics
Network
DBE-Database
Flarex (in development, not working)
Load Input File
3823 nodes
no edges

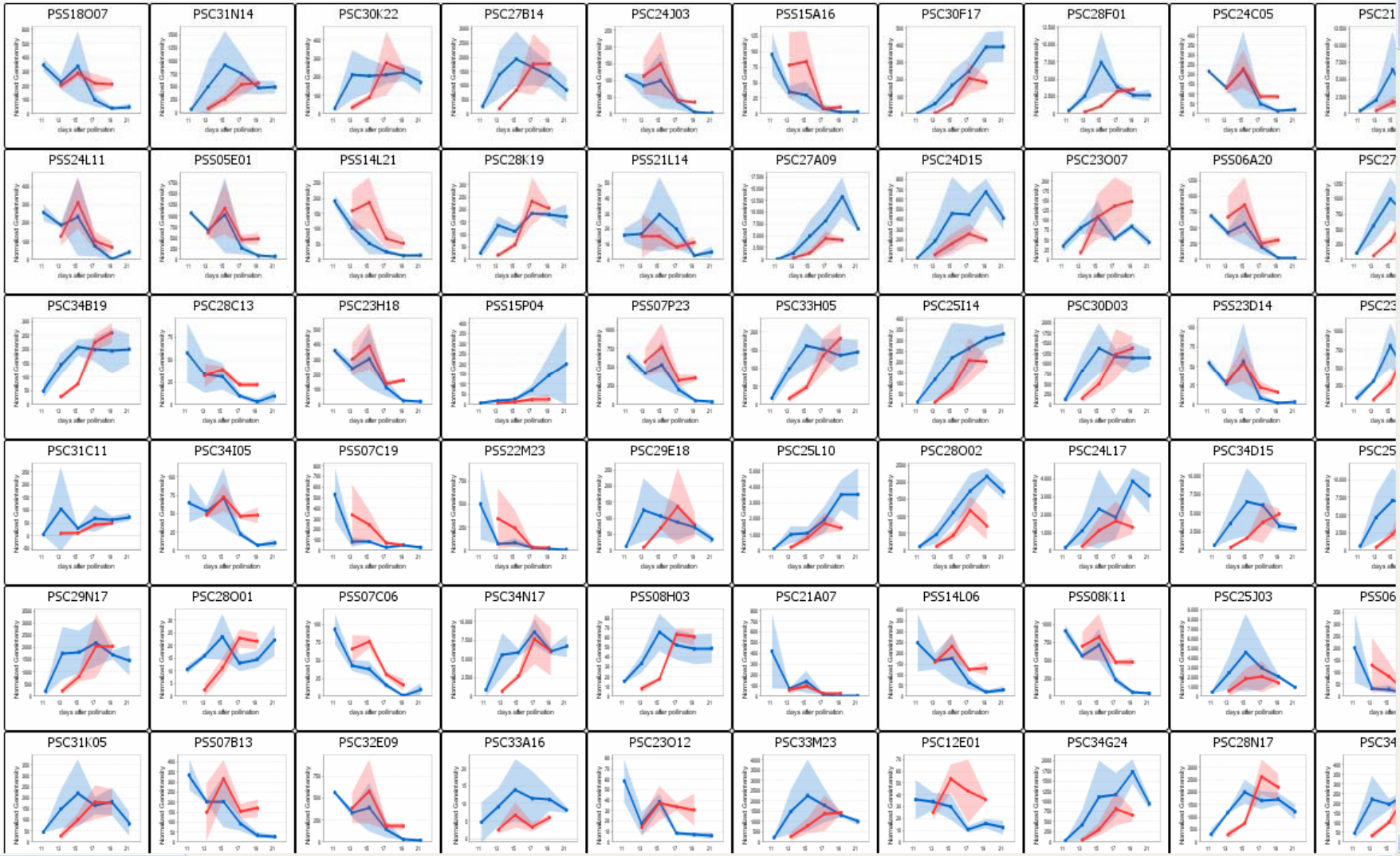

VANTED BETA
File Edit Analysis Edges Graph Layout Nodes Window Help
Search:
OK
demo_pgrc_all_data.gml* - view 1
PSS20C06
PSS07A14
PSC29C12
PSC32G12
PSS23H05
PSC25G04
PSS11C24
PSC31E09
PSS14N08
PSC22L04
PSC28A11
PSS11A11
PSC32C24
PSS05G21
PSS20O14
PSS15H23
PSC23J21
PSC28B22
PSC25I06
PSS09N05
PSS19B02
0 node(s) and 0 edges added to selection
1583 nodes
no edges

VANTED BETA

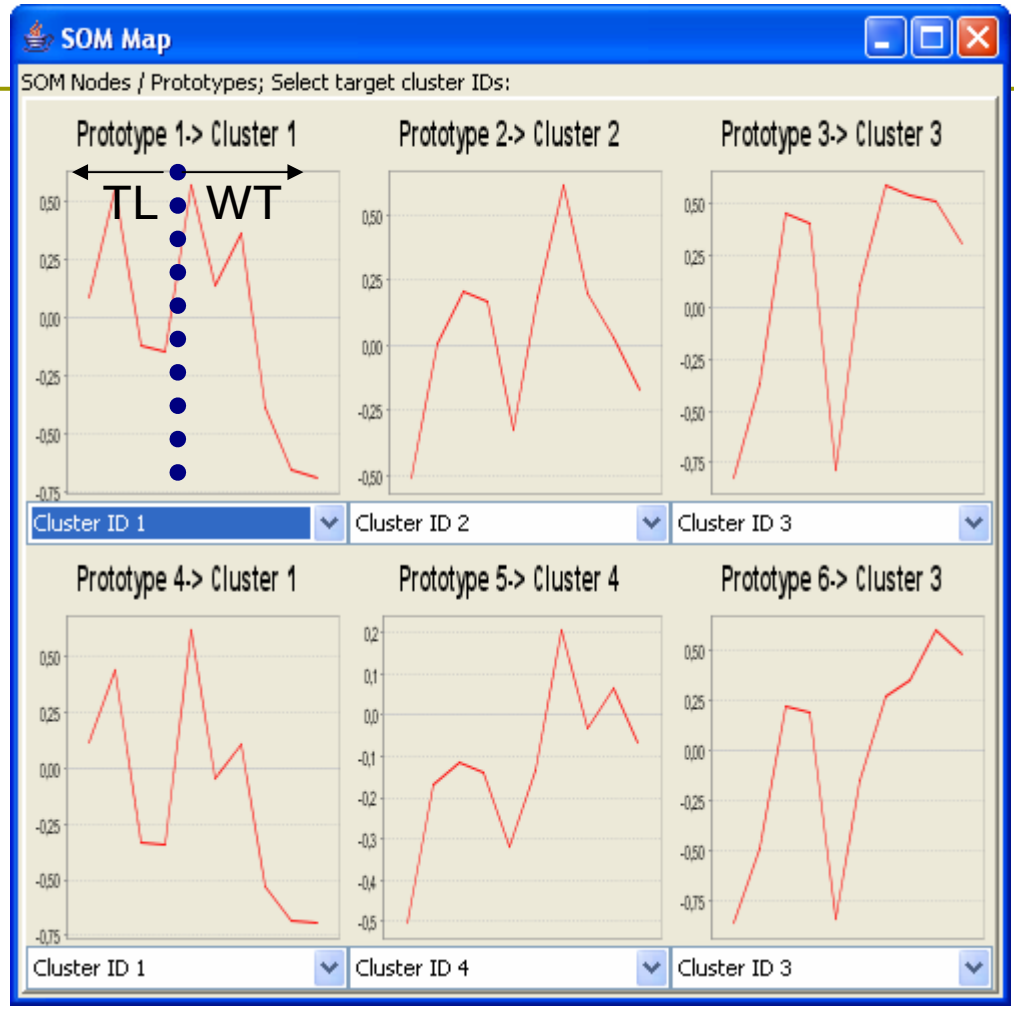
File Edit Analysis Edges Graph Layout Nodes Window Help

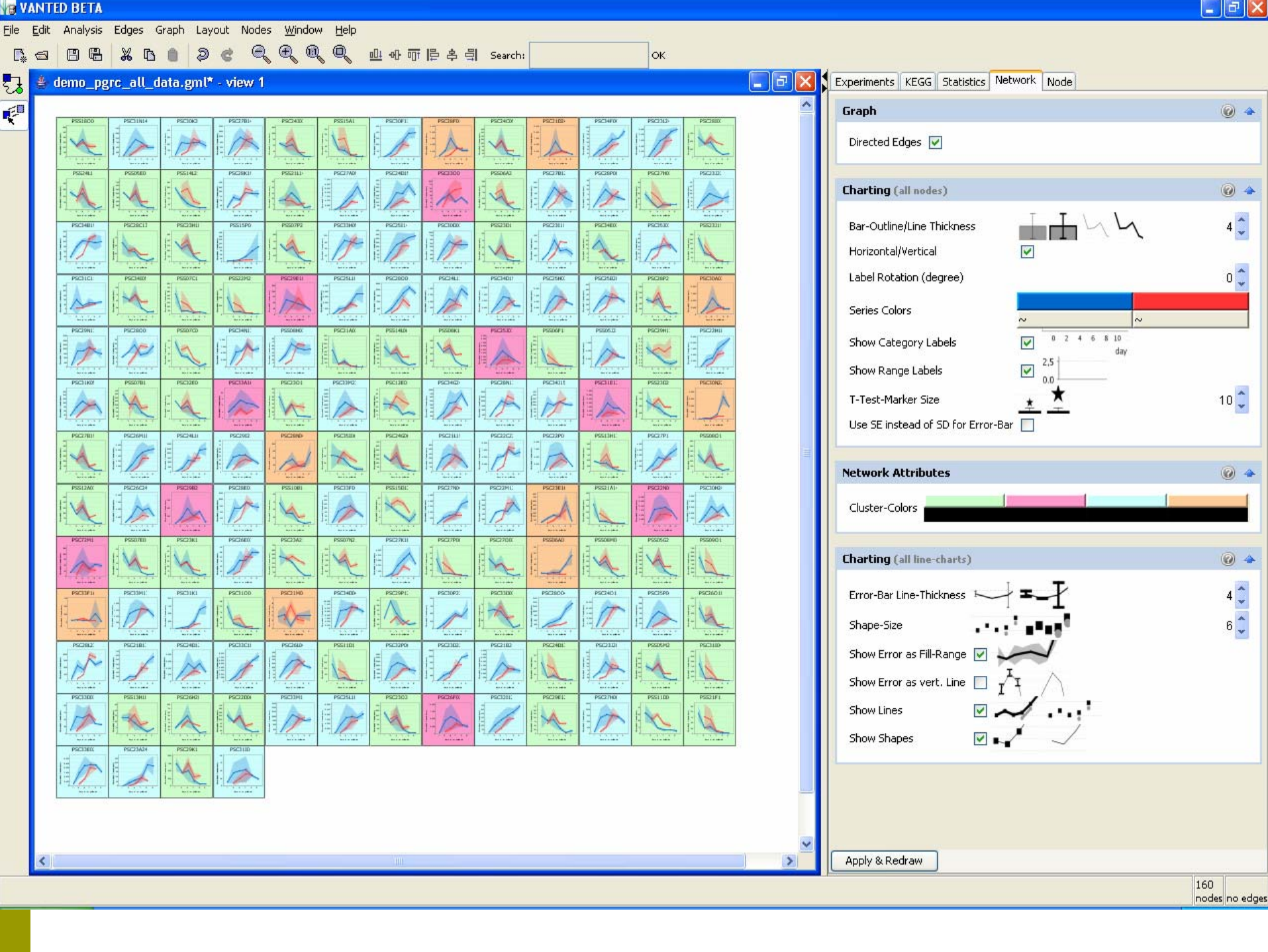
Search: OK

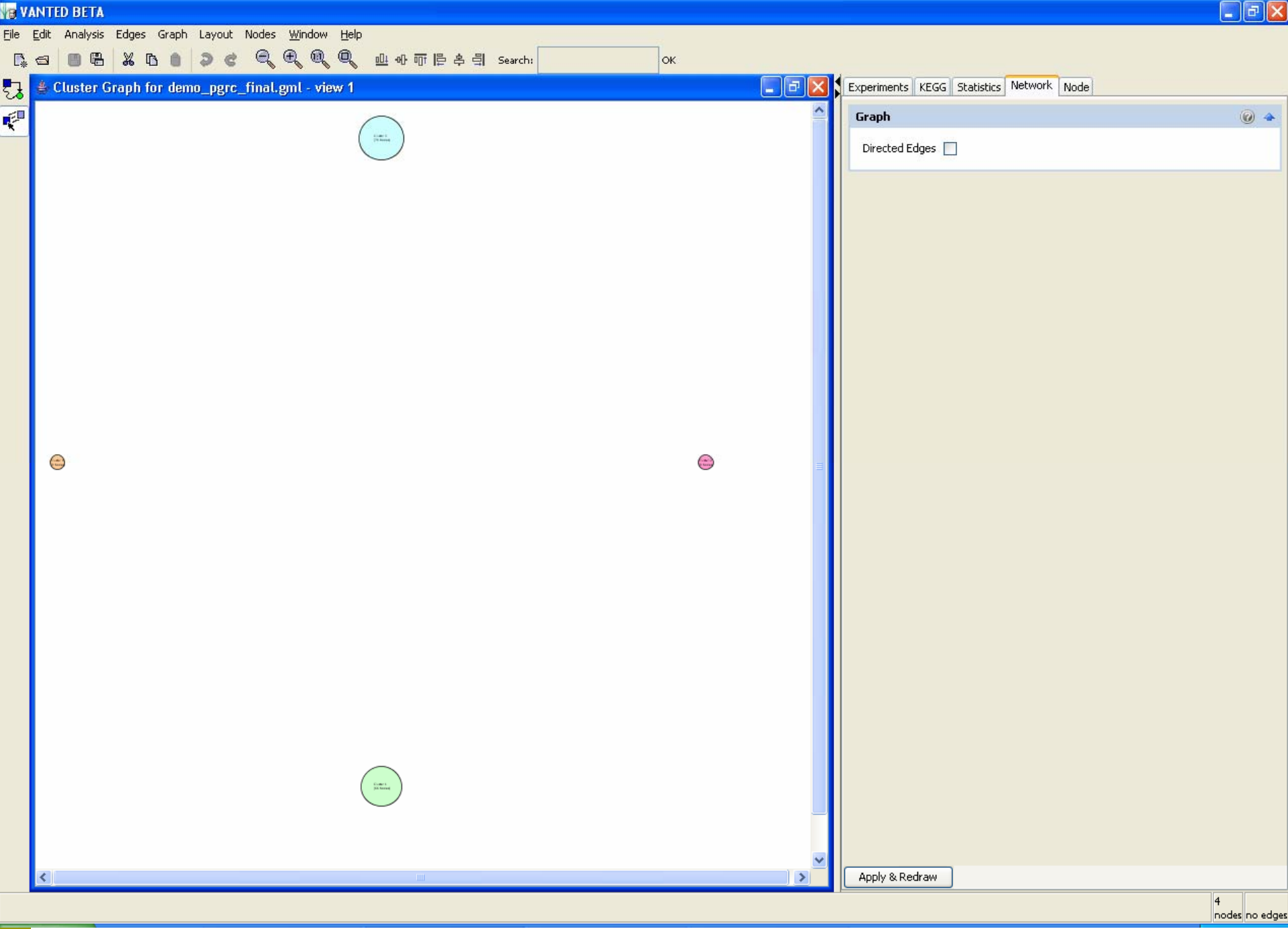
demo_pgrc_all_data.gml* - view 1

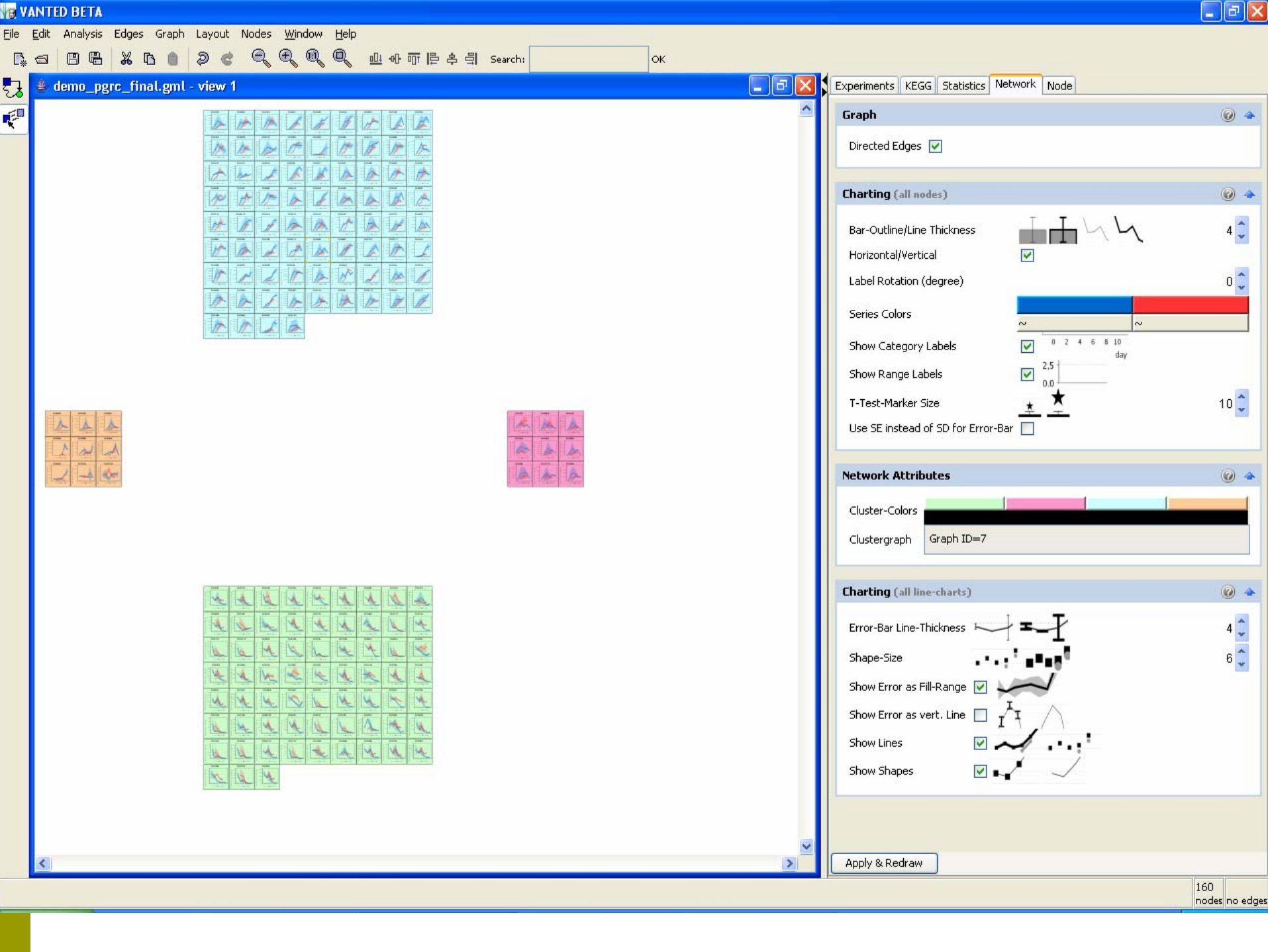


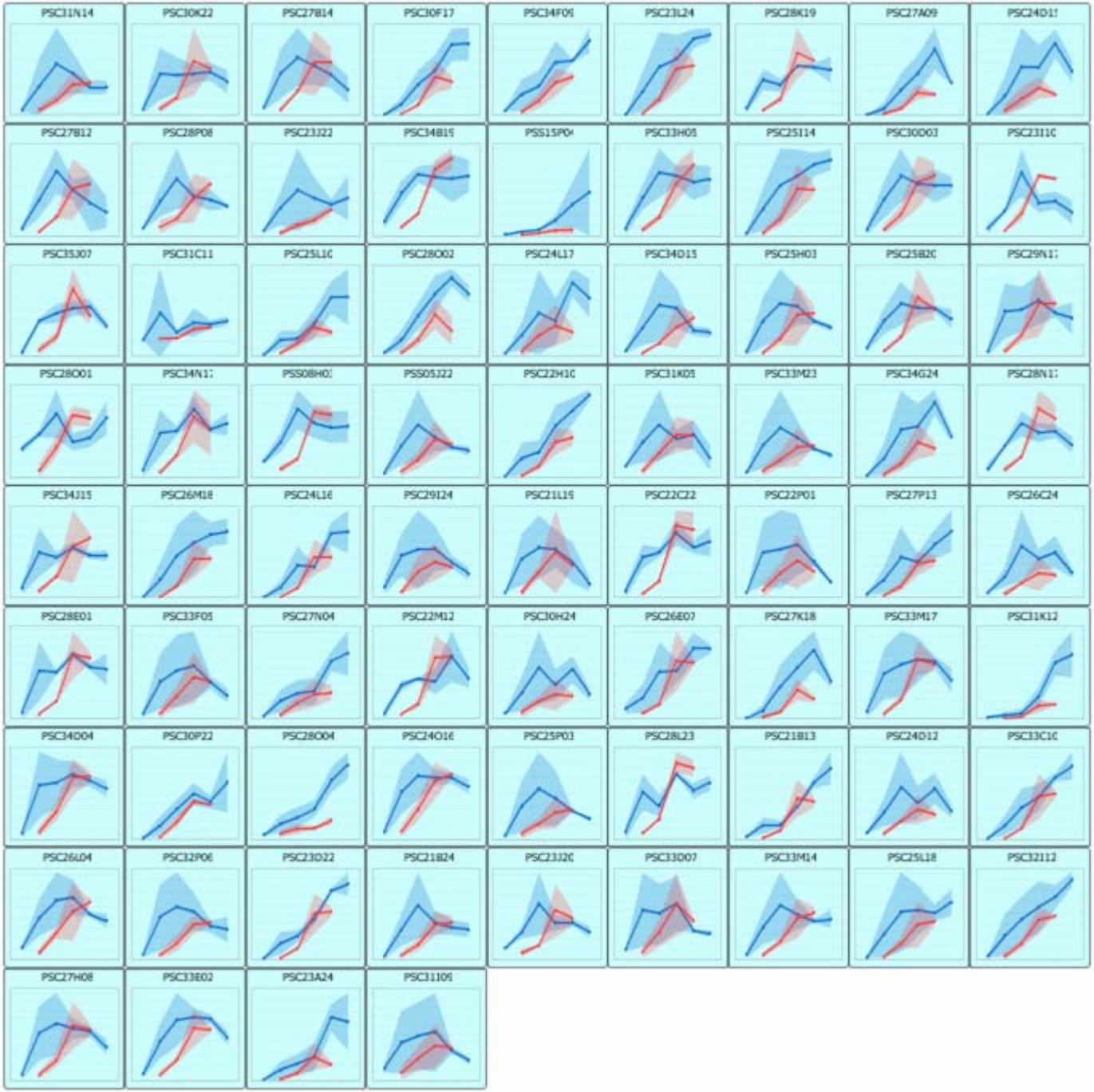
160 nodes no edges

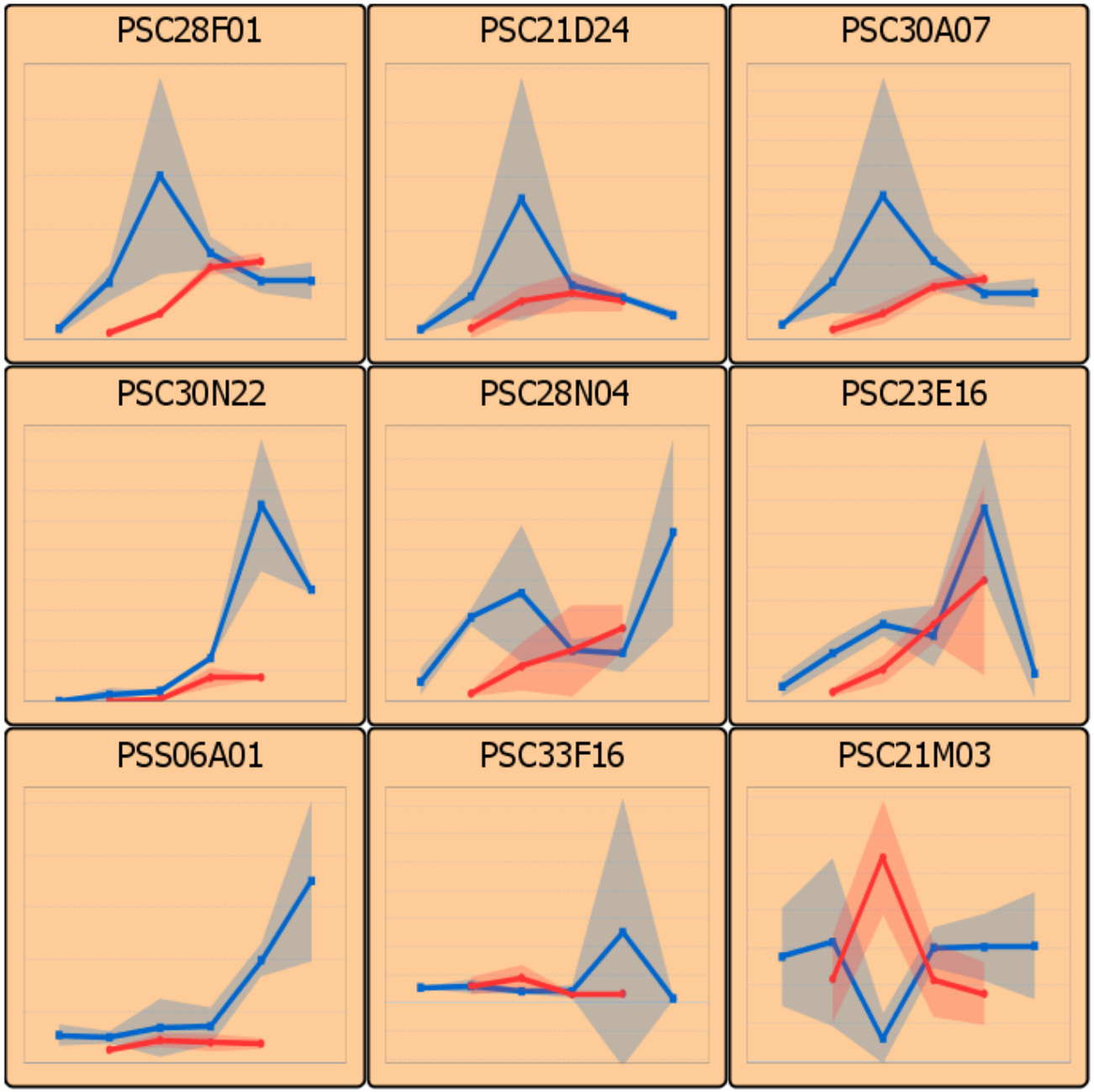

SOM Map
SOM Nodes / Prototypes; Select target cluster IDs:
Prototype 1-> Cluster 1
TL
WT
Prototype 2-> Cluster 2
Prototype 3-> Cluster 3
Cluster ID 1
Cluster ID 2
Cluster ID 3
Prototype 4-> Cluster 1
Prototype 5-> Cluster 4
Prototype 6-> Cluster 3
Cluster ID 1
Cluster ID 4
Cluster ID 3

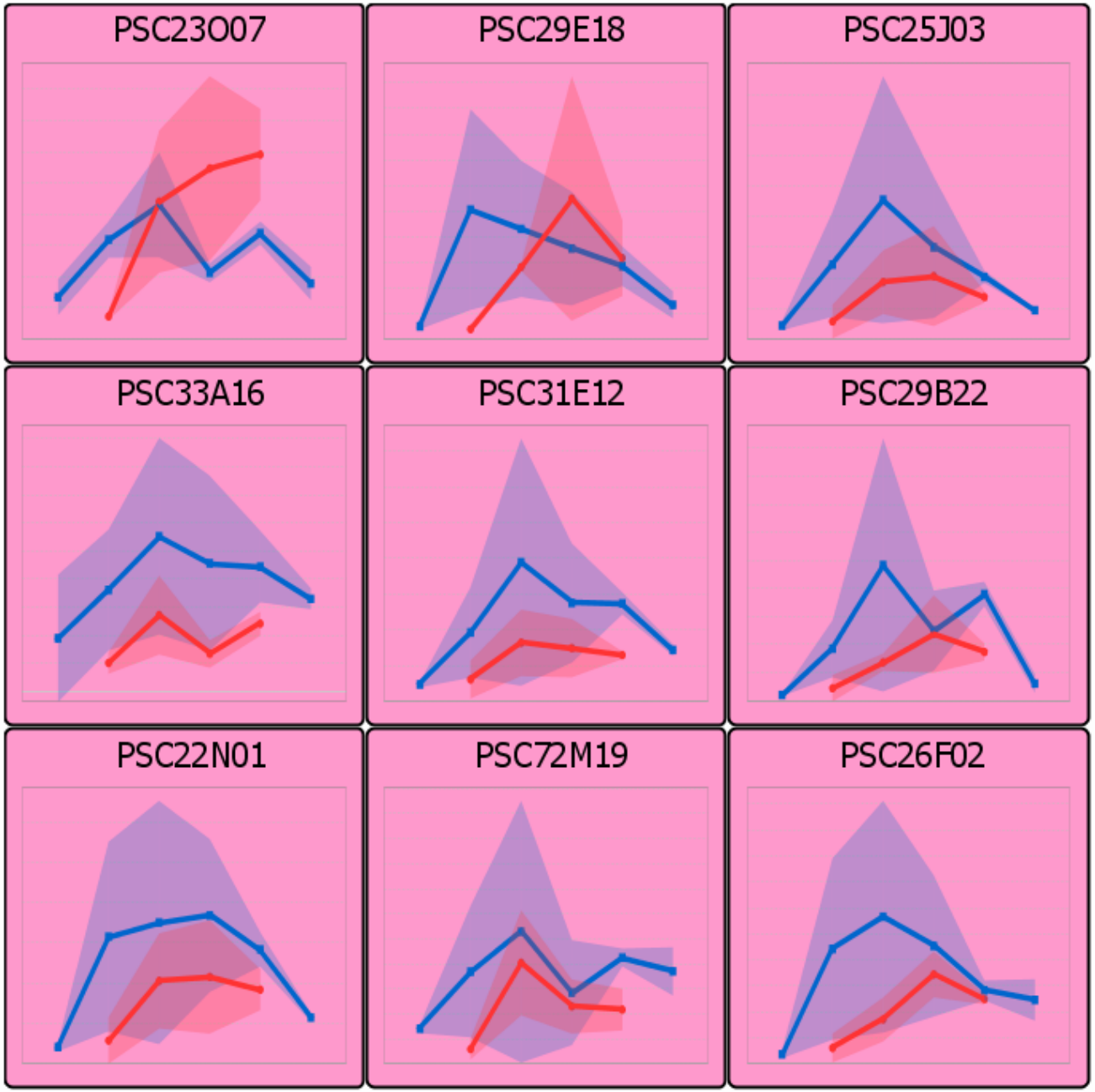

VANTED BETA
File Edit Analysis Edges Graph Layout Nodes Window Help
Search:
OK
demo_pgrc_all_data.gmt* - view 1
Experiments KEGG Statistics Network Node
Graph
Directed Edges
Charting (all nodes)
Bar-Outline/Line Thickness
4
Horizontal/Vertical
Label Rotation (degree)
0
Series Colors
Show Category Labels
Show Range Labels
T-Test-Marker Size
10
Use SE instead of SD for Error-Bar
Network Attributes
Cluster-Colors
Charting (all line-charts)
Error-Bar Line-Thickness
4
Shape-Size
6
Show Error as Fill-Range
Show Error as vert. Line
Show Lines
Show Shapes
Apply & Redraw
160 nodes
no edges


VANTED BETA
File Edit Analysis Edges Graph Layout Nodes Window Help
Search:
OK
Cluster Graph for demo_pgrc_final.gml - view 1
Experiments KEGG Statistics Network Node
Graph
Directed Edges
Apply & Redraw
4 nodes
no edges

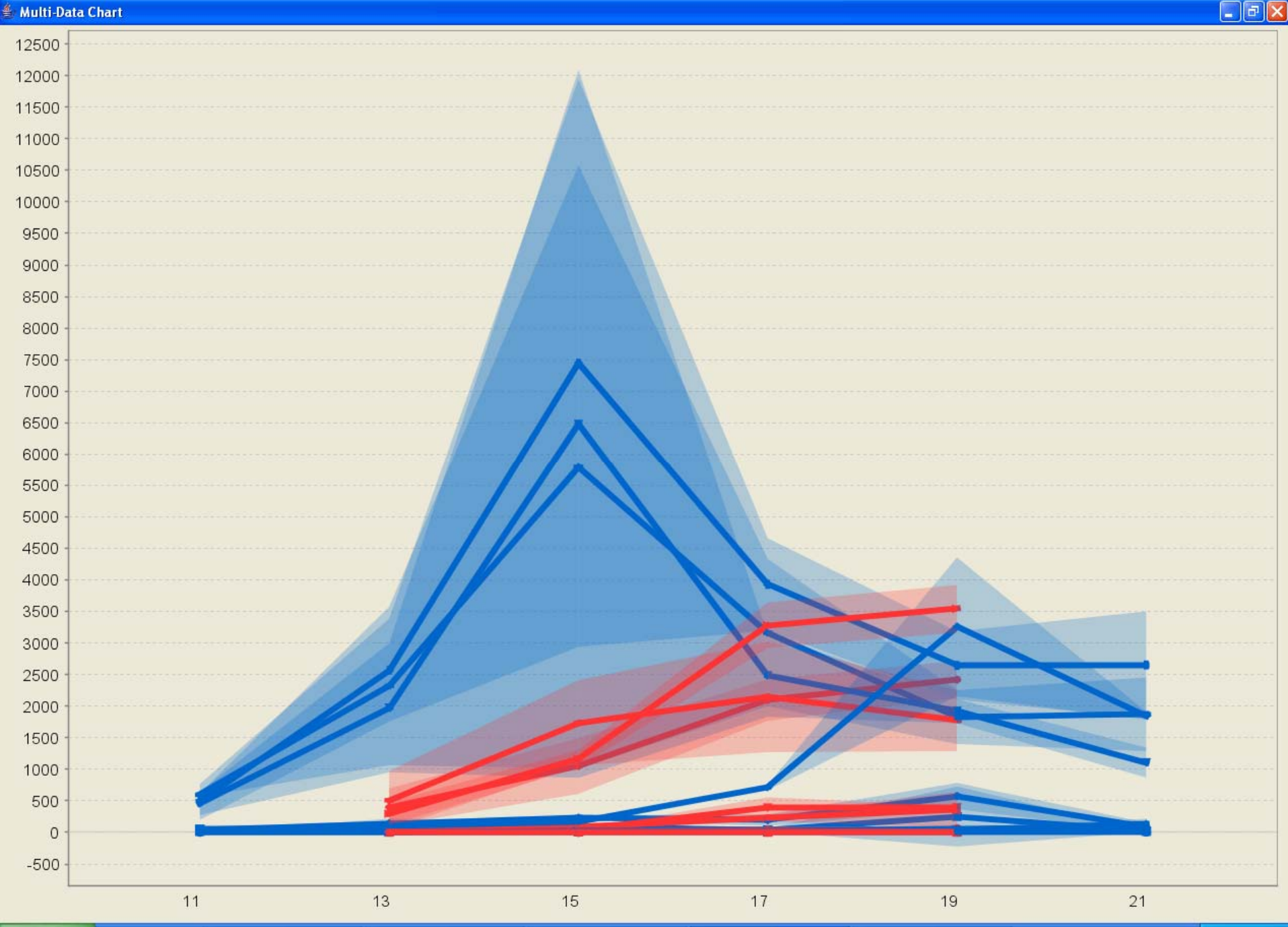

VANTED BETA
File Edit Analysis Edges Graph Layout Nodes Window Help
Search:
OK
demo_pgrc_final.gml - view 1
Experiments KEGG Statistics Network Node
Graph
Directed Edges
Charting (all nodes)
Bar-Outline/Line Thickness
4
Horizontal/Vertical
Label Rotation (degree)
0
Series Colors
Show Category Labels
Show Range Labels
T-Test-Marker Size
10
Use SE instead of SD for Error-Bar
Network Attributes
Cluster-Colors
Clustergraph
Graph ID=7
Charting (all line-charts)
Error-Bar Line-Thickness
4
Shape-Size
6
Show Error as Fill-Range
Show Error as vert. Line
Show Lines
Show Shapes
Apply & Redraw
160 nodes
no edges


PSC31N14
PSC30K22
PSC27B14
PSC30F17
PSC34F06
PSC23L24
PSC28K19
PSC27A09
PSC24D17
PSC27B12
PSC28P08
PSC23J22
PSC34B15
PSS15P04
PSC33H05
PSC25I14
PSC30D03
PSC23I1C
PSC35J07
PSC31C11
PSC25L1C
PSC28O02
PSC24L17
PSC34O15
PSC25H03
PSC25B2C
PSC29N17
PSC28O01
PSC34N17
PSS08H0
PSS05J22
PSC22H1C
PSC31K05
PSC33M23
PSC34G24
PSC28N17
PSC34J15
PSC26M18
PSC24L16
PSC29I24
PSC21L15
PSC22C22
PSC22P01
PSC27P13
PSC26C24
PSC28E01
PSC33F05
PSC27N04
PSC22M12
PSC30H24
PSC26E07
PSC27K18
PSC33M17
PSC31K12
PSC34O04
PSC30P22
PSC28O04
PSC24O16
PSC25P03
PSC28L23
PSC21B13
PSC24D12
PSC33C1C
PSC26L04
PSC32P0E
PSC23D22
PSC21B24
PSC23J2C
PSC33D07
PSC33M14
PSC25L18
PSC32112
PSC27H08
PSC33E02
PSC23A24
PSC31105


PSC28F01
PSC21D24
PSC30A07
PSC30N22
PSC28N04
PSC23E16
PSS06A01
PSC33F16
PSC21M03


PSC23O07
PSC29E18
PSC25J03
PSC33A16
PSC31E12
PSC29B22
PSC22N01
PSC72M19
PSC26F02


PSS18O07
PSC24J03
PSS15A16
PSC24C05
PSC28I03
PSS24L11
PSS05E01
PSS14L21
PSS21L14
PSS06A20
PSC27H07
PSC28C13
PSC23H18
PSS07P23
PSS23D14
PSC34E03
PSS23J19
PSC34I05
PSS07C19
PSS22M23
PSC28P23
PSS07C06
PSC21A07
PSS14L06
PSS08K11
PSS06F11
PSC29H17
PSS07B13
PSC32E09
PSC23O12
PSC12E01
PSS23D21
PSC27B19
PSC35I06
PSC24G06
PSS13H17
PSS08O11
PSS12A07
PSS10B16
PSS15D12
PSS21A14
PSS07E05
PSC23K18
PSC23A21
PSS07N21
PSC27P08
PSC27O03
PSS08M09
PSS05G21
PSS09O11
PSC31O09
PSC29P12
PSC33D03
PSC26O10
PSS11D15
PSC24D13
PSS05M22
PSC31I04
PSS13H18
PSC26H20
PSC22D06
PSC23O22
PSC29E12
PSS11D05
PSS21F16
PSC29K14


Multi-Data Chart
12500
12000
11500
11000
10500
10000
9500
9000
8500
8000
7500
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
-500
11
13
15
17
19
21