Analysis of Biological Networks and Related Data

Falk Schreiber

Summer school of the German-Russian Virtual Network on Computational Systems Biology
Novosibirsk, June 2008

Leibniz Institute of Plant Genetics
and Crop Plant Research (IPK)
Gatersleben

Institute of Computer Science
Martin-Luther-University Halle-Wittenberg
Halle

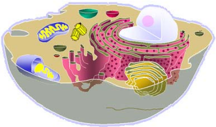
Analysis of Biological Networks

1. Motivation
2. Foundations
3. Network motifs
4. Network centralities
5. Network-related data analysis and visualisation

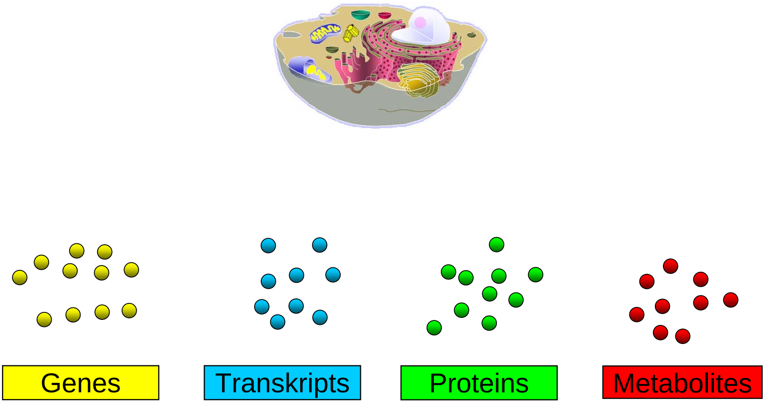
Analysis of Biological Networks

1. Motivation
2. Foundations
3. Network motifs
4. Network centralities
5. Network-related data analysis and visualisation

From Biological Building Blocks to Complex Systems



From Biological Building Blocks to Complex Systems



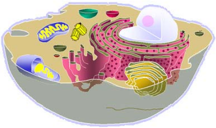
Genes

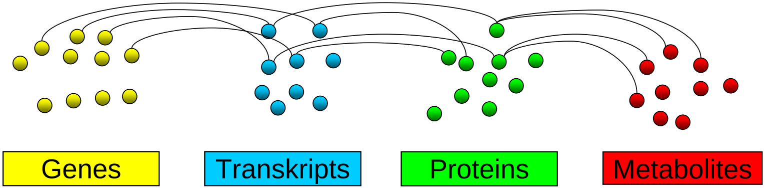
Transkripts

Proteins

Metabolites

From Biological Building Blocks to Complex Systems





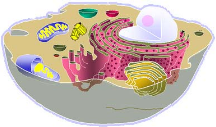
Genes

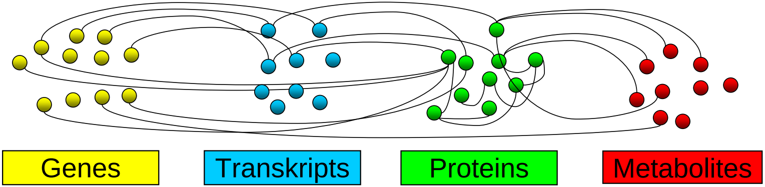
Transkripts

Proteins

Metabolites

From Biological Building Blocks to Complex Systems





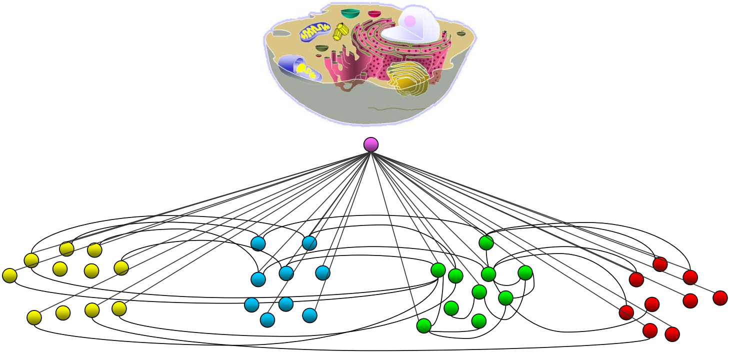
Genes

Transkripts

Proteins

Metabolites

From Biological Building Blocks to Complex Systems



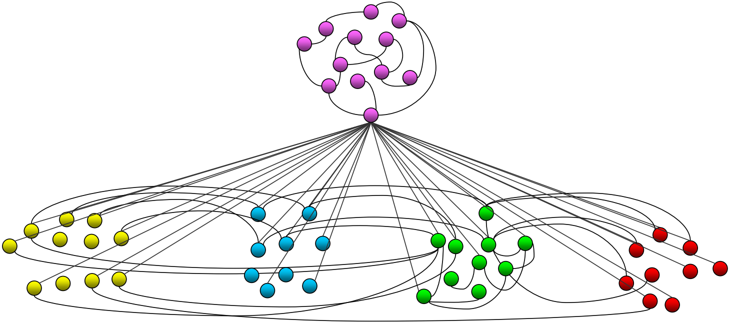
Genes

Transkripts

Proteins

Metabolites

From Biological Building Blocks to Complex Systems



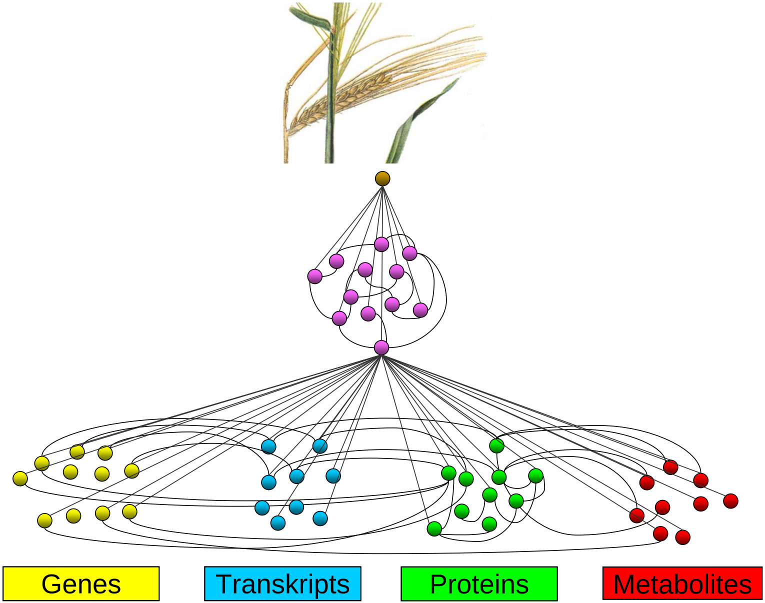
Genes

Transkripts

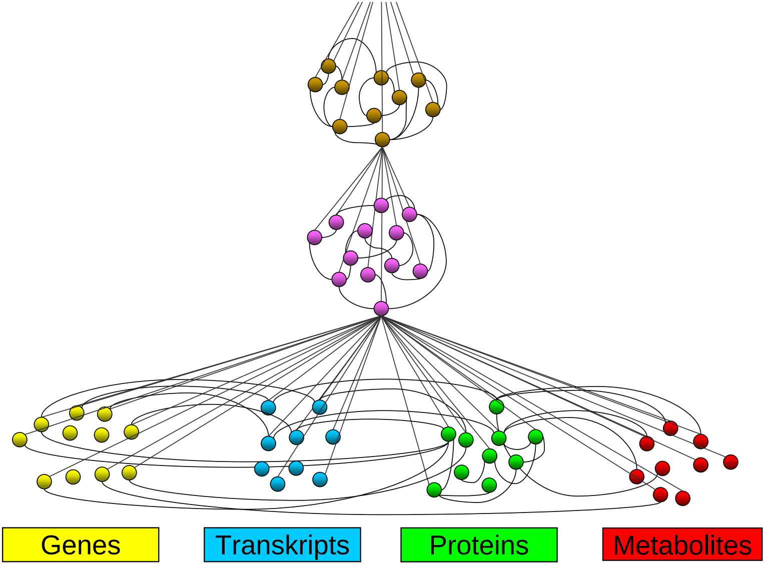
Proteins

Metabolites

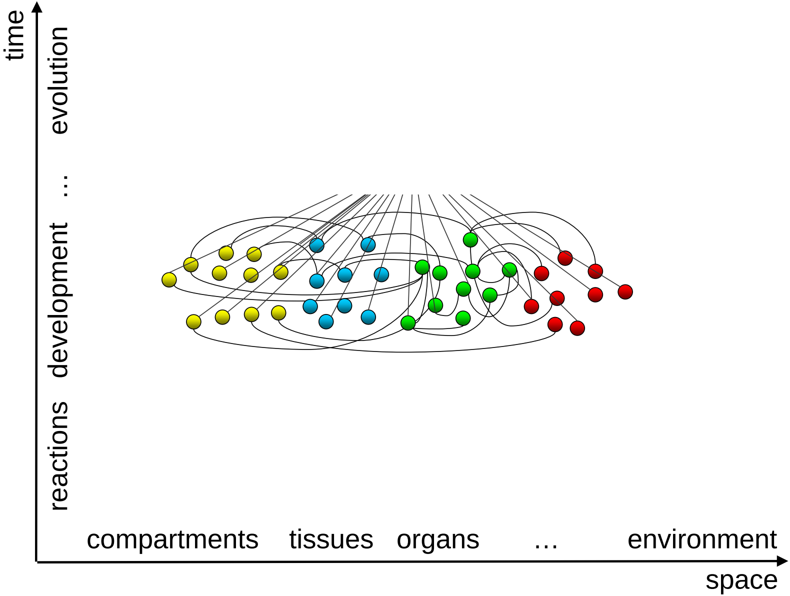
From Biological Building Blocks to Complex Systems



From Biological Building Blocks to Complex Systems



Embedding in Space and Time



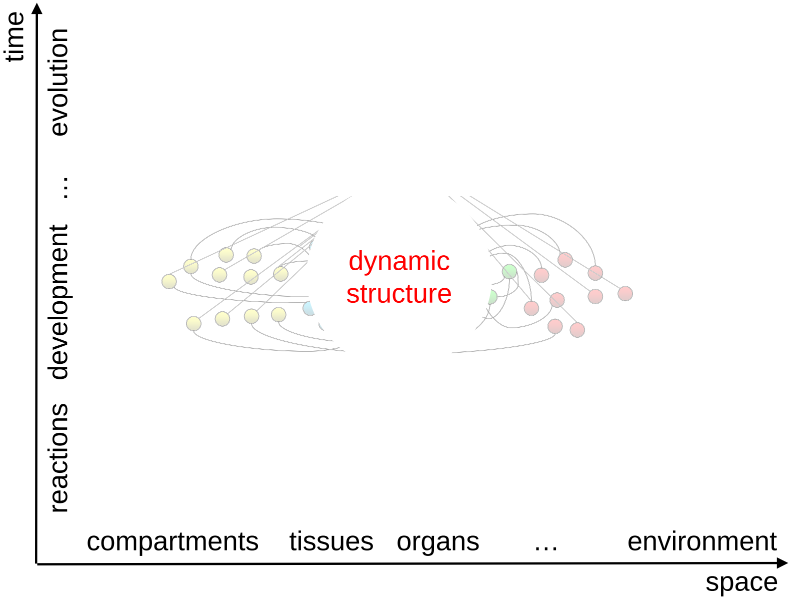
time

reactions development … evolution

compartments tissues organs … environment

space

Embedding in Space and Time



time

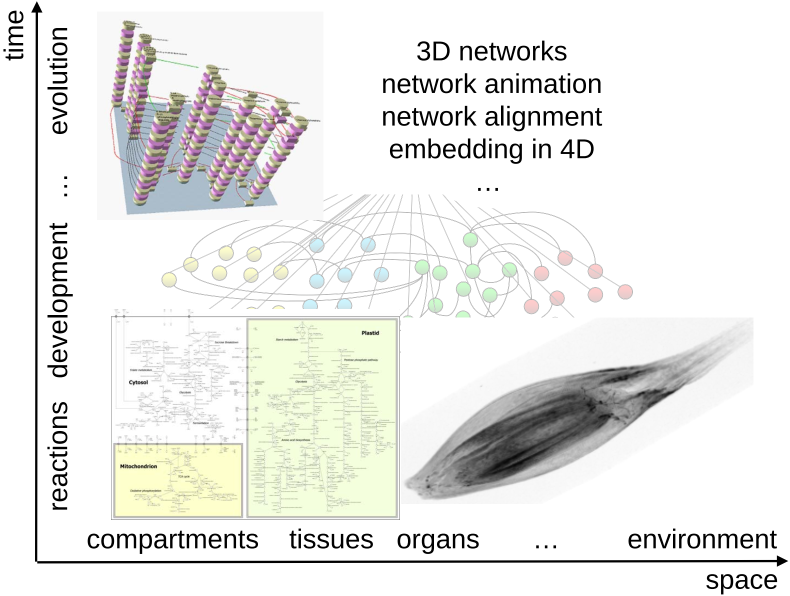
reactions development … evolution

dynamic structure

compartments tissues organs … environment

space

Embedding in Space and Time



time

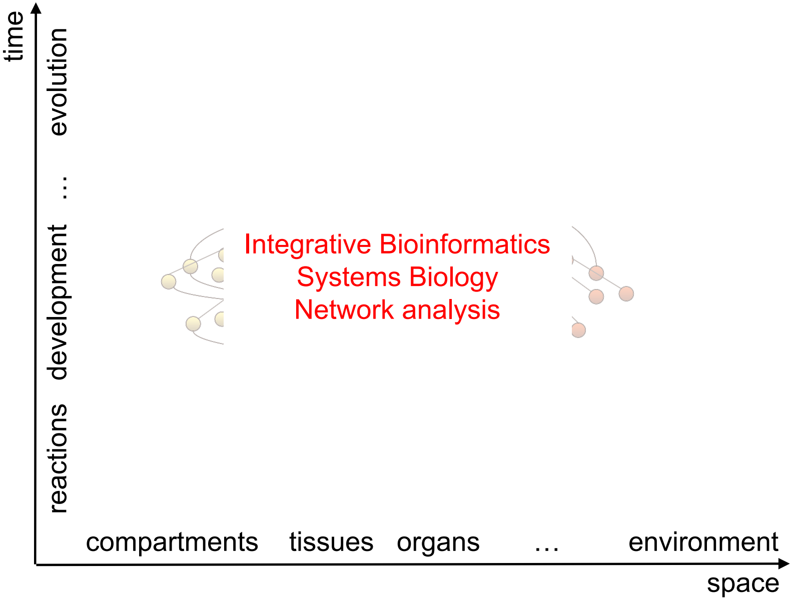
evolution … development reactions

3D networks
network animation
network alignment
embedding in 4D
…

compartments tissues organs … environment

space

Embedding in Space and Time



time

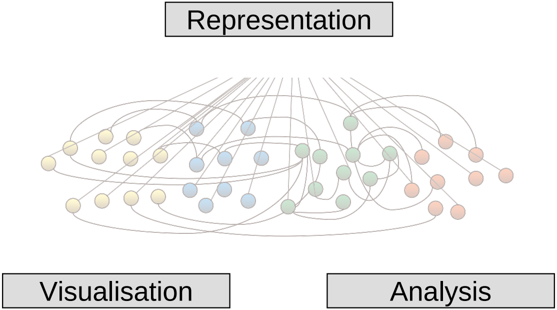
reactions development … evolution

Integrative Bioinformatics
Systems Biology
Network analysis

compartments tissues organs … environment

space

Bioinformatics Research Areas



Representation

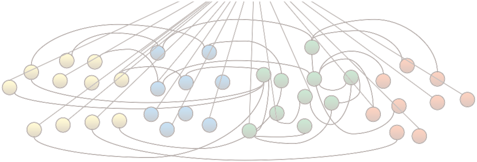
Visualisation

Analysis

Bioinformatics Research Areas

Data structures
Databases & information systems
Data integration

Representation



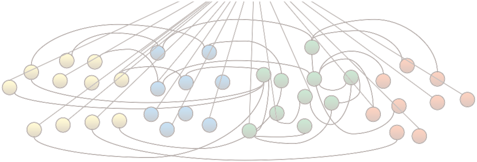
Visualisation

Analysis

Bioinformatics Research Areas

Data structures
Databases & information systems
Data integration

Representation



Visualisation

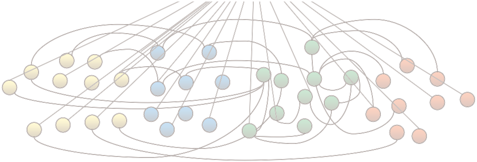
Analysis

Algorithms
Simulation
Theoretical CS
Statistics &
machine learning

Bioinformatics Research Areas

Data structures
Databases & information systems
Data integration

Representation



Visualisation

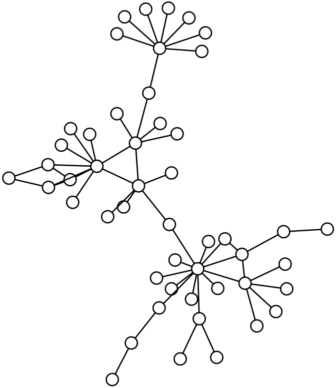
Analysis

Visualisation &
computer graphics
Human-computer-interaction

Algorithms
Simulation
Theoretical CS
Statistics &
machine learning

Biological Networks

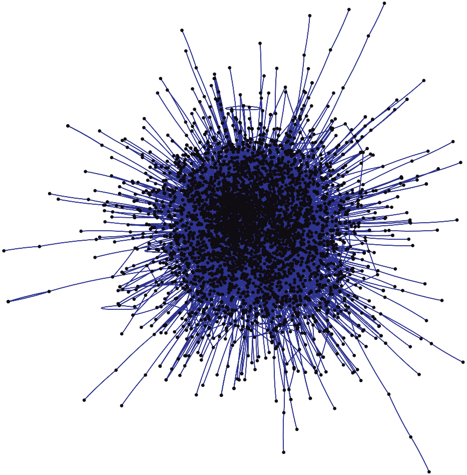
- Metabolic pathways
- Protein interaction networks
- Gene regulatory networks
- Signal transduction pathways
- Hormonal networks
- Food webs
- Evolutionary networks



Protein Interactions of *Mus musculus*
Source: DIP (Database of Interacting Proteins)

Biological Networks

- Networks get more complex
- Methods for the analysis of networks are required
- Several methods focus on *structural* analysis of networks



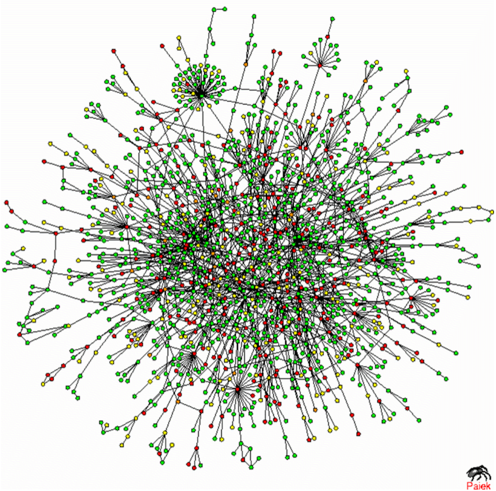
Protein Interactions of *Saccharomyces cerevisiae*
Source: DIP (Database of Interacting Proteins)

Example 1: Centralities in Biological Networks

Phenotypic effect of protein removal in *S. cerevisiae*: likelihood of lethal effects positively correlates with number of interactions

Effects:

- lethal (red)
- non-lethal (green)
- slow growth (orange)
- yellow (unknown)

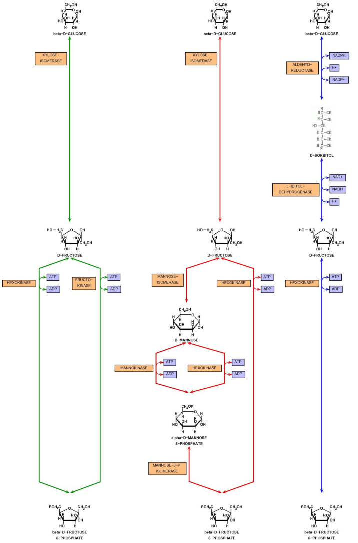


Lethality and centrality in protein networks [Jeong et al., 2001]

Example 2: Network Alignment and Comparison

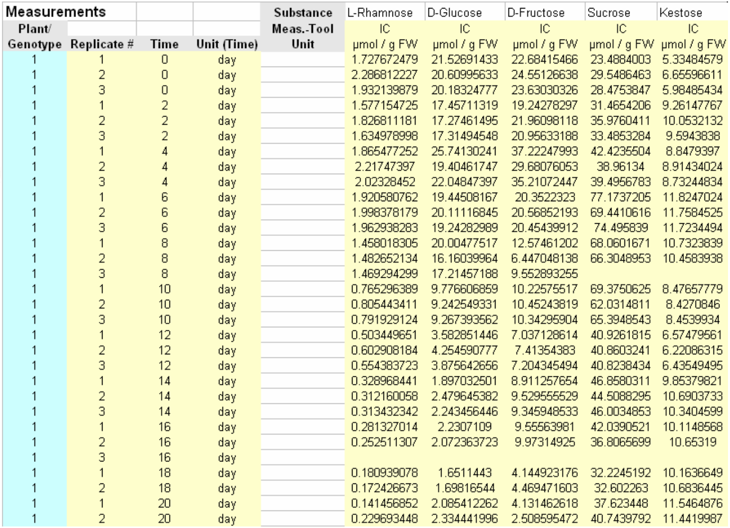
Applications for comparative analysis:

- Development of species-specific drug targets
- Identification of previously unknown parts of network in a species
- Understanding evolutional relationships between species



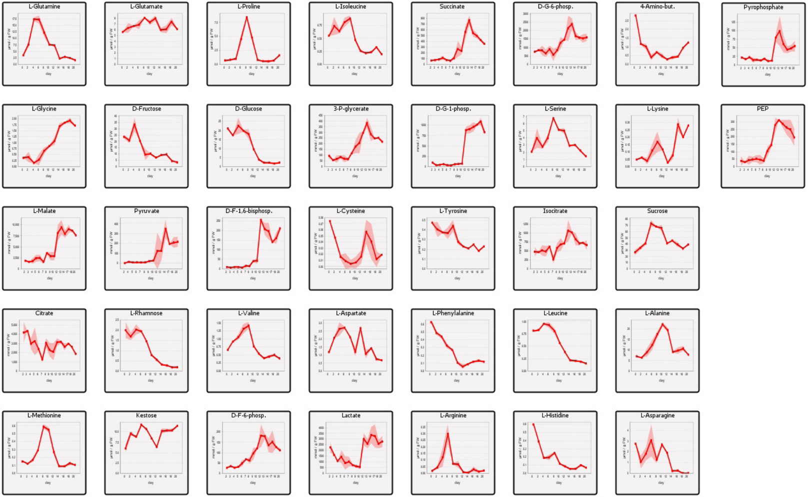
Visual comparison of metabolic pathways (BioPath)

Example 3: Network-related Data Analysis

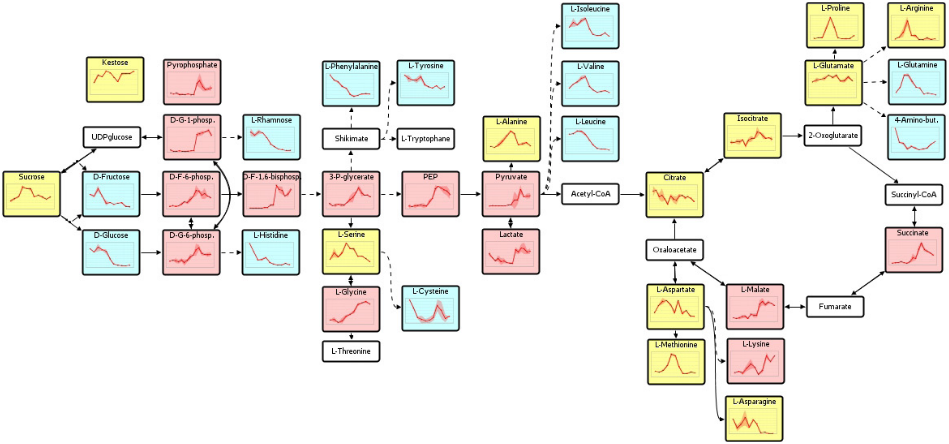


Measurements				Substance	L-Rhamnose	D-Glucose	D-Fructose	Sucrose	Kestose
Plant/				Meas.-Tool	IC	IC	IC	IC	IC
Genotype	Replicate #	Time	Unit (Time)	Unit	µmol / g FW	µmol / g FW	µmol / g FW	µmol / g FW	µmol / g FW
1	1	0	day		1.727672479	21.52691433	22.68415466	23.4884003	5.33484579
1	2	0	day		2.286812227	20.60995633	24.55126638	29.5486463	6.65596611
1	3	0	day		1.932139879	20.18324777	23.63030326	28.4753847	5.98485434
1	1	2	day		1.577154725	17.45711319	19.24278297	31.4654206	9.26147767
1	2	2	day		1.826811181	17.27461495	21.96098118	35.9760411	10.0532132
1	3	2	day		1.634978998	17.31494548	20.95633188	33.4853284	9.5943838
1	1	4	day		1.865477252	25.74130241	37.22247993	42.4235504	8.8479397
1	2	4	day		2.21747397	19.40461747	29.68076053	38.96134	8.91434024
1	3	4	day		2.02328452	22.04847397	35.21072447	39.4956783	8.73244834
1	1	6	day		1.920580762	19.44508167	20.3522323	77.1737205	11.8247024
1	2	6	day		1.998378179	20.11116845	20.56852193	69.4410616	11.7584525
1	3	6	day		1.962938283	19.24282989	20.45439912	74.495839	11.7234494
1	1	8	day		1.458018305	20.00477517	12.57461202	68.0601671	10.7323839
1	2	8	day		1.482652134	16.16039964	6.447048138	66.3048953	10.4583938
1	3	8	day		1.469294299	17.21457188	9.552893255		
1	1	10	day		0.765296389	9.776606859	10.22575517	69.3750625	8.47657779
1	2	10	day		0.805443411	9.242549331	10.45243819	62.0314811	8.4270846
1	3	10	day		0.791929124	9.267393562	10.34295904	65.3948543	8.4539934
1	1	12	day		0.503449651	3.582851446	7.037128614	40.9261815	6.57479561
1	2	12	day		0.602908184	4.254590777	7.41354383	40.8603241	6.22086315
1	3	12	day		0.554383723	3.875642656	7.204345494	40.8238434	6.43549495
1	1	14	day		0.328968441	1.897032501	8.911257654	46.8580311	9.85379821
1	2	14	day		0.312160058	2.479645382	9.529555529	44.5088295	10.6903733
1	3	14	day		0.313432342	2.243456446	9.345948533	46.0034853	10.3404599
1	1	16	day		0.281327014	2.2307109	9.55563981	42.0390521	10.1148568
1	2	16	day		0.252511307	2.072363723	9.97314925	36.8065699	10.65319
1	3	16	day						
1	1	18	day		0.180939078	1.6511443	4.144923176	32.2245192	10.1636649
1	2	18	day		0.172426673	1.69816544	4.469471603	32.602263	10.6836445
1	1	20	day		0.141456852	2.085412262	4.131462618	37.623448	11.5464876
1	2	20	day		0.229693448	2.334441996	2.508595472	40.7439792	11.4419987

Example 3: Data Analysis without the Network Context



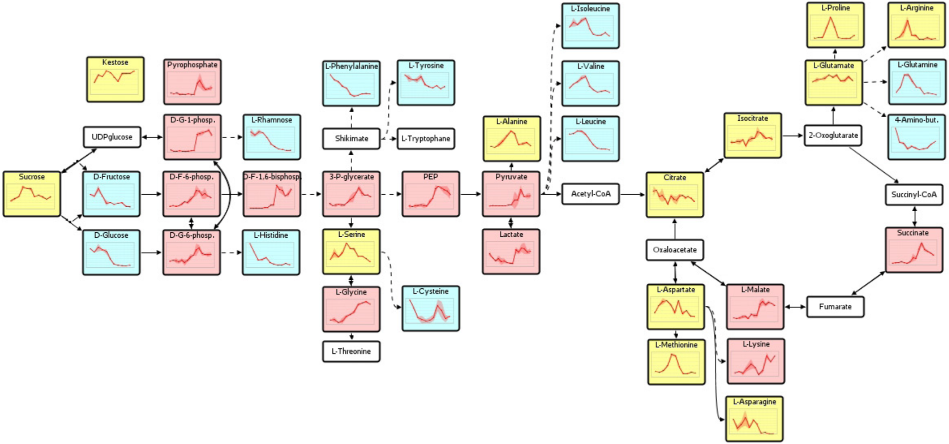
Example 3: Data Analysis within the Network Context



glycolysis TCA cycle

Clustering based on self-organising map (SOM)

Example 3: Data Analysis within the Network Context



glycolysis

TCA cycle

Three phases in seed development
(pre-storage, intermediate and main storage)

Analysis of Biological Networks

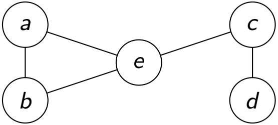
1. Motivation
2. Foundations
3. Network motifs
4. Network centralities
5. Network-related data analysis and visualisation

Graphs and Networks

- Network is an informal description for a set of elements with connections or interactions between them and data attached to them
- Graph is a formal description, it is a mathematical object consisting of vertices and edges representing elements and connections, respectively

Graphs

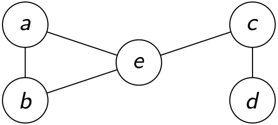
- *Graph:* $G = (V, E)$
- *Set of vertices:* V $(n = |V|)$
- *Set of edges:* $E \subseteq V \times V$ $(m = |E|)$
- *Neighbourhood of a vertex:* $N(u) = \{v : (u, v) \in E\}$
- *Adjacency matrix for* G*:* $(n \times n)$ matrix, where $a_{ij} = 1$ if and only if $(i, j) \in E$ and $a_{ij} = 0$ otherwise



$$A = \left(\begin{array}{cccccc} & a & b & c & d & e \\ a & 0 & 1 & 0 & 0 & 1 \\ b & 1 & 0 & 0 & 0 & 1 \\ c & 0 & 0 & 0 & 1 & 1 \\ d & 0 & 0 & 1 & 0 & 0 \\ e & 1 & 1 & 1 & 0 & 0 \end{array}\right)$$

Graphs

- *Degree of a vertex:* number of its incident edges ($d(v)$)
- *Walk:* sequence of edges connecting vertices $(e_1, \ldots, e_k)$
- *Length of a walk:* number of edges, $k = |(e_1, \ldots, e_k)|$
- *Path:* walk where edges are pairwise distinct
- *Shortest path:* a minimal length path between u and v

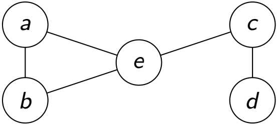


$d(a) = 2$
Walk:
$((a,b),(b,a),(a,b),(b,e),(e,c))$
Path: $((a,b),(b,e),(e,c))$
Shortest path: $((a,e),(e,c))$

Graphs

- *Distance:* length of a shortest path between two vertices ($\text{dist}(u, v)$)
- *Connected graph:* a walk exists between every two vertices
- *Random walk:* starting at vertex u chooses uniformly at random an incident edge until reaching v
- *Subgraph:* of $G = (V, E)$ is a graph $G' = (V', E')$, where $V' \subseteq V$, and $E' \subseteq E \cap V' \times V'$

Consider non-trivial, loop-free, connected graphs



$\text{dist}(a, c) = 2$

Random walk from a to c:

$((a, e), (e, a), (a, b), (b, e), (e, c))$

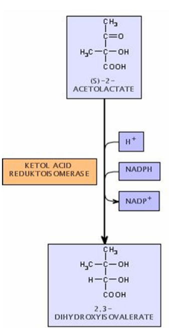
Subgraph: $V' = \{a, b, e\}$ and $E' = \{(a, b), (a, e)\}$

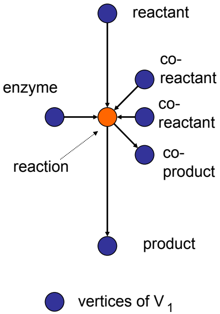
Graphs - Types

- Undirected graphs
- Directed graphs
- Mixed graphs
- Labelled graphs (vertices, edges)
- Multi-graphs

Graphs - Types

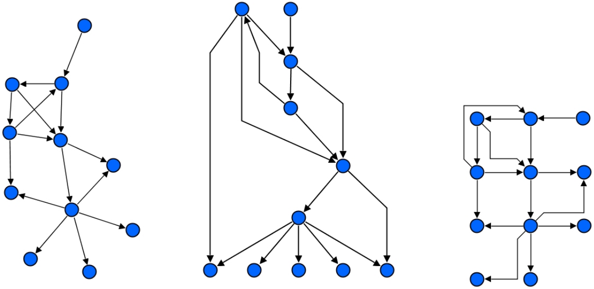
- Hyper-graphs
  consists of a set of vertices and a set of hyper-edges, each hyper-edge is a non-empty subsets of the node set V
- Bipartite graphs
  vertex set V can be partitioned in two disjoint, nonempty sets V_1 and V_2 such that each edge in E has exactly one end-vertex in V_1 and one end-vertex in V_2





Graph visualisation

- Graphical representation of a graph
- Draw a point for each vertex and a line for each edge which connects the corresponding points of its end-vertices
- The positions of the vertices and the drawing of the lines is called the layout of the graph



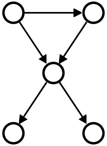
Analysis of Biological Networks

1. Motivation
2. Foundations
3. Network motifs
4. Network centralities
5. Network-related data analysis and visualisation

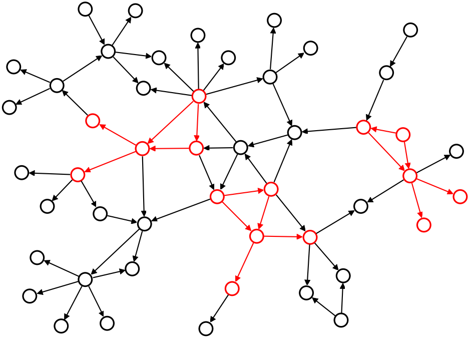
Motivation

Motif detection in networks:

- searching
- counting
- visual exploration



Motif
size: $|E_p|$



Target graph $G_t = (V_t, E_t)$ with highlighted motif matches
size: $|E_t|$

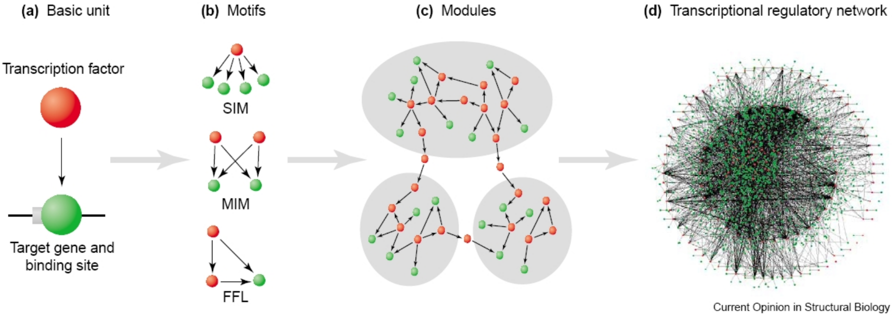
Motivation

Interesting motifs are found in biological networks:

- Gene regulatory networks
- Metabolic networks
- Protein-protein interaction networks
- Neuronal networks, food-webs

Network Motifs

- Particular subgraphs representing patterns of local interconnections between network elements
- May represent basic building blocks and design patterns of functional modules
- Overabundance may be a consequence of positive selection due to functional or structural properties

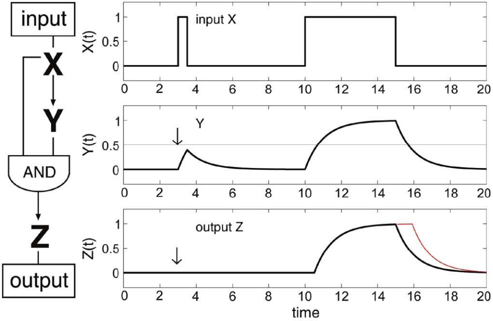

(a) Basic unit
(b) Motifs
(c) Modules
(d) Transcriptional regulatory network
Transcription factor
Target gene and binding site
SIM
MIM
FFL
Current Opinion in Structural Biology

[Babu *et al.*, 2004]

Motifs in Gene Regulatory Networks: Feed-forward Loop

Functional properties of the feed-forward loop motif in gene regulation

- Noise filtering: responds only to persistent activations



Network motifs in the transcriptional regulation network of *Escherichia coli* [Shen-Orr *et al.*, 2002]

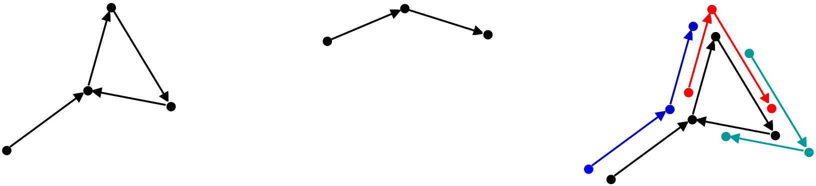
Analysing Networks and Finding Interesting Motifs

Frequent motifs in networks

- *Motif frequency*: number of matches in the target graph
- Motifs with high frequency are potential candidates for functional network motifs
- Different concepts for frequency determination as a result of different restrictions of the reuse of graph elements
  → *Motif Frequency Concepts*

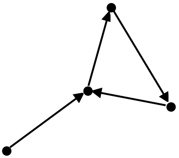
Concepts for Determination of Motif Frequency

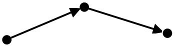
Concept	Graph element reuse		Frequency determination
	Vertices	Edges	
$\mathcal{F}_1$	yes	yes	All matches

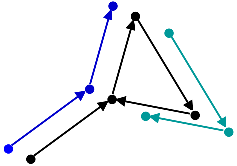


Concepts for Determination of Motif Frequency

Concept	Graph element reuse		Frequency determination
	Vertices	Edges	
$\mathcal{F}_1$	yes	yes	All matches
$\mathcal{F}_2$	yes	no	Maximum independent set

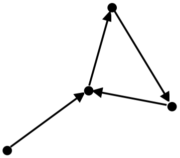


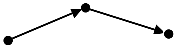




Concepts for Determination of Motif Frequency

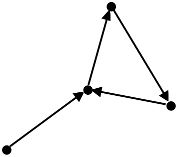
Concept	Graph element reuse		Frequency determination
	Vertices	Edges	
$\mathcal{F}_1$	yes	yes	All matches
$\mathcal{F}_2$	yes	no	Maximum independent set
-	no	yes	-

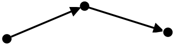


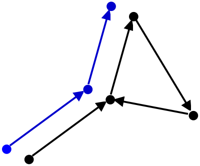


Concepts for Determination of Motif Frequency

Concept	Graph element reuse		Frequency determination
	Vertices	Edges	
$\mathcal{F}_1$	yes	yes	All matches
$\mathcal{F}_2$	yes	no	Maximum independent set
-	no	yes	-
$\mathcal{F}_3$	no	no	Maximum independent set







Properties of the Frequency Concepts

$\mathcal{F}_1$

- Does not exclude matches
- Shows the full potential of the motif

$\mathcal{F}_2$

- Matches does not share relation of elements
- Shows the maximum number of instances of a particular motif which can be active at the same time

$\mathcal{F}_3$

- Matches can be seen as non-overlapping clusters
- Allows specific analysis and navigation methods
  - Folding and unfolding of clusters
  - Motif preserving layout of the matches

Frequent Motif Search - Problems

- Many different motifs:

Rev. edges	-	+	-	+
Self loops	-	-	+	+
Motif size				
2	3	4	5	6
3	10	12	18	21
4	39	53	76	97
5	169	237	361	478
6	876	1306	1978	2762
7	4834	7537	11658	17002
8	29316	47913	74494	113528
9	189054	322253	505277	801966

- Many different matches for one motif: $O(|E_t|^{|E_p|})$

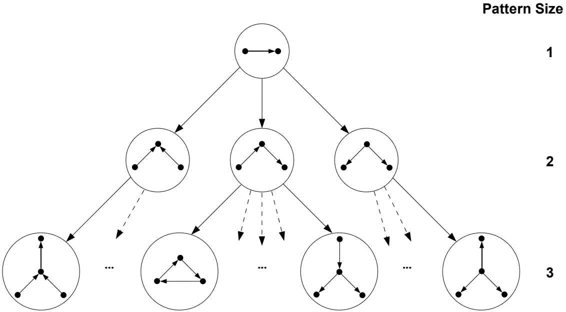
There is an giant number of possible motif matches!

Frequent Motif Finding Algorithm

- Search motifs of given size with maximum frequency:
  - Given: graph $G = (V, E)$, target size t, frequency concept $\mathcal{F}$
  - Result: Motif with maximum frequency (and frequency)
- Several extensions:
  - Application of different frequency concepts
  - Full control over the search, e.g. define frequency threshold
  - Parallel implementation of the search algorithm
- Idea:
  - Start of search with the motif of size 1
    $\longrightarrow$ each edge of target graph used to create a match
  - While (motifs for extension are left)
    $\longrightarrow$ extend next motif by addition of one edge (combine match with each incident edge to new motif)
    $\longrightarrow$ compute frequency for each new motif and if frequency is above threshold than keep motif (adjust threshold)

Traversal of the Space of all Motifs

- Each motif is assigned to parent motif of size $n-1$
- Only generation of the motifs supported by the target graph
- Depth first traversal of the motif tree
  - Allows pruning of infrequent branches for $\mathcal{F}_2$ and $\mathcal{F}_3$



Problems, Complexity and Solutions

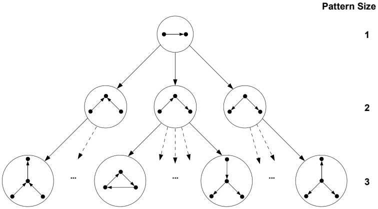
Problem

- Many different motifs

Rev. edges	-	+	...
Self loops	-	-	...
Motif size			
2	3	4	...
3	10	12	...
4	39	53	...
5	169	237	...
6	876	1306	...
7	4834	7537	...
8	29316	47913	...
9	189054	322253	...

Solution

- Only consider motifs supported by target graph
- Building and pruning of motif tree



- Parallel computation

Problems, Complexity and Solutions

Problem

- Many different matches for one motif: $O(|E_t|^{|E_p|})$
- Maximum independent set
- Graph isomorphism

Solution

- Only extension of matches of parent motif
- Use of heuristic
- Canonical labelling

Problems, Complexity and Solutions

Problem

- Many different matches for one motif: $O(|E_t|^{|E_p|})$
- Maximum independent set
- Graph isomorphism
- In worst case computational very expensive

Solution

- Only extension of matches of parent motif
- Use of heuristic
- Canonical labelling
- In practise for moderate sized networks applicable

Analysis of Biological Networks

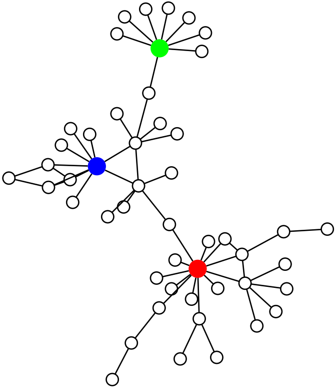
1. Motivation
2. Foundations
3. Network motifs
4. Network centralities
5. Network-related data analysis and visualisation

Network Centralities

Ranking of vertices according to importance based on the network structure

Applications of centralities:

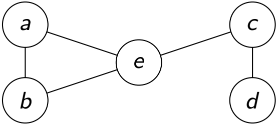
- Prioritisation of potential drug targets
- Hypothesis generation for experiments
- Exploration of a network
- Determination which patients should be vaccinated first



Protein Interactions of *Mus musculus*
Source: DIP (Database of Interacting Proteins)

Informal Definition of Centrality

- Let $G = (V, E)$ be a graph
- A function $\mathcal{C}\colon V \mapsto \mathbb{R}$ is called a *centrality*
- We say $u \in V$ is more important than $v \in V$ with respect to a given centrality $\mathcal{C}$ if $\mathcal{C}(u) > \mathcal{C}(v)$
- A centrality allows us to order the vertices
- Convention: “Important” vertices get a high centrality value



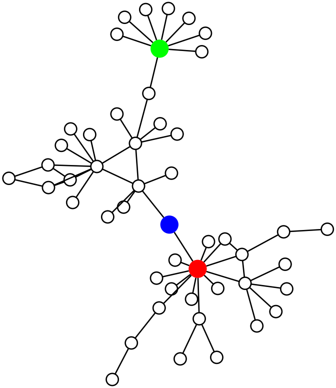
v	$\mathcal{C}_d(v)$	Order
a	2	e
b	2	a
c	2	b
d	1	c
e	3	d

Centralities in Biological Networks

- Ranking of vertices according to importance
- Based on the network structure
- Many different (> 20) centrality measures exist
- Examples: Degree-Centrality, Eccentricity-Centrality, Closeness-Centrality, Random Walk Betweenness-Centrality, Eigenvector-Centrality

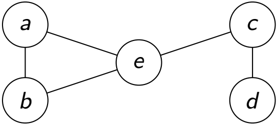
	Degree	Closeness	SPBetw
1	DIP:320N	DIP:369N	DIP:320N
2	DIP:24169N	DIP:1048N	DIP:369N
3	DIP:493N	DIP:320N	DIP:1048N
4	DIP:24196N	DIP:24196N	DIP:24196N
5	DIP:442N	DIP:24169N	DIP:24169N

DIP:320N	protein-tyrosine kinase JAK2
DIP:493N	transcription factor IID chain
DIP:1048N	protein kinase raf-1



Degree-Centrality

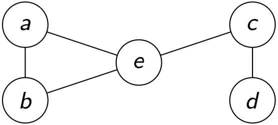
- Number of incident edges to v
- *degree-centrality:* $\mathcal{C}_d(v) := d(v)$
- Jeong *et al.* reports the correlation of removal of high degree proteins with lethality for the organism (Jeong *et al.*, 2002)



v	$\mathcal{C}_d(v)$
a	2
b	2
c	2
d	1
e	3

Eccentricity-Centrality

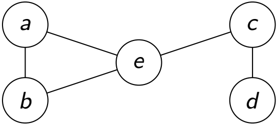
- *eccentricity ecc* of a vertex u is defined as $\text{ecc}(u) := \max_{v \in V} \text{dist}(u, v)$
- *eccentricity-centrality:* $\mathcal{C}_e(u) := \frac{1}{\text{ecc}(u)}$
- Applied by Wuchty *et al.* to compute the "central" metabolites of the metabolic network (Wuchty *et al.*, 2003)



v	$\mathcal{C}_e(v)$
a	0.333
b	0.333
c	0.5
d	0.333
e	0.5

Closeness-Centrality

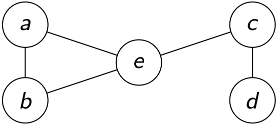
- Use sum of distances to all other vertices, i.e. $\mathrm{sumdist}(u) = \sum_{v \in V} \mathrm{dist}(u, v)$
- *closeness-centrality:* $\mathcal{C}_c(u) := \frac{1}{\mathrm{sumdist}(u)}$
- Also applied by Wuchty *et al.* to compute the central metabolites of the metabolic network (Wuchty *et al.*, 2003)



v	$\mathcal{C}_c(v)$
a	0.143
b	0.143
c	0.167
d	0.111
e	0.2

Random Walk Betweenness-Centrality

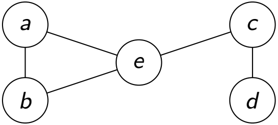
- Betweenness: Observe communication in the network
  - A vertex u can observe the communication between v and w if u lies in the path of the communication
  - Different methods to model communication
- *random-walk betweenness centrality* $\mathcal{C}_r(u)$ is equal to the number of times a random walk from v to w goes through u, averaged over all v and w



v	$\mathcal{C}_r(v)$
a	0.5
b	0.5
c	0.7
d	0.4
e	0.833

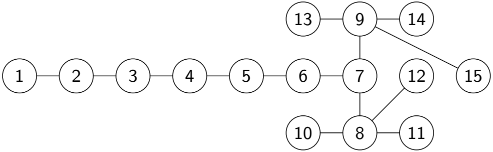
Eigenvector-Centrality

- Value of a single vertex is determined by the values of the neighbouring vertices
- *eigenvector-centrality:* $\mathcal{C}_\lambda(u) := \sum_{v \in N(u)} \mathcal{C}_\lambda(v)$
  - Equivalent to: $\mathcal{C}_\lambda(v_i) := \sum_{j=1}^{n} a_{ij} \mathcal{C}_\lambda(v_j)$
  - Well known problem of eigenvector computation $\lambda S = AS$
  - We use the eigenvector for the largest eigenvalue



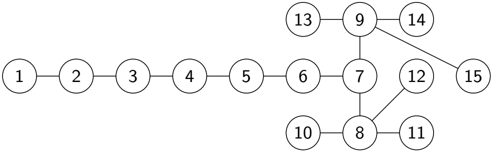
v	$\mathcal{C}_\lambda(v)$
a	0.497
b	0.497
c	0.342
d	0.155
e	0.604

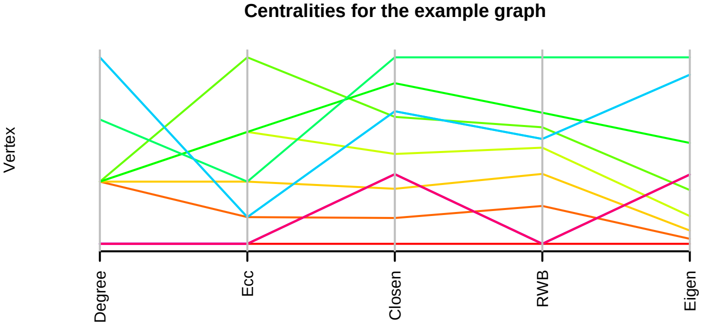
Comparing Centrality Values



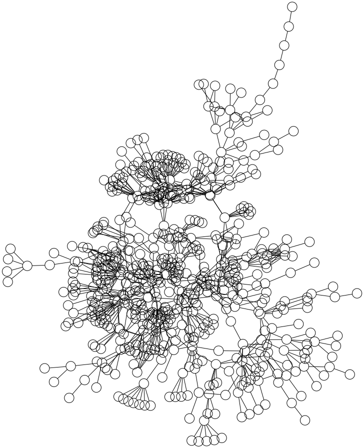
Vertex	$\mathcal{C}_d$	Vertex	$\mathcal{C}_e$	Vertex	$\mathcal{C}_c$	Vertex	$\mathcal{C}_r$	Vertex	$\mathcal{C}_\lambda$
8	4	5	0.2500	7	0.0286	7	0.7429	7	0.5021
9	4	4	0.2000	6	0.0263	6	0.5619	8	0.4563
7	3	6	0.2000	8	0.0238	5	0.5143	9	0.4563
2	2	3	0.1667	9	0.0238	8	0.4762	6	0.2761
3	2	7	0.1667	5	0.0233	9	0.4762	10	0.1927
4	2	2	0.1429	4	0.0200	4	0.4476	11	0.1927
5	2	8	0.1429	10	0.0182	3	0.3619	12	0.1927
6	2	9	0.1429	11	0.0182	2	0.2571	13	0.1927
1	1	1	0.1250	12	0.0182	1	0.1333	14	0.1927
10	1	10	0.1250	13	0.0182	10	0.1333	15	0.1927
11	1	11	0.1250	14	0.0182	11	0.1333	5	0.1517
12	1	12	0.1250	15	0.0182	12	0.1333	4	0.0830
13	1	13	0.1250	3	0.0169	13	0.1333	3	0.0448
14	1	14	0.1250	2	0.0143	14	0.1333	2	0.0230
15	1	15	0.1250	1	0.0120	15	0.1333	1	0.0097

Comparing Centrality Values



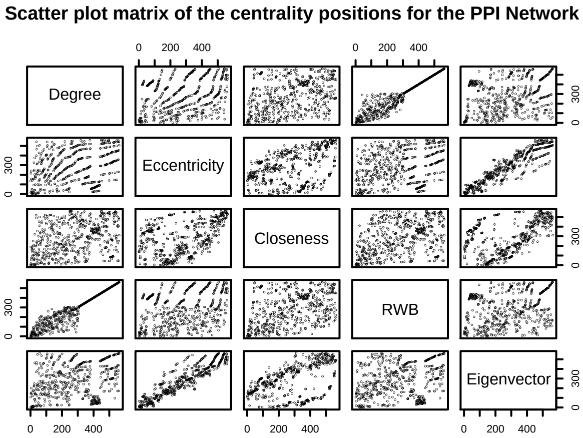


Protein Protein Interactions for *Homo sapiens*



- Vertices denote proteins
- Edges denote interactions
- Undirected network
- Vertex and edge labels not shown
- Only the giant component, 563 vertices, 870 edges

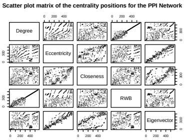
Protein Protein Interactions for *Homo sapiens*

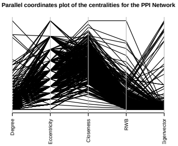


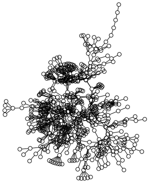
Protein Protein Interactions for *Homo sapiens*

Correlation coefficients for the centrality positions for the PPI-network

	$\mathcal{C}_d$ Degree	$\mathcal{C}_e$ Eccentricity	$\mathcal{C}_c$ Closeness	$\mathcal{C}_r$ RWB	$\mathcal{C}_\lambda$ Eigenvector
$\mathcal{C}_d$	–	0.2794	0.3396	0.9534	0.2703
$\mathcal{C}_e$	0.2794	–	0.4231	0.2776	0.9248
$\mathcal{C}_c$	0.3396	0.4231	–	0.3843	0.4726
$\mathcal{C}_r$	0.9534	0.2776	0.3843	–	0.2627
$\mathcal{C}_\lambda$	0.2703	0.9248	0.4726	0.2627	–


Scatter plot matrix of the centrality positions for the PPI Network


Parallel coordinates plot of the centralities for the PPI Network

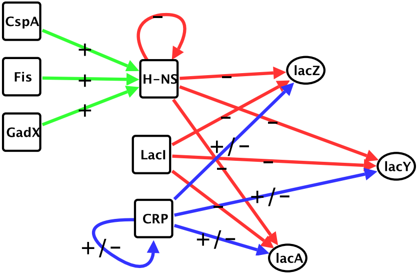


Centralities in Biological Networks

- Ranking of vertices according to importance
- Based on the network structure
- Many different (> 20) centrality measures exist
- None of them uses biological relevant information
- New or adapted centralities necessary

Gene Regulatory Networks

- Genes are either transcription factors or targets genes
- Transcription factors regulate genes
- Regulatory interactions between genes form a network



Regulation of lactose degradation in *E. coli*
Source: RegulonDB

Global Regulators in Gene Regulatory Networks

Understanding gene regulation ⇔ identification of global regulators

Different criteria for identification of global regulators:

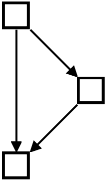
- Number of regulated genes
- Number and type of co-regulators
- Number of other regulators they control
- Size of their evolutionary family
- Number of growth conditions under which they are active

Centralities for Gene Regulatory Networks

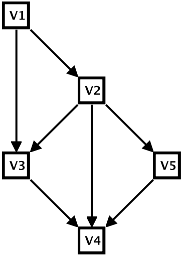
- Specific centrality for the analysis of GRN necessary
- Should help in identifying global regulators
- Network motifs are relevant for GRNs!

Motif-based Centrality

- Combination of centrality measures and network motifs
- Use the occurrences of a motif in the network
- Incorporation of functional substructures into centrality analysis



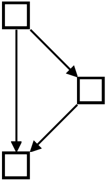
Feed-forward loop (FFL)



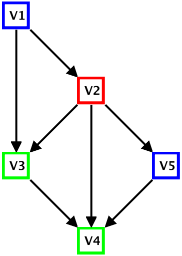
Target graph

Motif-based Centrality

- Combination of centrality measures and network motifs
- Use the occurrences of a motif in the network
- Incorporation of functional substructures into centrality analysis



Feed-forward loop (FFL)



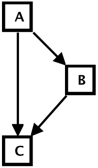
Target graph

Vertex	Centrality
V_2	3
V_3	2
V_4	2
V_1	1
V_5	1

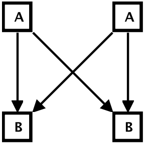
Centr. for FFL motif

Roles of Motif Vertices

Different vertices have different roles



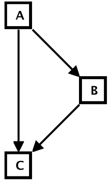
Feed-forward loop



Bi-fan

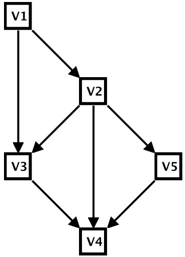
Extended Motif-based Centrality

Count the number of matches according to roles



Feed-forward loop

Vertex	Centrality value		
	Role A	Role B	Role C
v_2	2	1	0
v_1	1	0	0
v_3	0	1	1
v_5	0	1	0
v_4	0	0	2



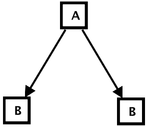
Target graph

Motif Classes

Motifs with similar structure grouped into motif classes

Example: Single Input Motif (SIM)

Role A	regulator for a set of genes
Role B	exclusively regulated genes

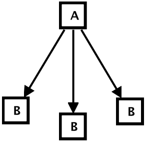


Motif Classes

Motifs with similar structure grouped into motif classes

Example: Single Input Motif (SIM)

Role A	regulator for a set of genes
Role B	exclusively regulated genes

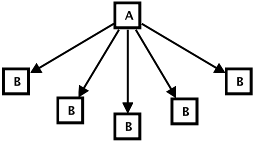


Motif Classes

Motifs with similar structure grouped into motif classes

Example: Single Input Motif (SIM)

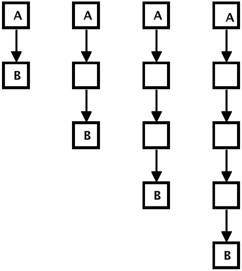
Role A	regulator for a set of genes
Role B	exclusively regulated genes



Regulatory Chains

- One regulator regulating another, which in turn regulates a third one and so forth
- Regulators at the top of chains start regulatory cascades

Role A	regulator starting a regulatory cascade
Role B	target gene of regulatory cascade
Other	intermediate regulators



Summary: Varieties of Motif-based Centralities

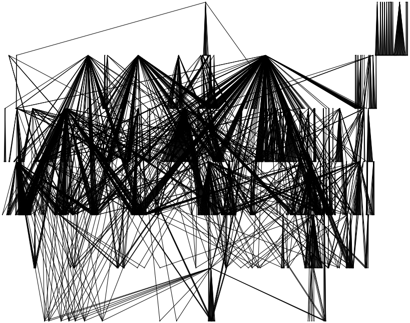
Three varieties of motif-based centralities

- Plain
- With roles for the vertices of the motif
- Using classes of similar motifs

All based on the same concept:
count the matches of functional substructures in the target graph

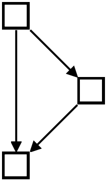
Gene Regulatory Network of *E. coli*

- Based on data from RegulonDB
- 1250 vertices and 2515 edges
- Global regulators?



Motif-based Centrality for *E. coli*

Gene	Cent.
crp	254
fnr	203
arcA	111
fis	110
narL	100
ihfAB	61
hns	53
fur	43
gadX	34
hyfR	33
marA	29
flhD	21
nagC, *soxS*	19
modE, *tdcA*, *yiaJ*	18
gutM, *ompR*, *srlR*	17

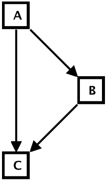


Top 20: 10 of the 18 global regulators

Martínez-Antonio & Collado-Vides: Identifying global regulators in transcriptional regulatory networks in bacteria, Current Opinion in Microbiology, 2003

Extended Motif-based Centrality for *E. coli*

	Cent.		
Gene	*A*	*B*	*C*
crp	254	0	0
fnr	150	53	0
ihfAB	61	0	0
arcA	58	53	0
fis	40	70	0
modE	18	0	0
soxS	18	1	0
hns	14	39	0
fhlA	11	0	0
gadE	11	0	0
cpxR	11	0	0
rob	10	0	0
galR	8	0	0
gadX	8	26	0
gntR	6	0	0
fur	6	36	1
oxyR	6	1	0
tdcR	6	0	0
narL	5	95	0
nagC	5	14	0



Top 20: 11 of the 18 global regulators

Motif-class Centrality for *E. coli*

Gene	c_{mcc}	Length of chain					
		2	3	4	5	6	7
crp	1592	359	525	436	212	60	0
ihfAB	667	186	215	156	82	28	0
fnr	470	206	237	27	0	0	0
arcA	470	111	215	127	17	0	0
fis	387	156	121	82	28	0	0
evgA	325	4	27	90	125	51	28
ydeO	322	1	27	90	125	51	28
gadE	321	27	90	125	51	28	0
soxR	213	2	24	92	91	4	0
soxS	211	24	92	91	4	0	0
torR	191	10	15	87	51	28	0
gadW	185	4	15	87	51	28	0
cspE	184	1	2	88	65	28	0
cspA	183	2	88	65	28	0	0
gadX	181	15	87	51	28	0	0
hns	181	88	65	28	0	0	0
oxyR	166	15	73	74	4	0	0
fur	151	73	74	4	0	0	0
modE	141	32	94	15	0	0	0
narL	109	94	15	0	0	0	0

Top 20: 11 of the 18

Other effects:
evgA, *ydeO* and *gadE*

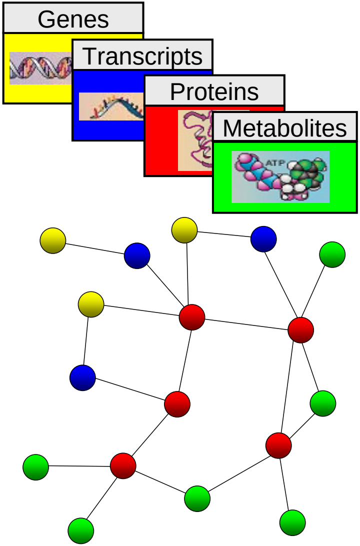
Analysis of Biological Networks

1. Motivation
2. Foundations
3. Network motifs
4. Network centralities
5. Network-related data analysis and visualisation

Analysis in Network Context - Motivation

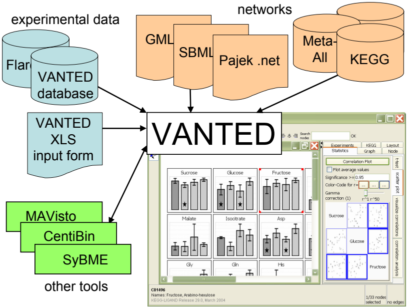
Integration and analysis of high-throughput data in the context of underlying processes

- Show large amounts of data in a readable and understandable form
- Consideration related networks
- Fast data evaluation (statistic tests, correlation analysis, data clustering)



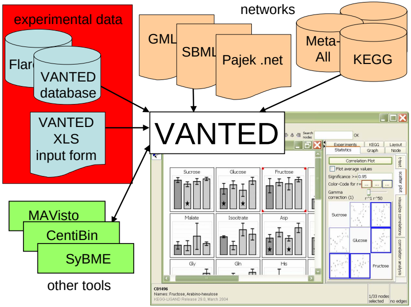
Analysis and Exploration of Data in Network Context

- Vanted features
  - Creation/derivation of networks
  - Data mapping onto dynamic networks
  - Data analysis and visualisation methods



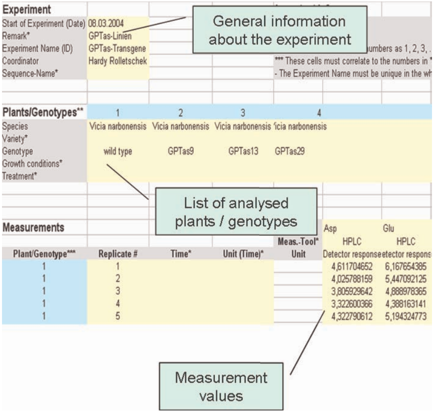
Experimental Data

- Databases
    - DBE
    - FLAREX



Experimental Data

- Databases
  - DBE
  - FLAREX
- Excel/CSV Files
  - `Vanted` input file (metabolite, proteomics and expression data)
  - J-Express files (expression data)



Experiment	
Start of Experiment (Date)	08.03.2004
Remark*	GPTas-Linien
Experiment Name (ID)	GPTas-Transgene
Coordinator	Hardy Rolletschek
Sequence-Name*	

umbers as 1, 2, 3, .

*** These cells must correlate to the numbers in *

- The Experiment Name must be unique in the wh

General information about the experiment

Plants/Genotypes**	1	2	3	4
Species	Vicia narbonensis	Vicia narbonensis	Vicia narbonensis	ʻicia narbonensis
Variety*				
Genotype	wild type	GPTas9	GPTas13	GPTas29
Growth conditions*				
Treatment*				

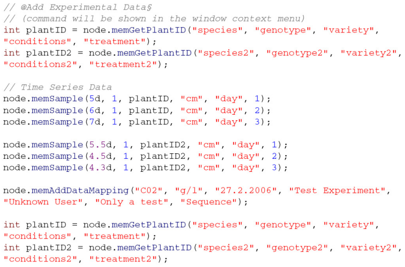
List of analysed plants / genotypes

Measurements					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

Measurement values

Experimental Data

- Databases
  - DBE
  - FLAREX
- Excel/CSV Files
  - Vanted input file (metabolite, proteomics and expression data)
  - J-Express files (expression data)
- Programmatic
  - Script API: Java/Ruby code

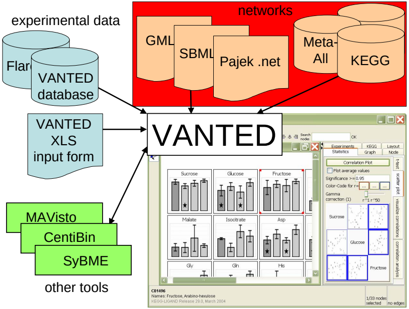


```
// @Add Experimental Data§
// (command will be shown in the window context menu)
int plantID = node.memGetPlantID("species", "genotype", "variety",
"conditions", "treatment");
int plantID2 = node.memGetPlantID("species2", "genotype2", "variety2",
"conditions2", "treatment2");

// Time Series Data
node.memSample(5d, 1, plantID, "cm", "day", 1);
node.memSample(6d, 1, plantID, "cm", "day", 2);
node.memSample(7d, 1, plantID, "cm", "day", 3);

node.memSample(5.5d, 1, plantID2, "cm", "day", 1);
node.memSample(4.5d, 1, plantID2, "cm", "day", 2);
node.memSample(4.3d, 1, plantID2, "cm", "day", 3);

node.memAddDataMapping("C02", "g/l", "27.2.2006", "Test Experiment",
"Unknown User", "Only a test", "Sequence");

int plantID = node.memGetPlantID("species", "genotype", "variety",
"conditions", "treatment");
int plantID2 = node.memGetPlantID("species2", "genotype2", "variety2",
"conditions2", "treatment2");
```

Network Data

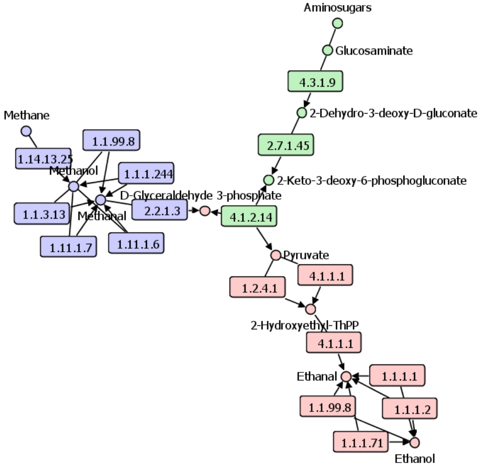
- Databases
  - MetaCrop (SBML)
  - KEGG Pathway

    - Reference-Pathway/ Organism-specific
    - Bottom-Up
    - Top-Down
    - Super-Pathway

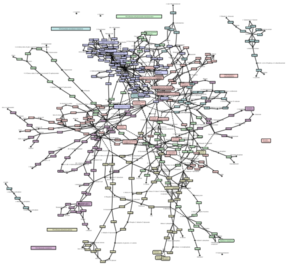


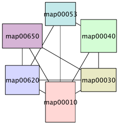
Network Data - KEGG Specific Pathway Parts

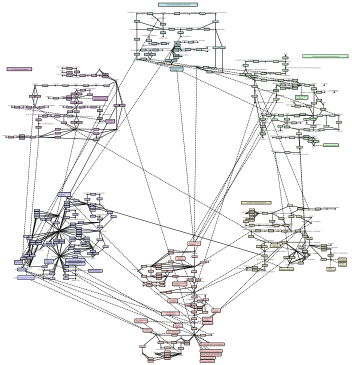
Example: Shortest paths between the substances methane, amino-sugars and ethanol



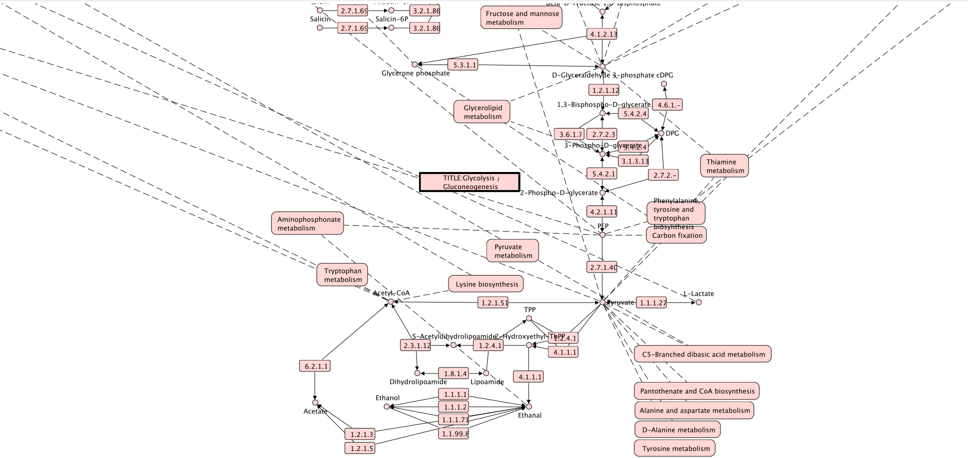
Network Data - Exploration of KEGG Pathways





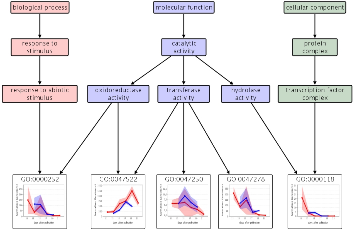


Network Data - Exploration of KEGG Pathways



Network Data

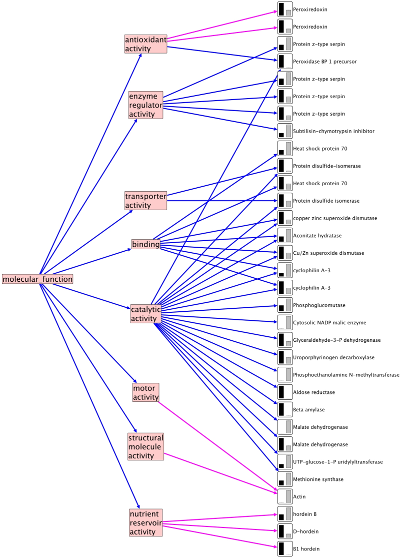
- Databases
  - MetaCrop (SBML)
  - KEGG Pathway
    - Reference-Pathway/ Organism-specific
    - Bottom-Up
    - Top-Down
    - Super-Pathway
- Ontologies / functional hierarchies
  - Full GO tree
  - Relevant subset
- Files
  - GML, Pajek-.NET, SBML



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

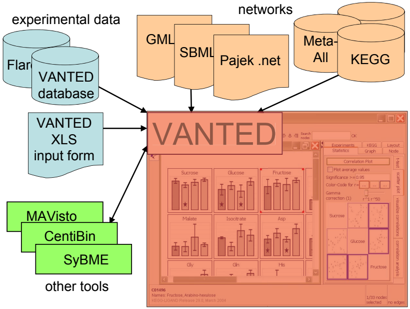
Network Data

- Databases
  - MetaCrop (SBML)
  - KEGG Pathway
    - Reference-Pathway/ Organism-specific
    - Bottom-Up
    - Top-Down
    - Super-Pathway
- Ontologies / functional hierarchies
  - Full GO tree
  - Relevant subset
- Files
  - GML, Pajek-.NET, SBML

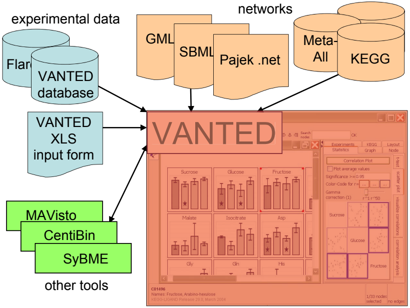


Data Integration

- Knowledge about identifiers, synonyms, associated genes and annotations
  - KEGG compounds
  - Expasy Enzymes
  - KEGG KO
  - User defined (e. g. Affymetrix)



Data Analysis



experimental data

Flar

VANTED database

VANTED XLS input form

MAVisto

CentiBin

SyBME

other tools

networks

GML

SBML

Pajek .net

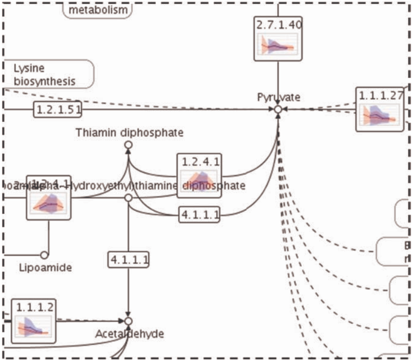
Meta-All

KEGG

VANTED

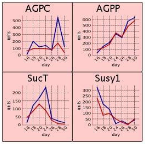
Data Analysis

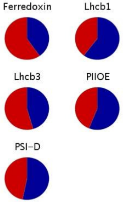
- Visualisation methods

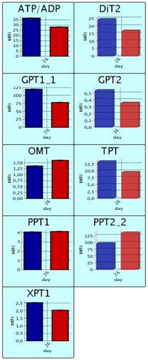


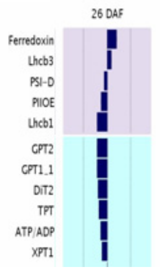
Data Analysis

- Visualisation methods
- Data Charting
  - Data mapping, graph-embedded view of experimental data
  - Time series data (line chart)
  - Non-time series data (bar charts, pie charts, ratio view)
  - Filter operations (show/analyze subset, e.g. selected plant lines, selected time point(s))



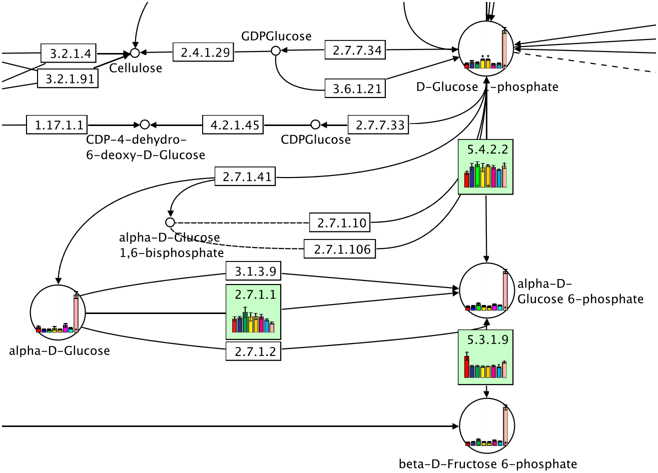






Data Analysis

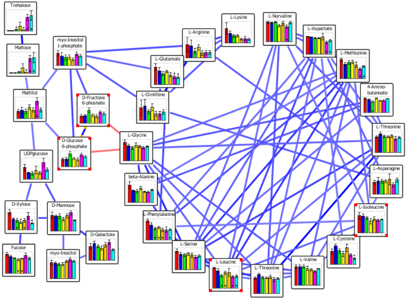
- Combination of different -omics data



Compound / Enzyme information, mapped onto a KEGG pathway

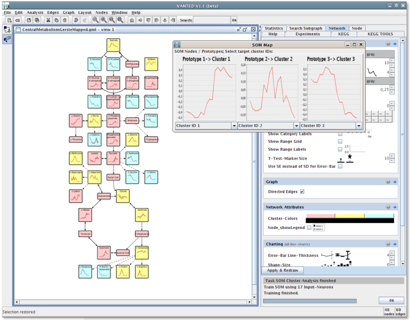
Data Analysis

- Statistical analysis
  (e.g. outliers removal, t-test)
- Correlation analysis
  - Correlate time-series profile
  - Correlate samples (replicate data)

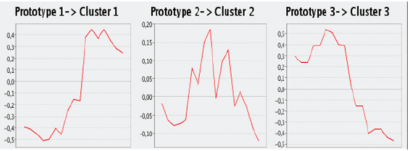


Data Analysis

- Data clustering (e.g. SOM)
  - Detection of a given number of common time series patterns in the data
- Connection to external data clustering approaches

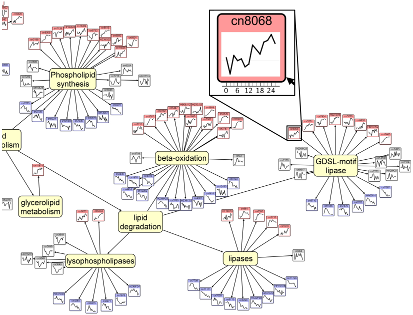


Task SOM Cluster Analysis finished
Train SOM using 17 Input-Neurons
Training finished.



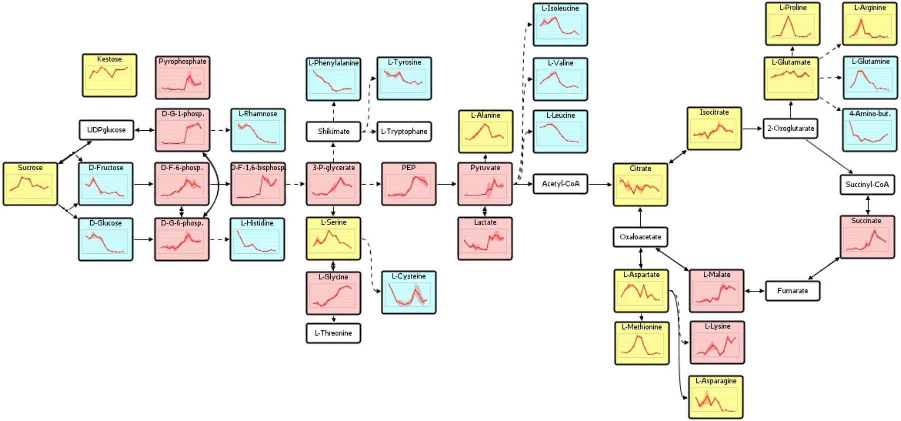
Data Analysis

- Data clustering (e.g. SOM)
  - Detection of a given number of common time series patterns in the data
- Connection to external data clustering approaches



Expression data map onto MapMan hierarchy

Examples - Data Analysis (Barley)

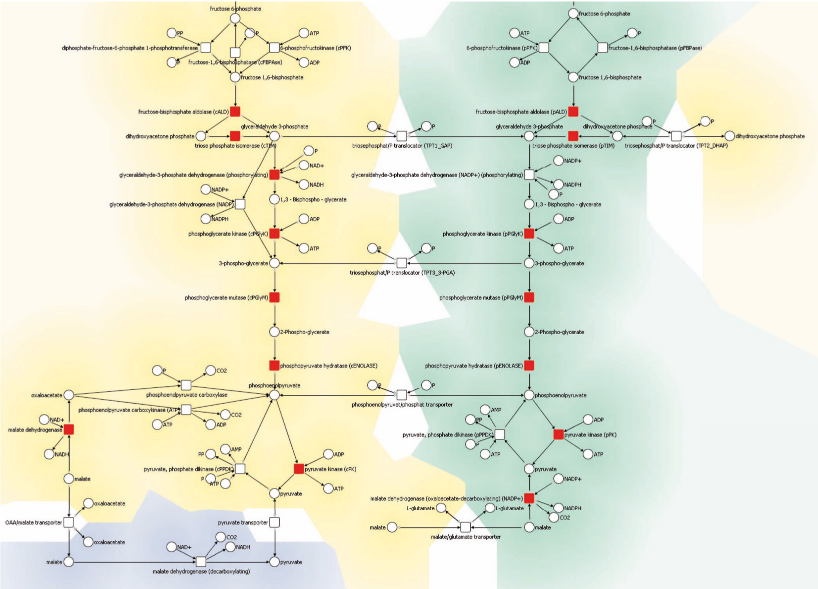


glycolysis

TCA cycle

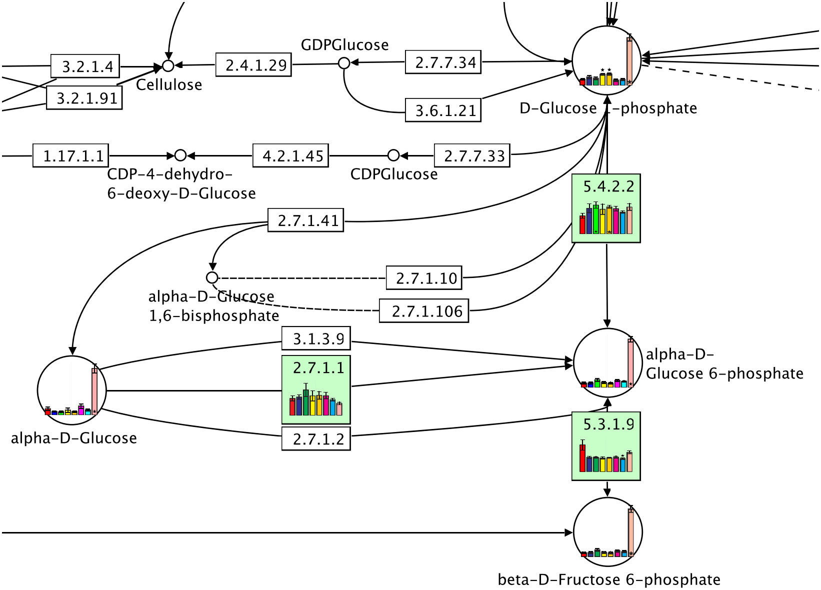
Three phases in seed development
(pre-storage, intermediate and main storage)

Examples - Coverage of Processes (Barley)

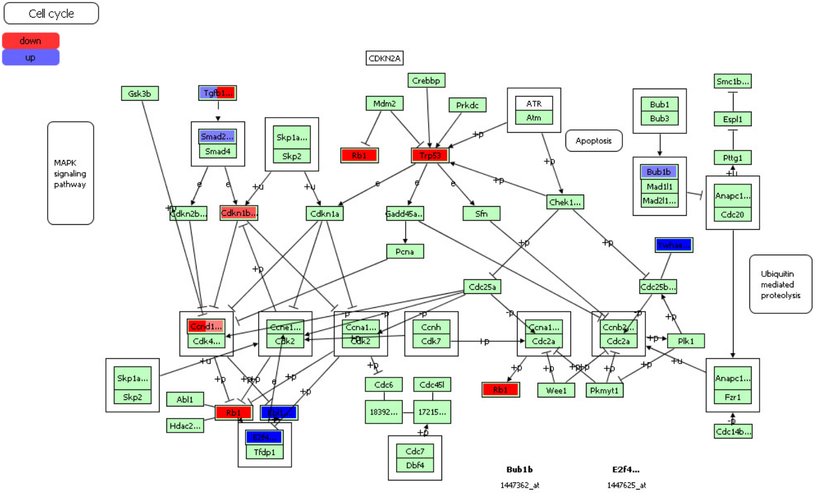


fructose 6-phosphate
PP
P
ATP
diphosphate-fructose-6-phosphate 1-phosphotransferase
6-phosphofructokinase (cPFK)
P
fructose-1,6-bisphosphatase (cFBPAse)
ADP
fructose 1,6-bisphosphate
fructose-bisphosphate aldolase (cALD)
glyceraldehyde 3-phosphate
dihydroxyacetone phosphate
triose phosphate isomerase (cTIM)
P
NAD+
glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)
NADH
NADP+
1,3 - Bisphospho - glycerate
glyceraldehyde-3-phosphate dehydrogenase (NADP)
NADPH
ADP
phosphoglycerate kinase (cPGlyK)
ATP
3-phospho-glycerate
phosphoglycerate mutase (cPGlyM)
2-Phospho-glycerate
phosphopyruvate hydratase (cENOLASE)
P
CO2
phosphoenolpyruvate
oxaloacetate
phosphoendpyruvate carboxylase
phosphoenolpyruvate carboxykinase (ATP)
CO2
NAD+
ATP
ADP
malate dehydrogenase
NADH
AMP
PP
ADP
pyruvate, phosphate dikinase (cPPDK)
pyruvate kinase (cPK)
P
ATP
ATP
malate
pyruvate
oxaloacetate
OAA/malate transporter
pyruvate transporter
oxaloacetate
CO2
NAD+
NADH
malate
pyruvate
malate dehydrogenase (decarboxylating)
P
P
triosephosphat/P translocator (TPT1_GAP)
P
P
triosephosphat/P translocator (TPT3_3-PGA)
P
P
phosphoenolpyruvat/phosphat transporter
fructose 6-phosphate
ATP
P
6-phosphofructokinase (pPFK)
fructose-1,6-bisphosphatase (pFBPase)
ADP
fructose 1,6-bisphosphate
fructose-bisphosphate aldolase (pALD)
P
P
glyceraldehyde 3-phosphate
dihydroxyacetone phosphate
dihydroxyacetone phosphate
triose phosphate isomerase (pTIM)
triosephosphat/P translocator (TPT2_DHAP)
NADP+
glyceraldehyde-3-phosphate dehydrogenase (NADP+) (phosphorylating)
NADPH
P
1,3 - Bisphospho - glycerate
ADP
phosphoglycerate kinase (pPGlyK)
ATP
3-phospho-glycerate
phosphoglycerate mutase (pPGlyM)
2-Phospho-glycerate
phosphopyruvate hydratase (pENOLASE)
phosphoenolpyruvate
AMP
PP
ADP
pyruvate, phosphate dikinase (pPPDK)
pyruvate kinase (pPK)
P
ATP
ATP
pyruvate
NADP+
malate dehydrogenase (oxaloacetate-decarboxylating) (NADP+)
L-glutamate
L-glutamate
NADPH
CO2
malate
malate
malate/glutamate transporter

Examples - Data Integration (Potato)



Examples - Data Exploration (Mouse)



Examples - Visualisation [Tognetti et al. (2007)]

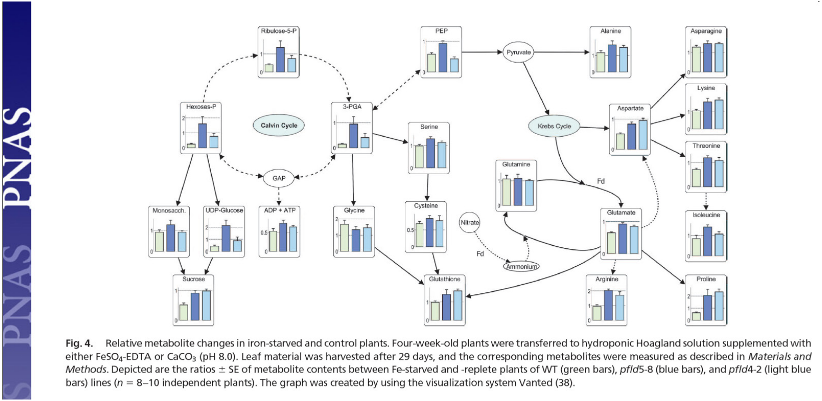
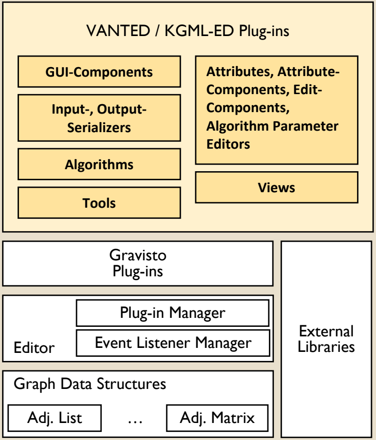


Fig. 4. Relative metabolite changes in iron-starved and control plants. Four-week-old plants were transferred to hydroponic Hoagland solution supplemented with either $FeSO_4$-EDTA or $CaCO_3$ (pH 8.0). Leaf material was harvested after 29 days, and the corresponding metabolites were measured as described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

System Architecture

- Based on the extensible, plugin-based graph visualization toolkit Gravisto
- Event management (observer design pattern)
- MVC concept



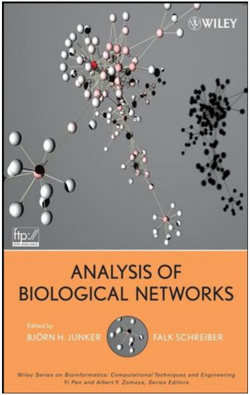
System architecture

Analysis of Biological Networks

1. Motivation
2. Foundations
3. Network motifs
4. Network centralities
5. Network-related data analysis and visualisation

Tools

- Mavisto - motif analysis and visualisation tool
  http://mavisto.ipk-gatersleben.de
- CentiBiN - centrality analysis in biological networks
  http://centibin.ipk-gatersleben.de
- Vanted - analysis and visualisation of experimental data in the network context
  http://vanted.ipk-gatersleben.de
- MetaCrop - information system for plant specific metabolism
  http://metacrop.ipk-gatersleben.de
- KGML-ED - KEGG pathway explorer and editor
  http://kgml-ed.ipk-gatersleben.de


WILEY
ftp://
ANALYSIS OF
BIOLOGICAL NETWORKS
Edited by
BJÖRN H. JUNKER
FALK SCHREIBER
Wiley Series on Bioinformatics: Computational Techniques and Engineering
Yi Pan and Albert Y. Zomaya, Series Editors