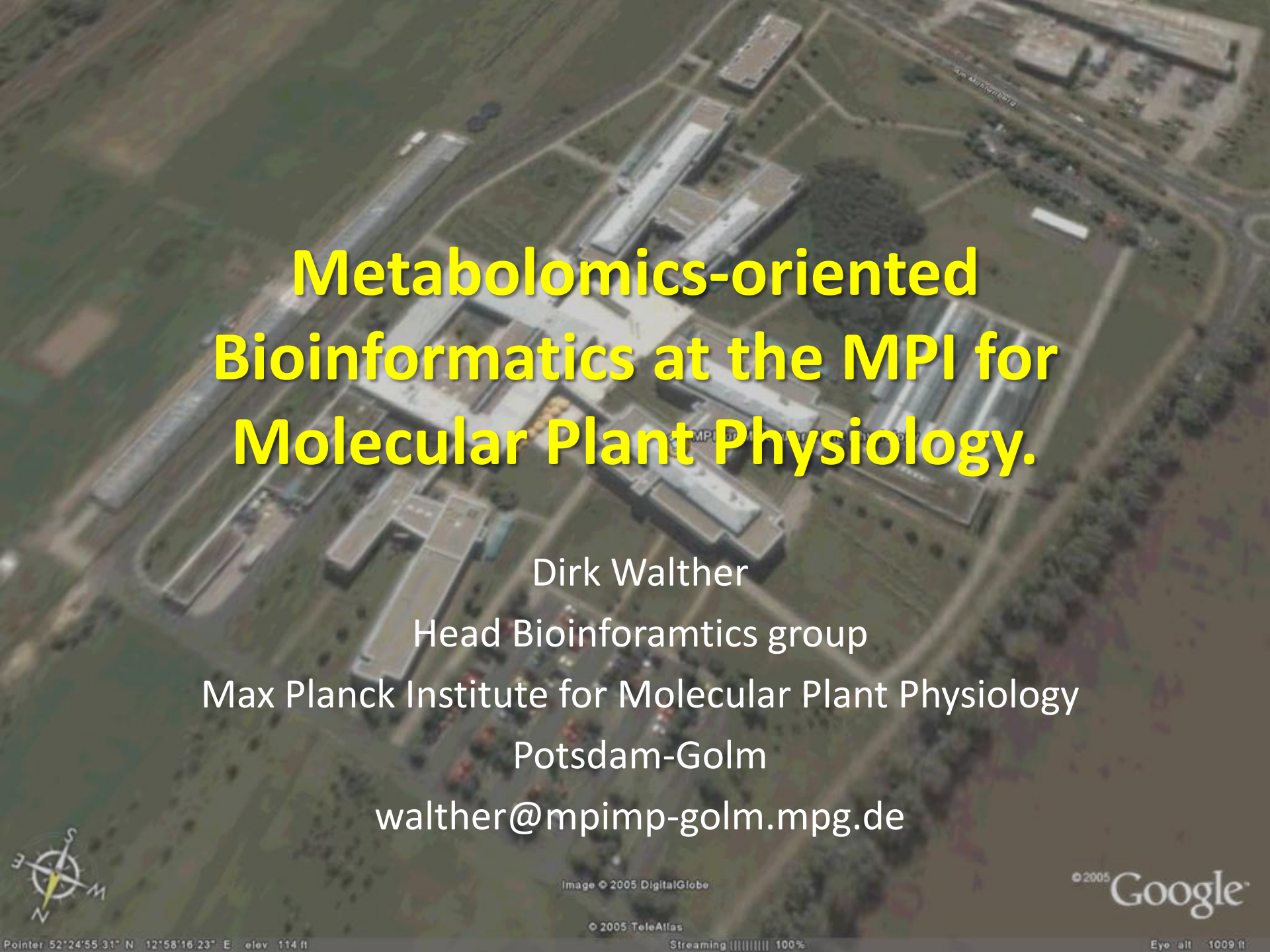


Metabolomics-oriented Bioinformatics at the MPI for Molecular Plant Physiology.

Dirk Walther

Head Bioinforamtics group

Max Planck Institute for Molecular Plant Physiology

Potsdam-Golm

walther@mpimp-golm.mpg.de

Overview

GC/MS Data

- Metabolomics databases: The Golm Metabolome Database (GMD)
- Automated GC/MS-spectra interpretation

Metabolites

- Metabolite-transcript causal relationships from time course data

Pathways

- Metabolomics-assisted genome annotation in *Chlamydomonas reinhardtii*

Genomes

From GC/MS Data to Metabolites

GC/MS Data

- **Metabolomics databases: The Golm Metabolome Database (GMD)**
- **Automated GC/MS-spectra interpretation**

Metabolites

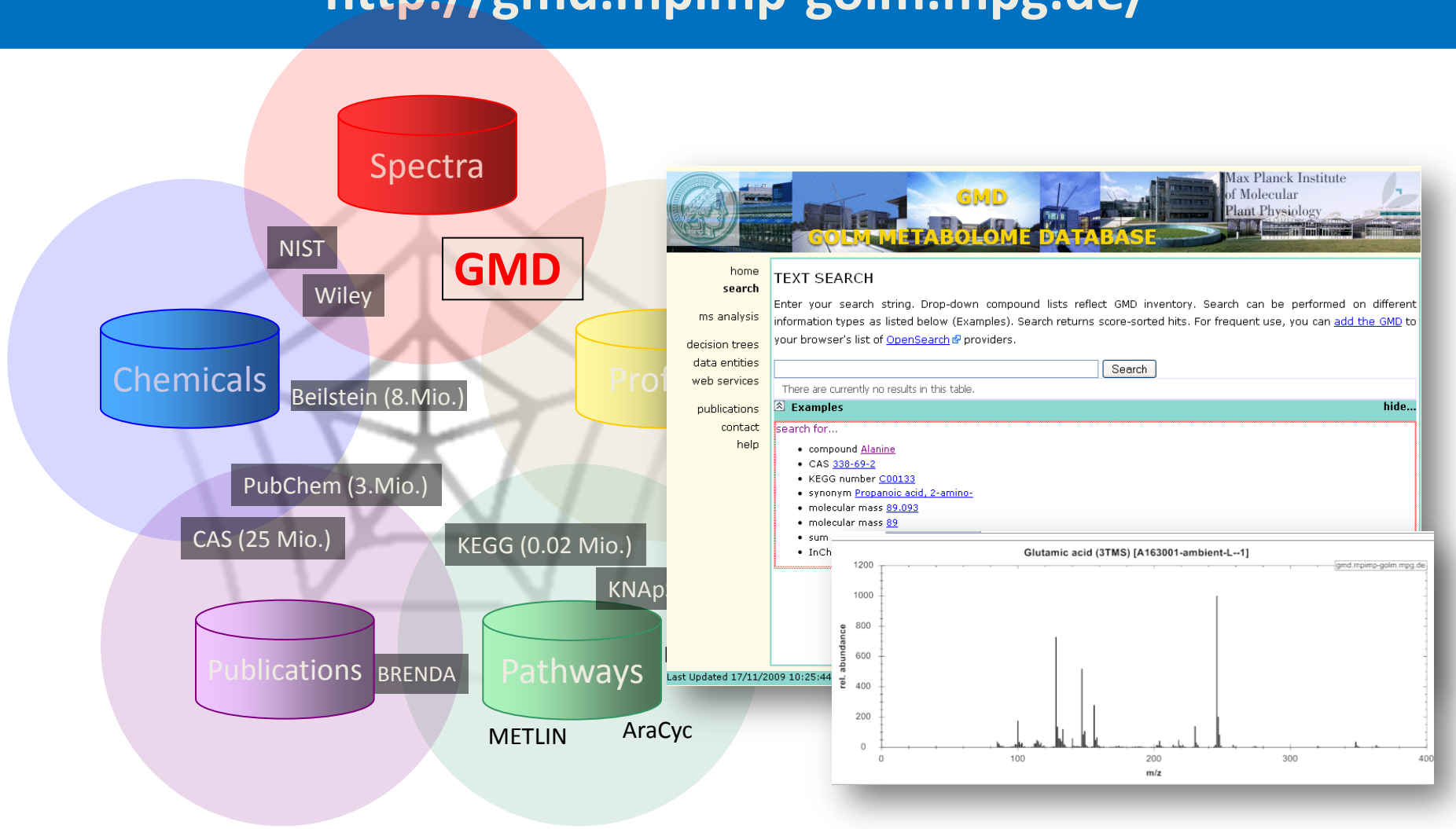
- Metabolite-transcript causal relationships from time course data

Pathways

- Metabolomics-assisted genome annotation in *Chlamydomonas reinhardtii*

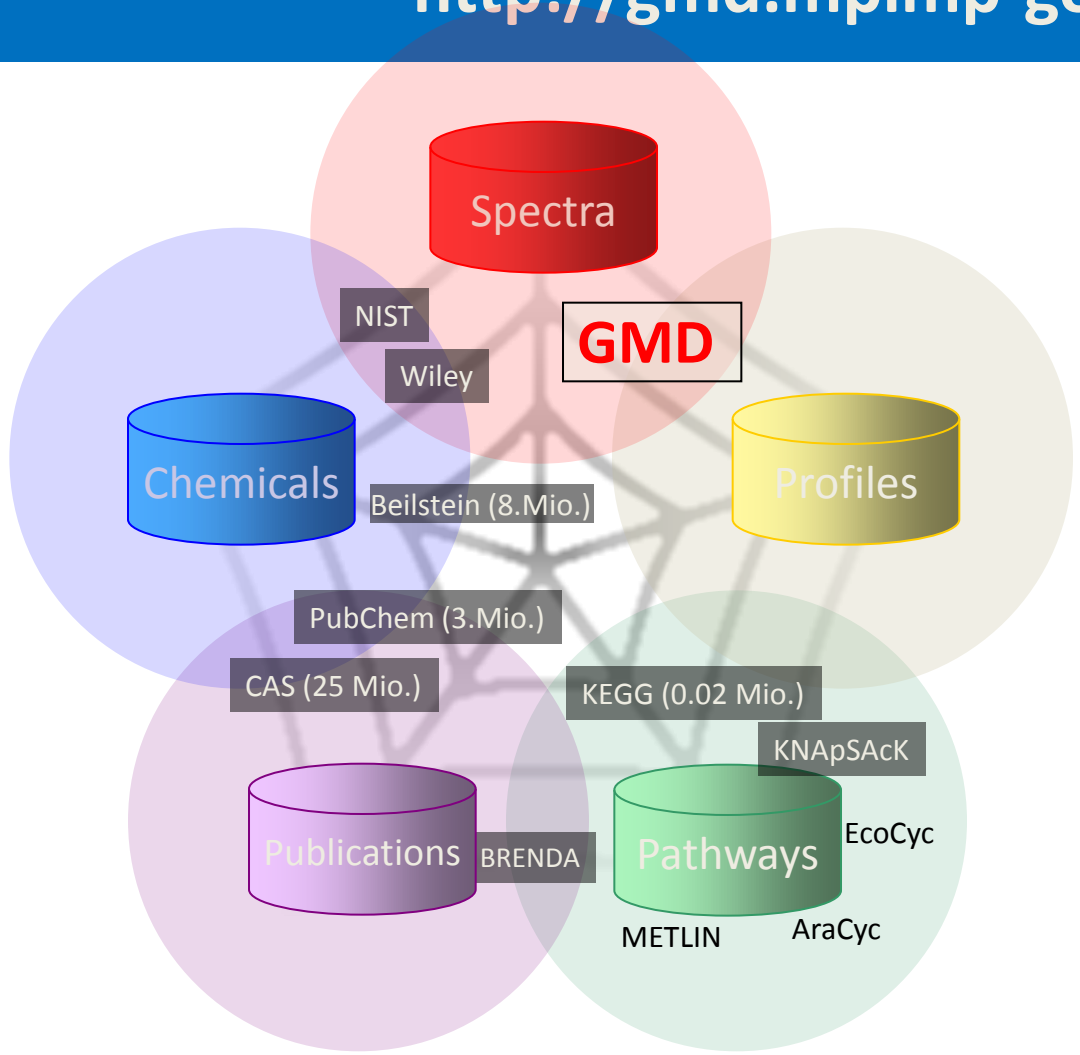
Genomes

GMD – Golm Metabolome Database
http://gmd.mpimp-golm.mpg.de/



- Kopka et al., (2005) GMD@CSBDB: The Golm metabolome database. Bioinformatics 21:1635-1638
- Hummel et al. (2007) *The Golm Metabolome Database: a Database for GC-MS based Metabolite Profiling*. In: Hohmann, S. (ed) Topics in Current Genetics: Nielsen, J., Jewett, M. (eds) Metabolomics. Springer-Verlag, Berlin Heidelberg New York,

GMD – Golm Metabolome Database
http://gmd.mpimp-golm.mpg.de/

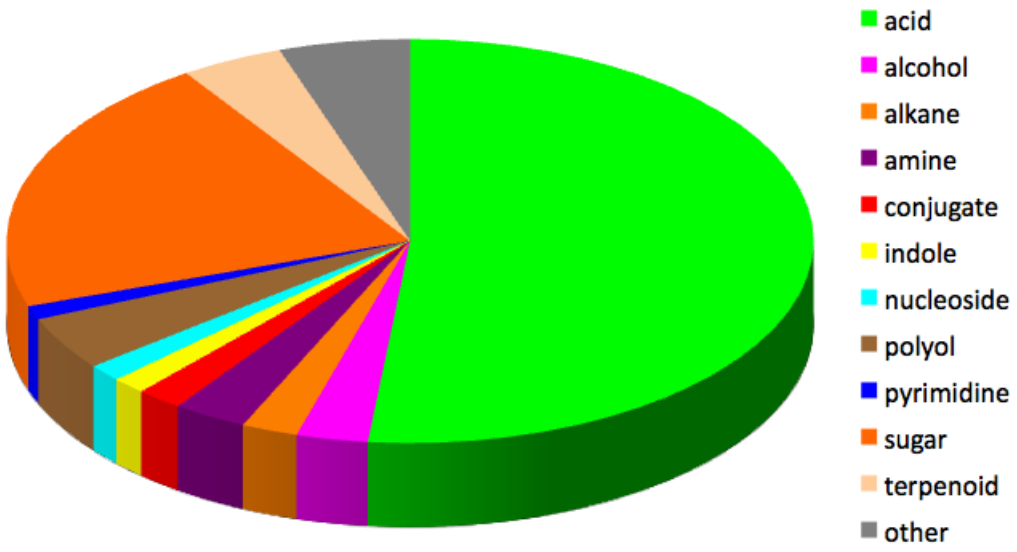


GMD in numbers

Reference substances:	3,056
Metabolites:	830
Analytes (derivatized compound):	3,175
Number of spectra:	24,383
Profiles (samples):	668

- Kopka et al., (2005) GMD@CSBDB: The Golm metabolome database. Bioinformatics 21:1635-1638
- Hummel et al. (2007) *The Golm Metabolome Database: a Database for GC-MS based Metabolite Profiling*. In: Hohmann, S. (ed) Topics in Current Genetics: Nielsen, J., Jewett, M. (eds) Metabolomics. Springer-Verlag, Berlin Heidelberg New York,

GMD – Inventory, Reference substances

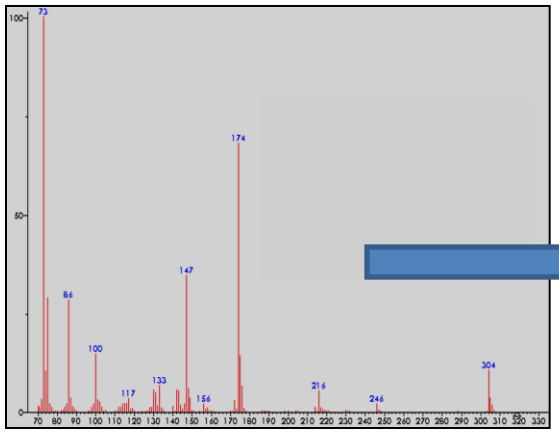


acid	52%
sugar	21%
polyol	4%
terpenoid	4%
alcohol	3%
amine	3%
alkane	2%
conjugate	2%
indole	1%
nucleoside	1%
pyrimidine	1%

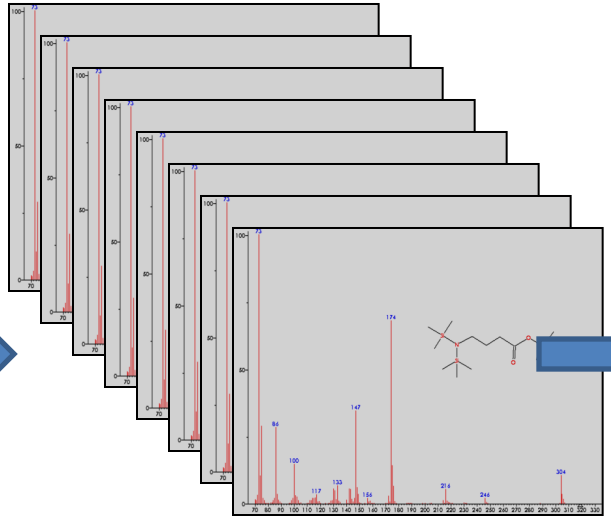
others less than 1%:

aldehyde, alkaloid, amide, calystegine, chalcone, flavonoid, imide, lactam, nucleotide, purine, stilbene

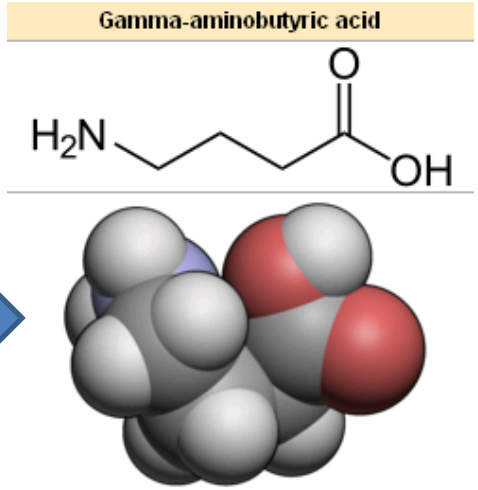
Annotation of GC/MS-spectra



Recorded GC/MS spectrum

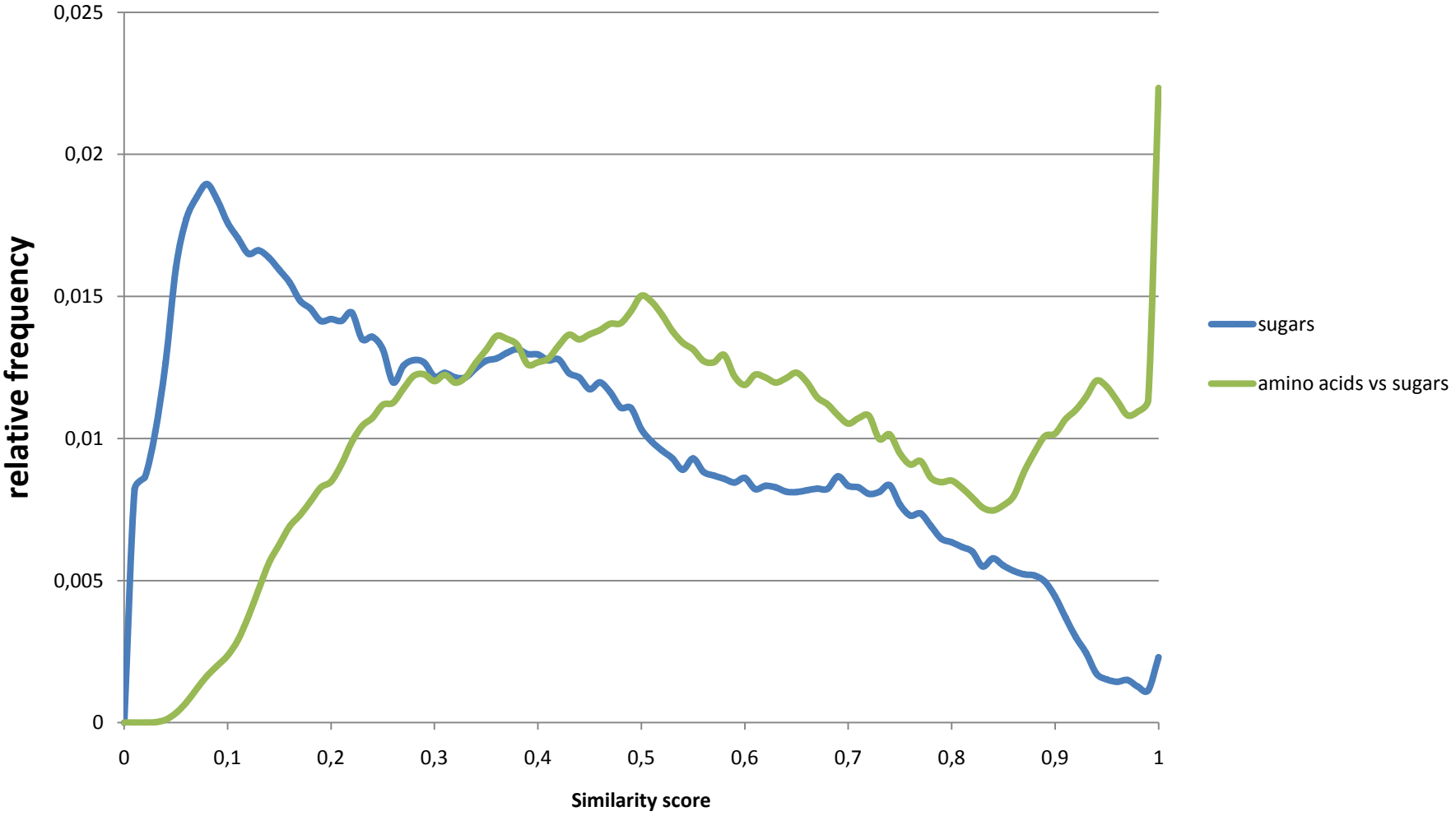


Library of GC/MS spectra for reference substances



Identified Compound

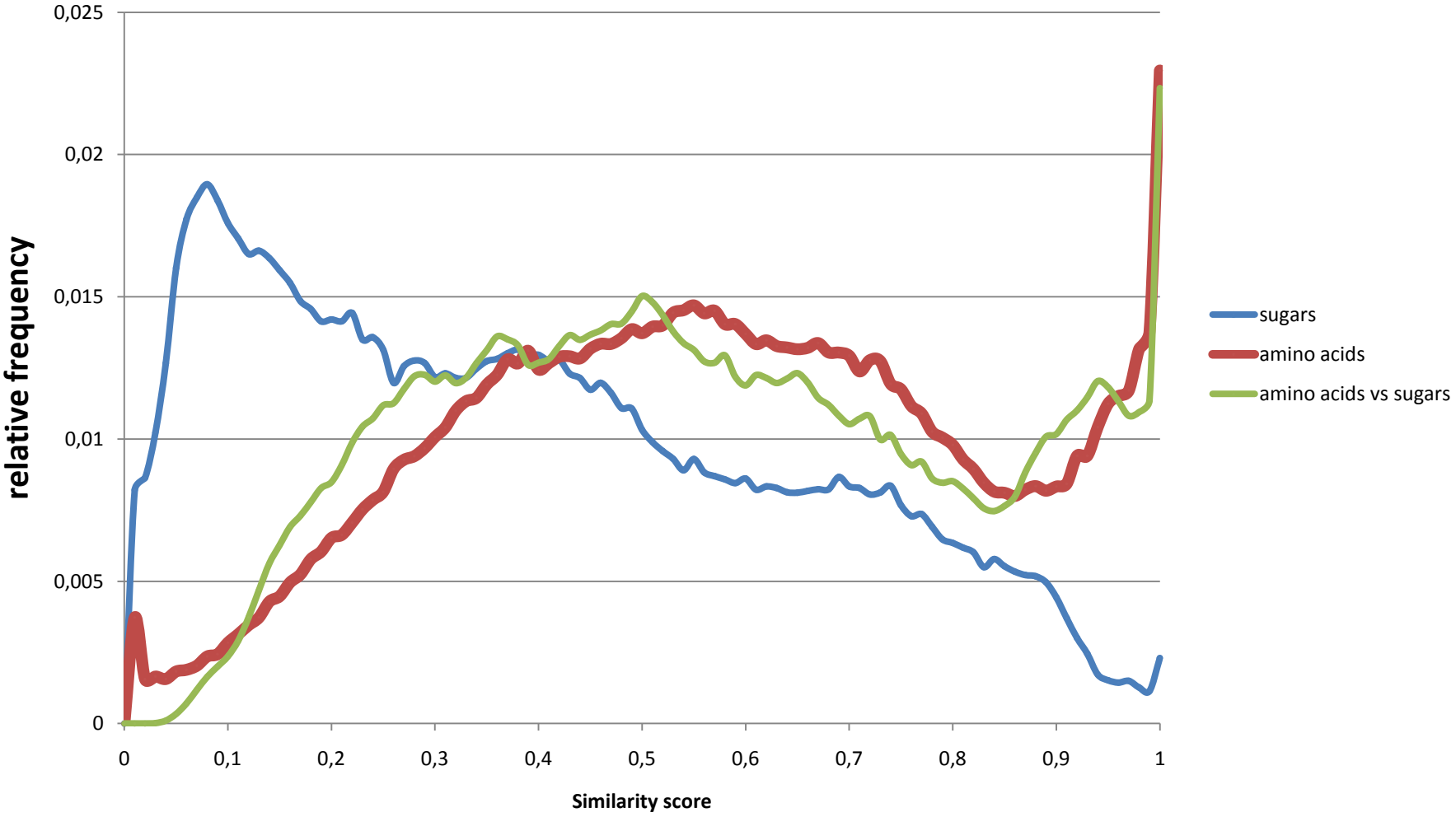
Annotation via GC/MS-spectra comparisons



Similar spectra ⟷ Dissimilar spectra

Without RI information

Annotation via GC/MS-spectra comparisons

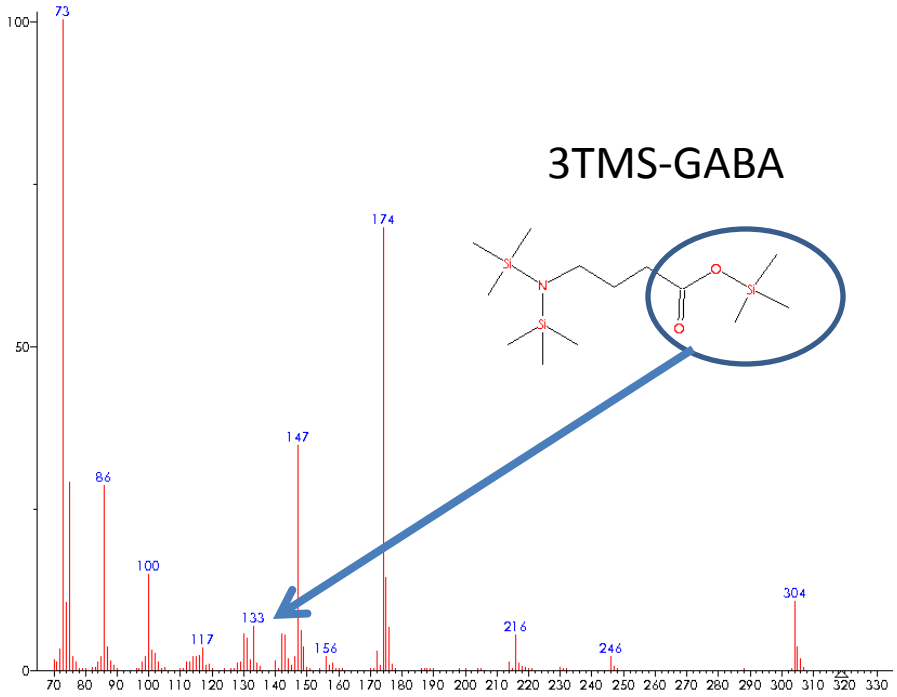


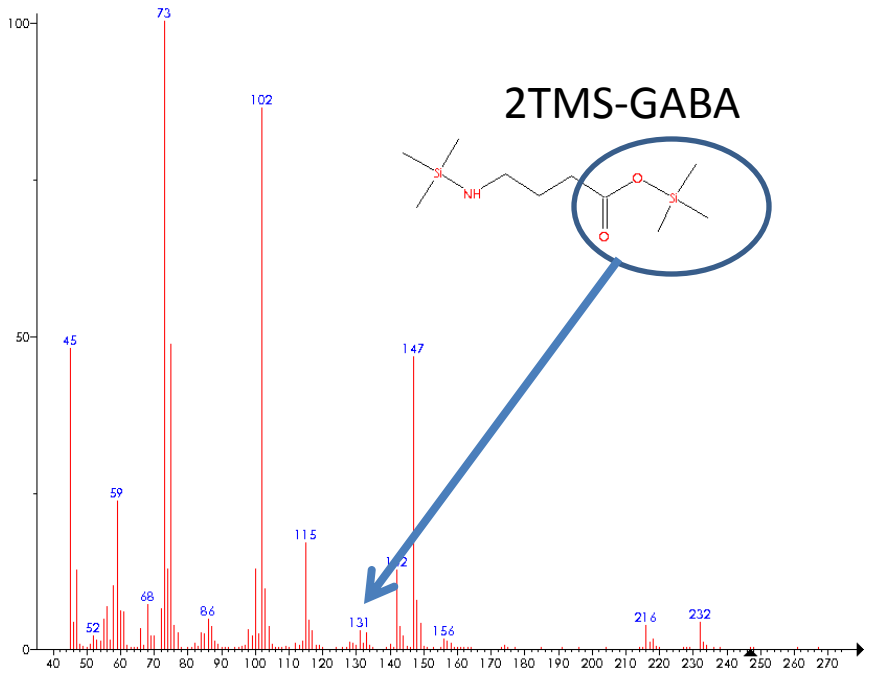
Similar spectra ⟷ Dissimilar spectra

Without RI information

Knowledge-based annotation of GC/MS-spectra

Same metabolite, but different derivatives





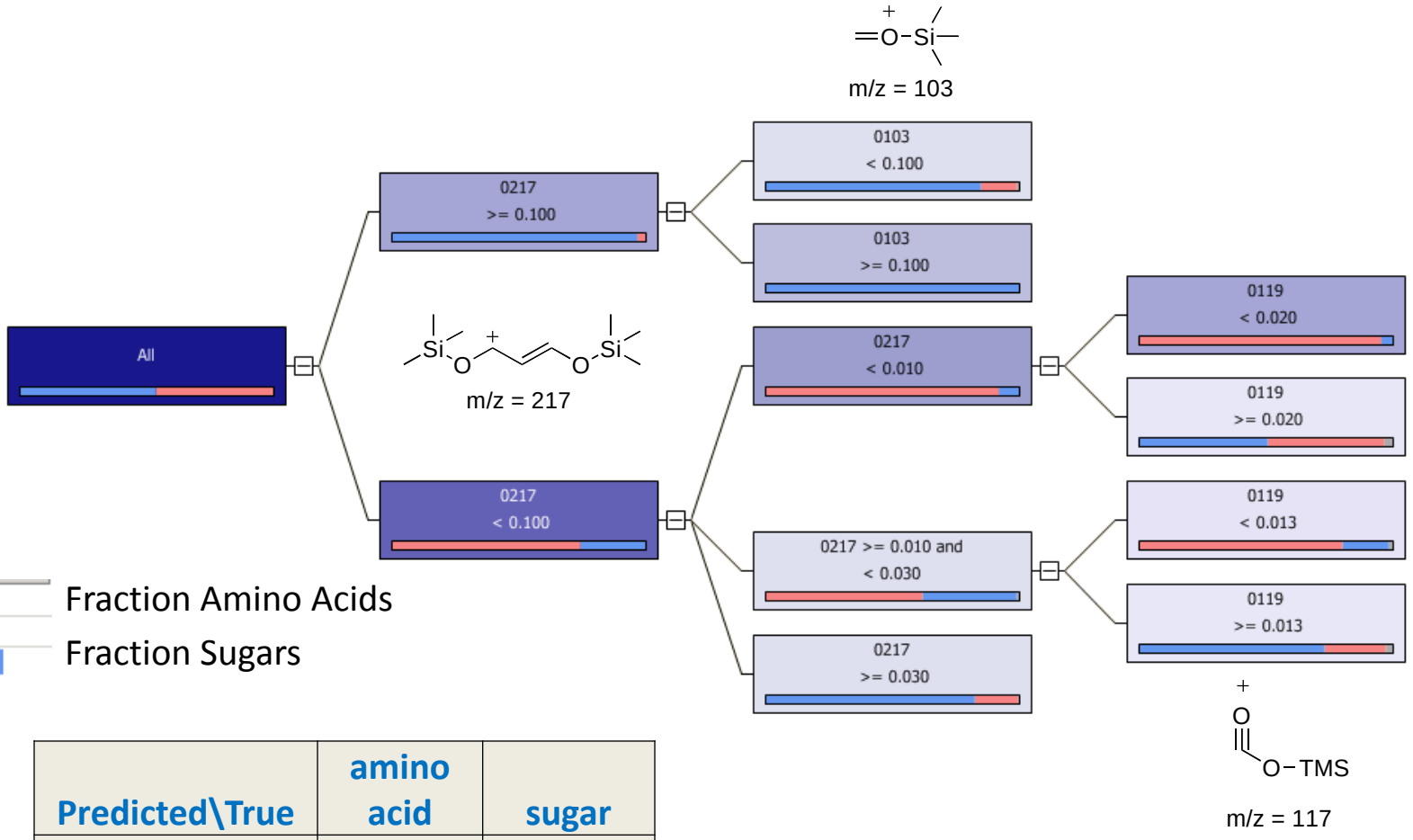
Identification via characteristic mass peaks

Spectra interpretation

Scsibrany & Varmuza (1992) J Anal Chem

Varmuza & Werther (1996). Chem. Inf. Comput. Sci.

Annotation of GC/MS-spectra using decision trees

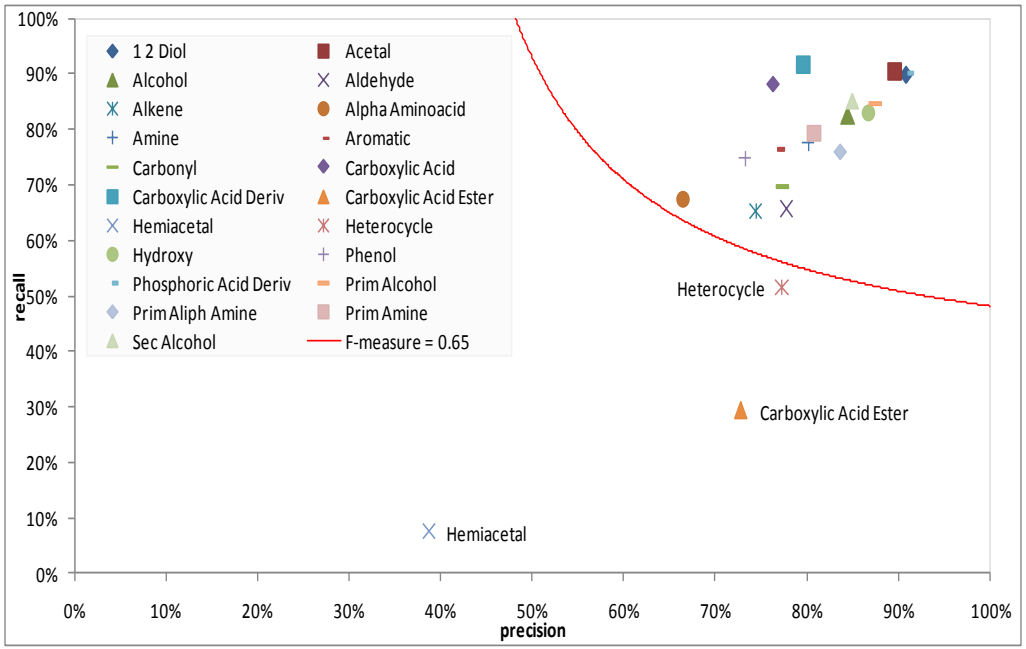


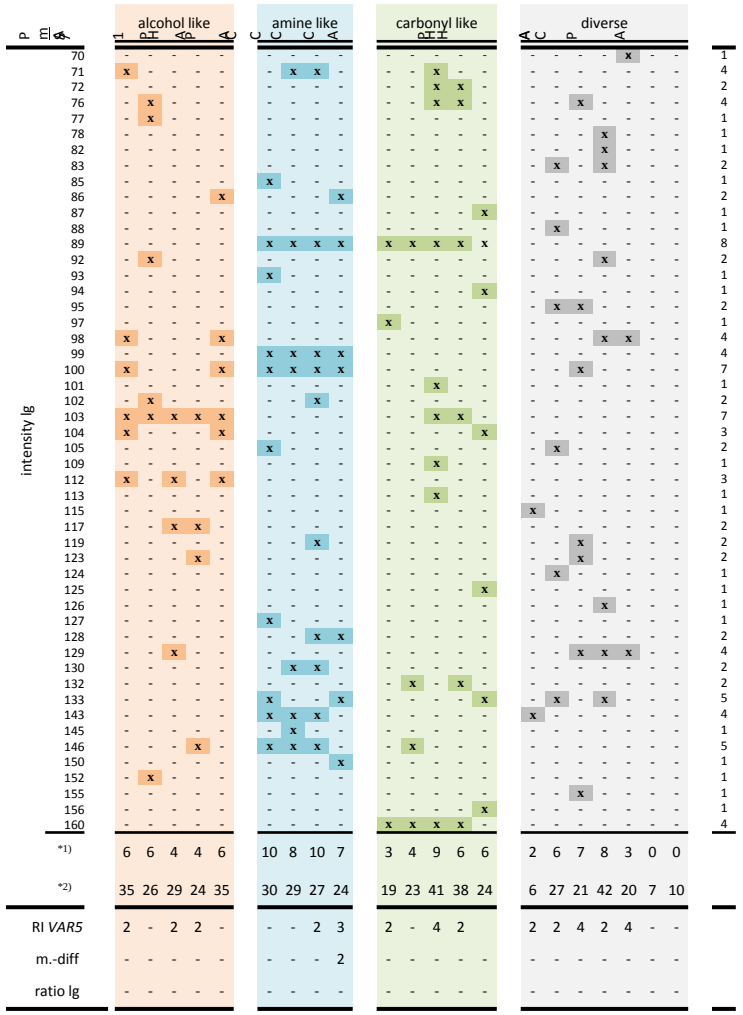
Predicted\True	amino acid	sugar
amino acid	574	42
sugar	89	579

89% correct predictions!

Compound class predictions based on fragments

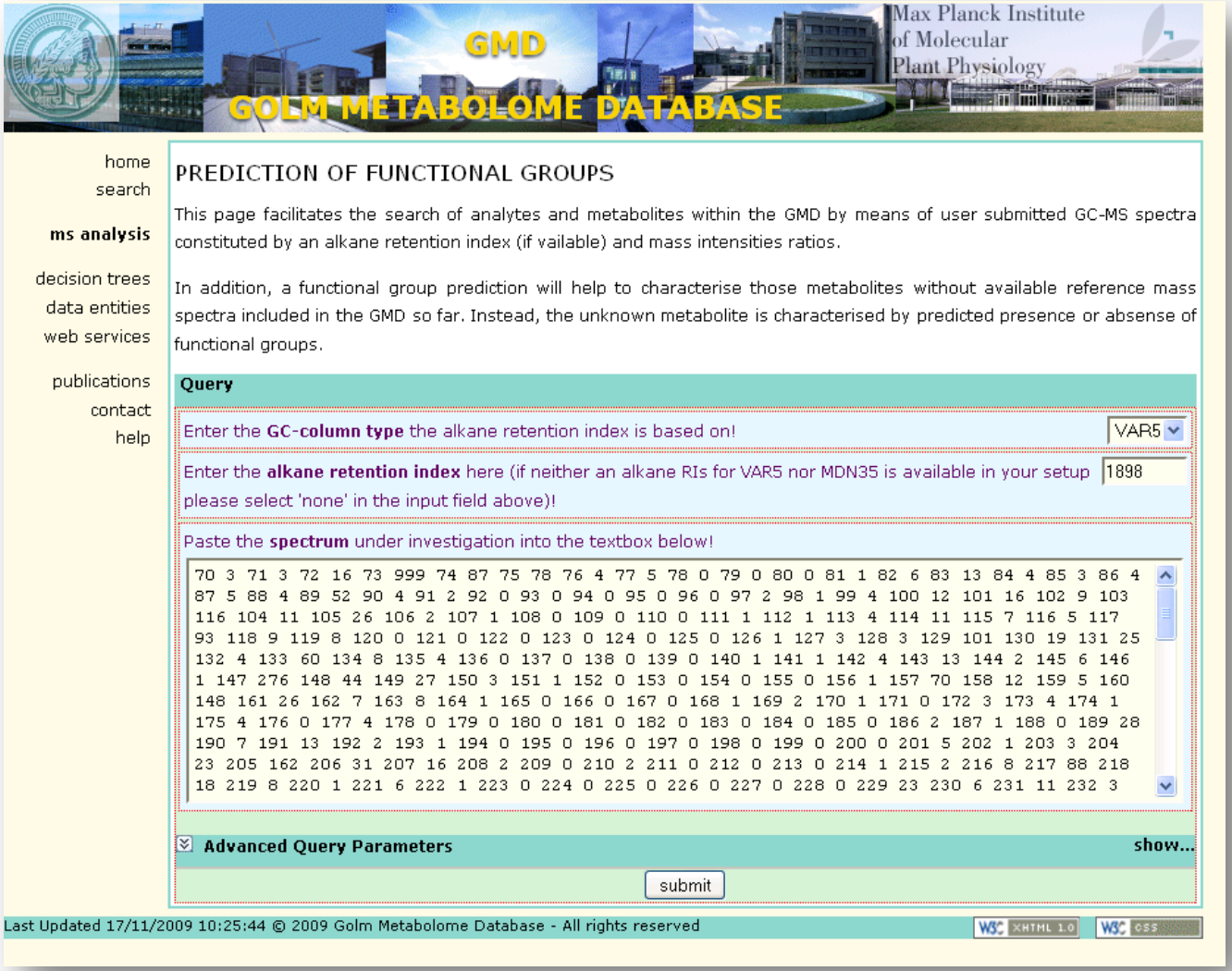
Classification tree models for 20 functional groups





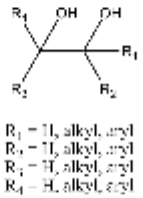
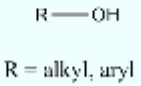
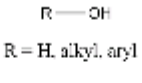
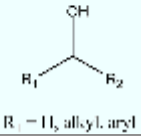
Hummel et al., Submitted

Annotation of user-submitted GC/MS spectra at gmd.mpimp-golm.mpg.de



Annotation of user-submitted GC/MS spectra at gmd.mpimp-golm.mpg.de

Functional Group Prediction Results — hide 8 predicted functional groups ...

functional group	image	prediction	probability	adjusted probability	support	description	contained in Idose
1 2 Diol		present	100.00%	9.71%	651	intensity lg - 217 >= 2.3706843138 and intensity lg - 103 >= 2.5928983808 and intensity ...	true
Hydroxy		present	100.00%	0.89%	779	intensity lg - 191 >= 1.4219502926 and intensity lg - 217 >= 2.3706843138 and intensity ...	true
Alcohol		present	99.98%	1.64%	515	intensity lg - 217 >= 2.3706843138 and intensity lg - 319 >= 0.5253005028	true
Sec Alcohol		present	99.23%	4.98%	900	intensity lg - 217 >= 2.3706843138 and intensity lg - 103 >= 2.067753911 and intensity	true

From Metabolites to Pathways

GC/MS Data

- Metabolomics databases: The Golm Metabolome Database (GMD)
- Automated GC/MS-spectra interpretation

Metabolites

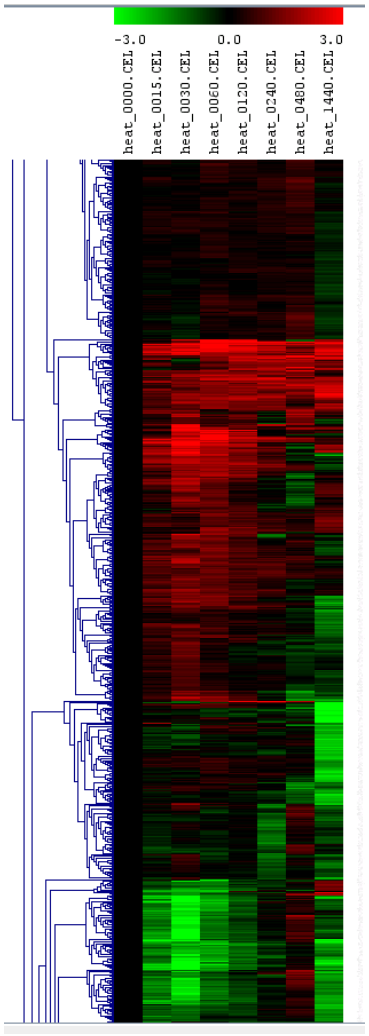
- **Metabolite-transcript causal relationships from time course data**

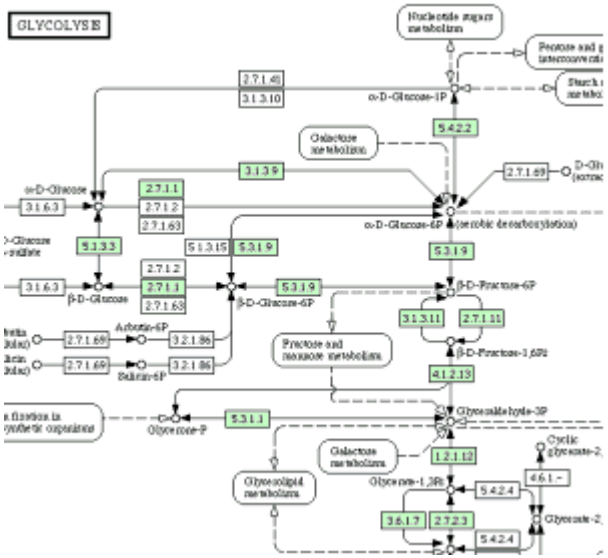
Pathways

- Metabolomics-assisted genome annotation in *Chlamydomonas reinhardtii*

Genomes

Can we infer causal and, thus, pathway relationships from metabolomics data?

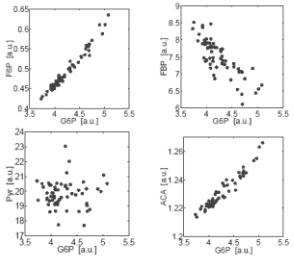




Which molecules function as **cause**, which as **effect**?

From Correlation Networks to Pathways

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_X S_Y}$$

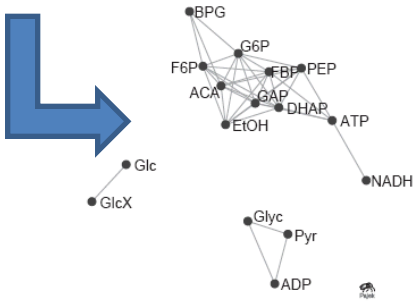


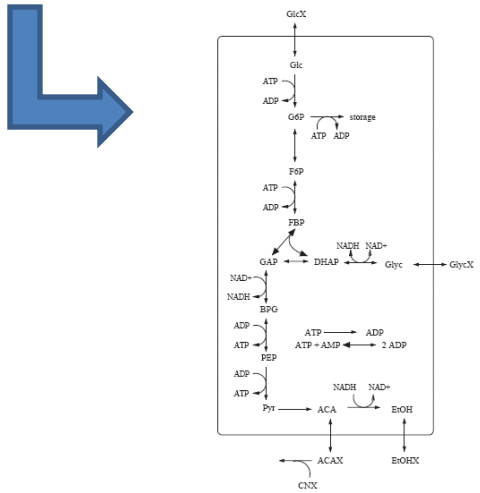
X,Y…. gene expression level, metabolite level etc.

X,Y are connected (edge) if $r > r_c$

Is there a straightforward connection between the underlying system and the observed correlations?

Can we deduce novel pathways based on the observed correlation matrix?





Steuer et al. (2003) Bioinformatics 19: 1019

Yeast Transcripts & Metabolite Profiles in Response to Temperature Stress: Time series data

Temperature shifts:

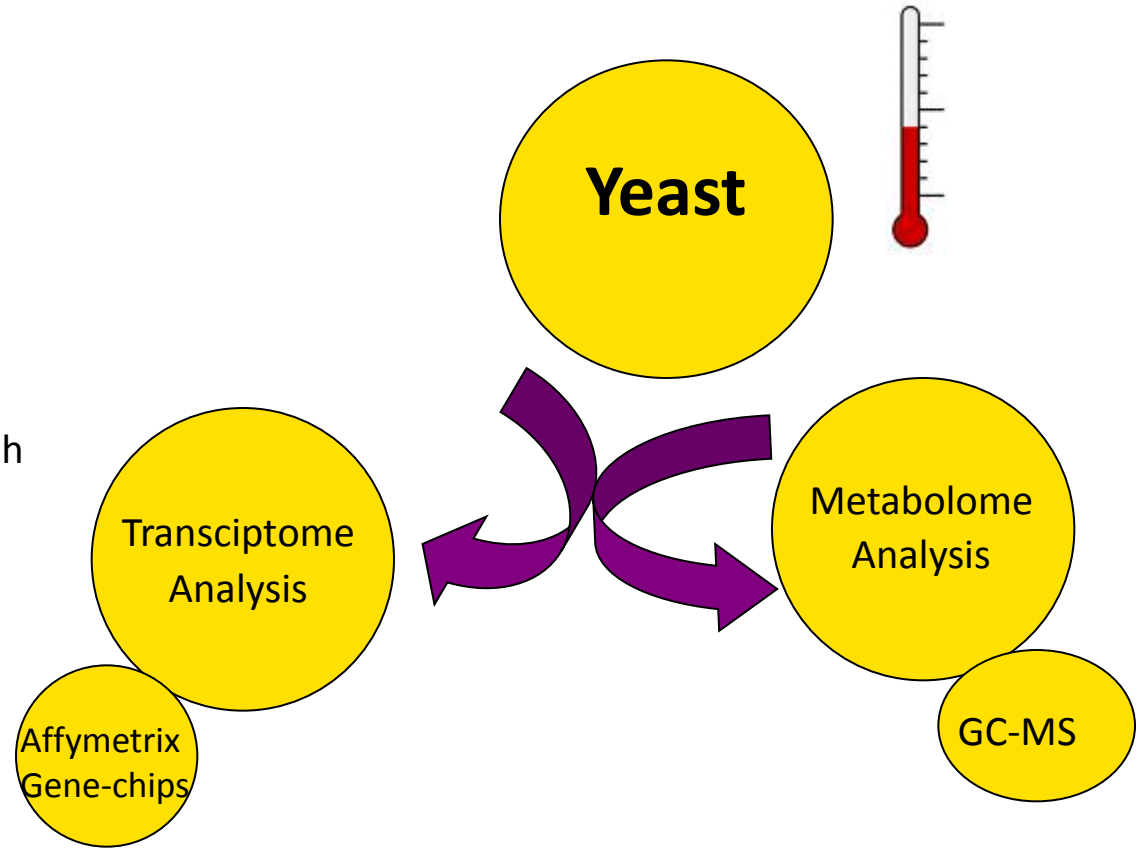
28°C → 37 C heat

28°C → 10°C cold

28°C → 28°C control

Time series: 8 time points

0, 0.25, 0.5, 1, 2, 4, 8 and 24 h



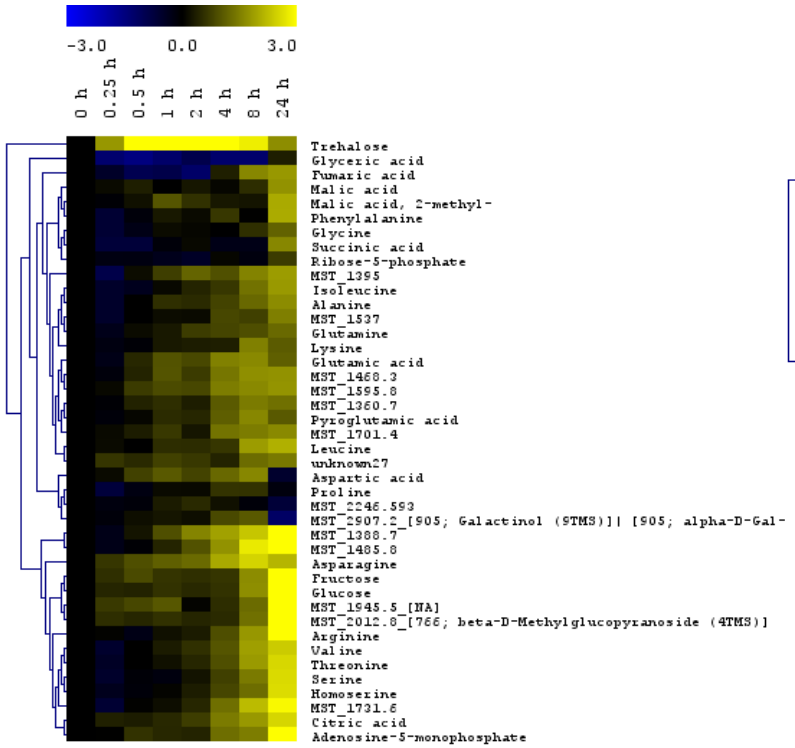
Katrin Straßburg, Joachim Kopka

Metabolite Profiles

GC/MS with *in vivo* stable isotope labeling for accurated quantification

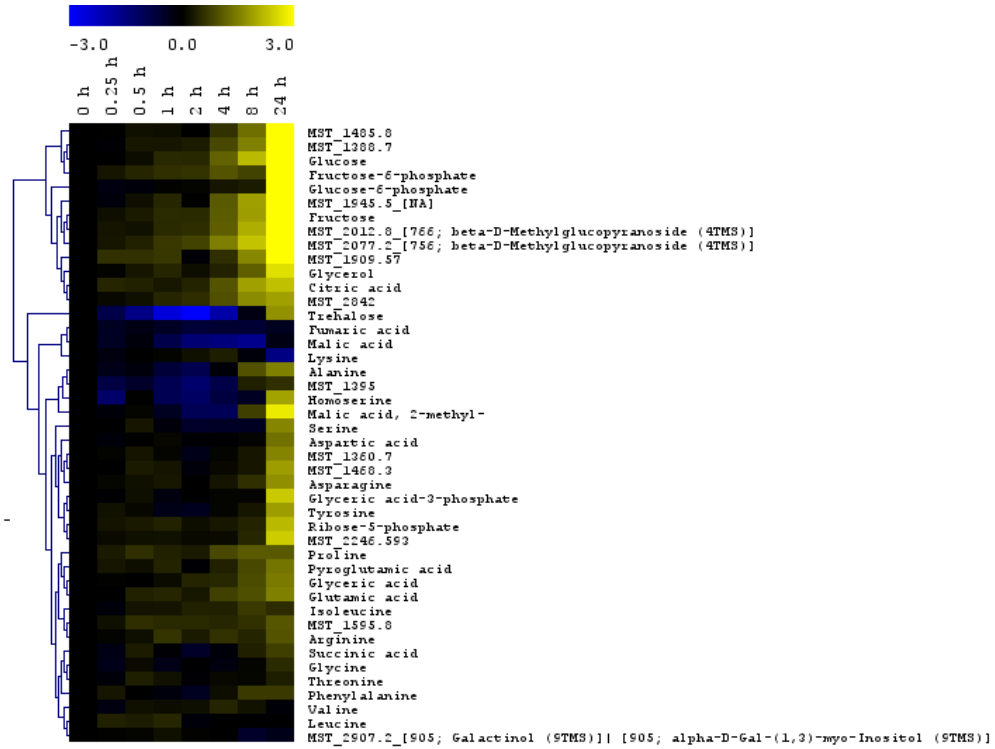
Heat Stress

- 42 metabolites
- 11 unknowns



Cold Stress

- 44 metabolites
- 13 unknowns

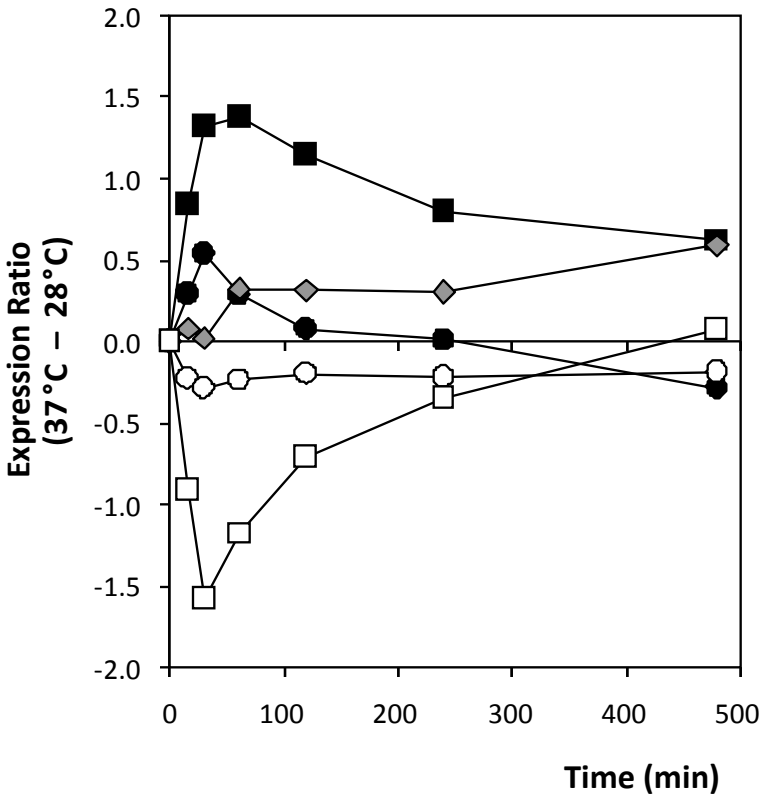


Katrin Straßburg, Joachim Kopka

Transcript Profiles

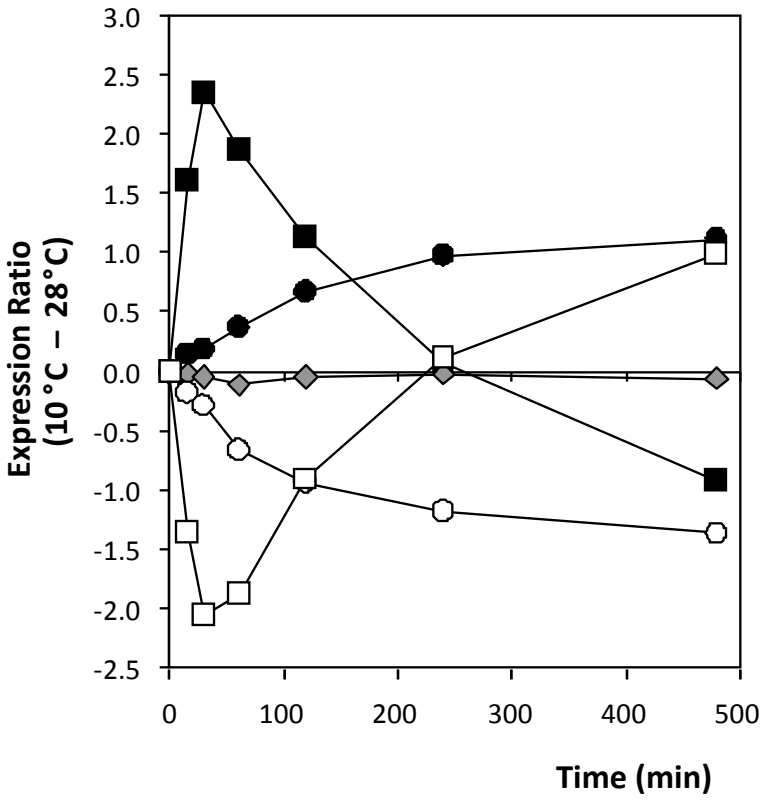
Affymetrix whole yeast genome microarray

Heat Stress



- Cluster 2 427 genes [7%]
- Cluster 3 1507 genes [26%]
- Cluster 4 1392 genes [24%]
- Cluster 5 1876 genes [33%]
- Cluster 1 514 genes [9%]

Cold Stress



- Cluster 1 572 genes [10%]
- Cluster 2 1110 genes [19%]
- Cluster 5 2487 genes [44%]
- Cluster 3 942 genes [16%]
- Cluster 4 605 genes [11%]

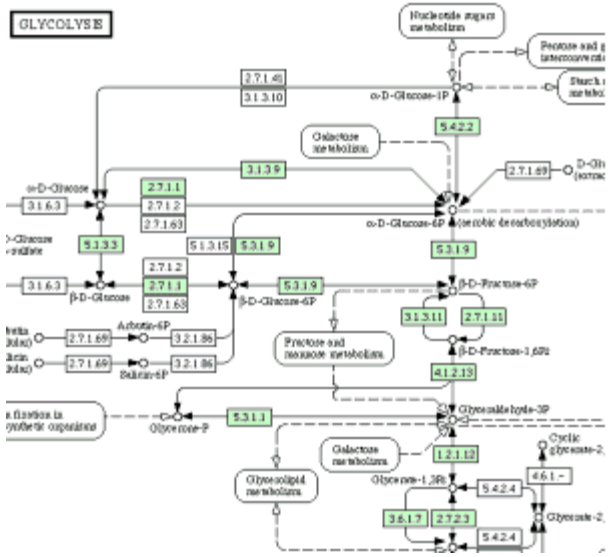
Katrin Straßburg, Joachim Kopka

Deducing pathways from time-course data?

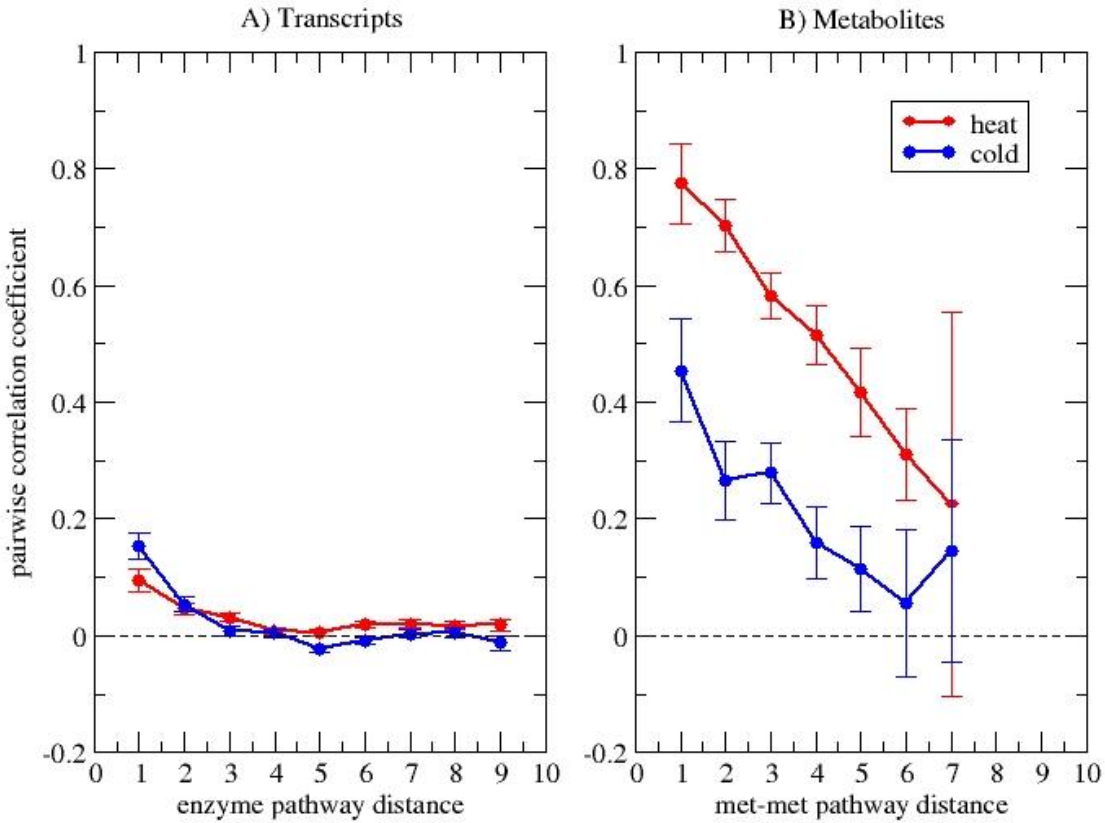
Omics-viewer from YeastCyc

Patrick May

Metabolite correlations are more informative than transcript correlations

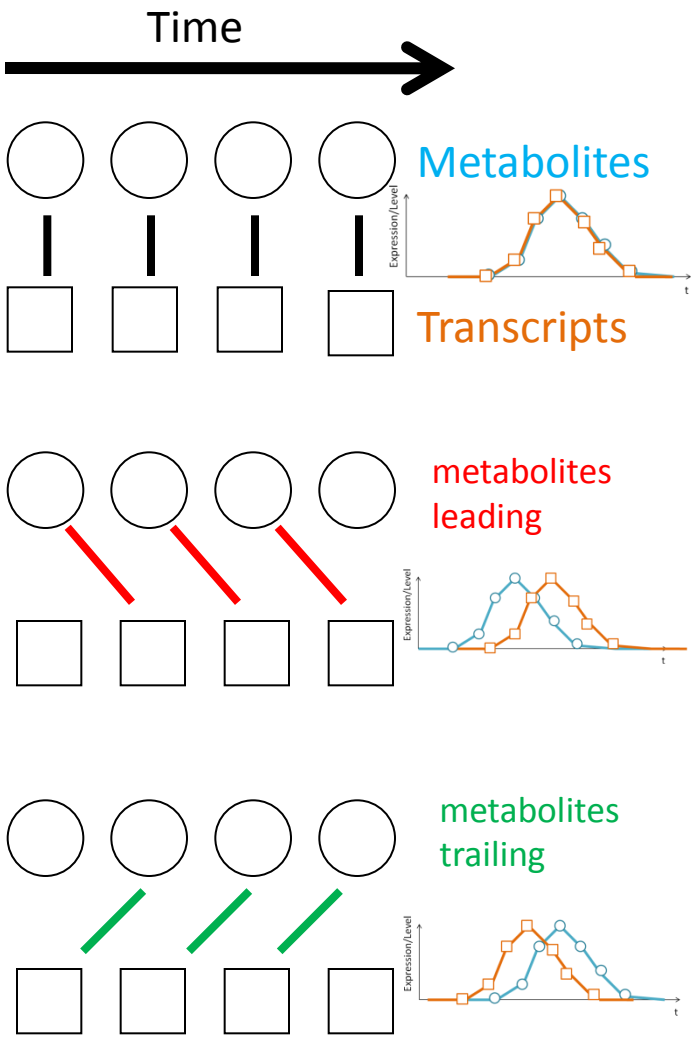


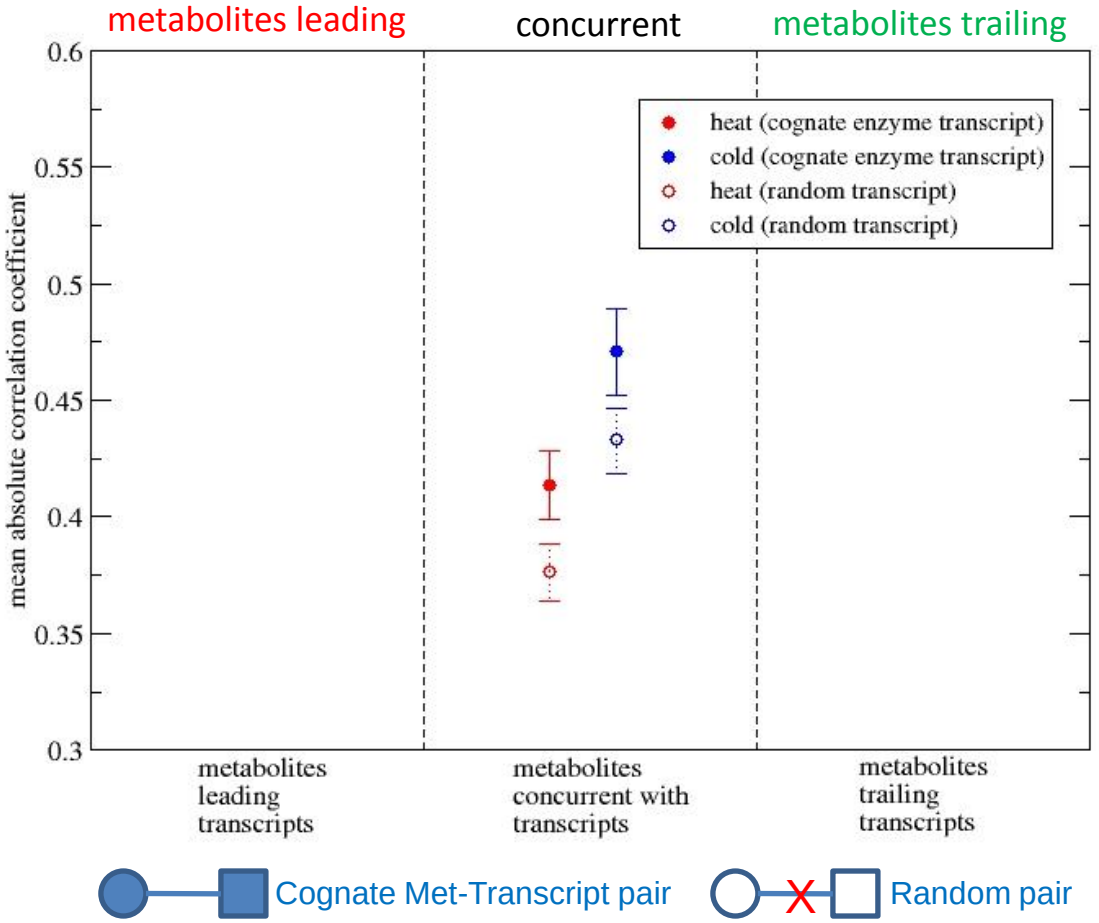
YeastCyc & KEGG yeast pathway maps



Katrin Straßburg, Joachim Kopka

Causal relationships via time-delayed correlations





Stronger correlations when metabolites are considered leading

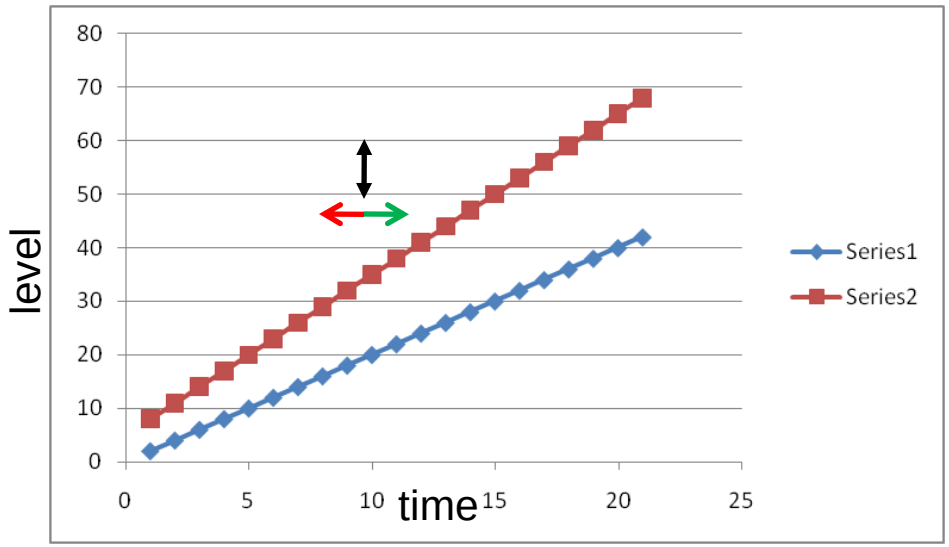
Causal relationships using the concept of Granger causality

$$T(t) = \sum_{i=1}^{d} A_{T,i} T(t-i) + \sum_{i=1}^{d} A_{MT,i} M(t-i) + E_T(t)$$

$$M(t) = \sum_{i=1}^{d} A_{TM,i} T(t-i) + \sum_{i=1}^{d} A_{M,i} M(t-i) + E_M(t)$$

T variable 1 (transcript)
M variable 2 (metabolite)
A parameters
E residual error

Monotonic Signals:



High Pearson Correlation for all three relative time shifts, r=1

Low Granger causality, p-value=1 data not informative,

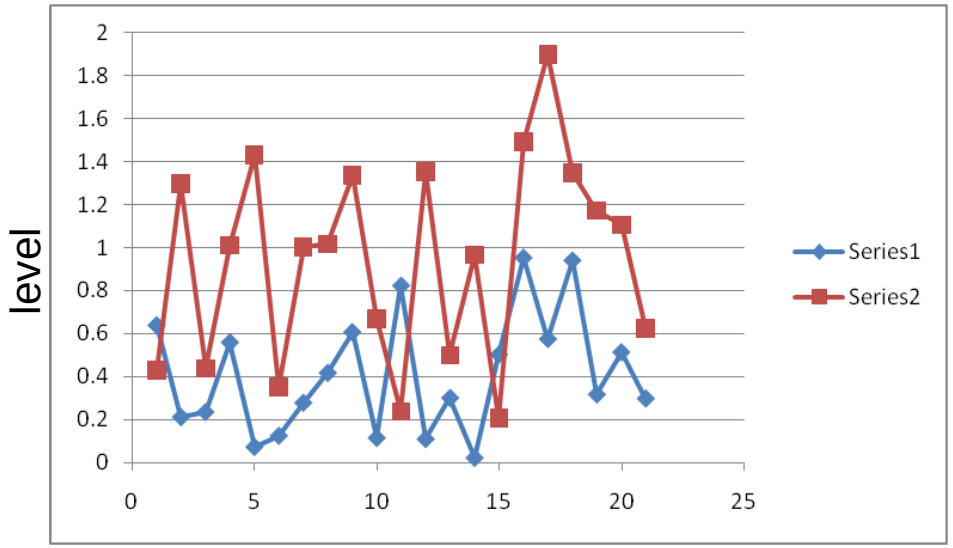
Causal relationships using the concept of Granger causality

$$T(t) = \sum_{i=1}^{d} A_{T,i} T(t-i) + \sum_{i=1}^{d} A_{MT,i} M(t-i) + E_T(t)$$

$$M(t) = \sum_{i=1}^{d} A_{TM,i} T(t-i) + \sum_{i=1}^{d} A_{M,i} M(t-i) + E_M(t)$$

T variable 1 (transcript)
M variable 2 (metabolite)
A parameters
E residual error

Artificial data: s2(t)=f[s1(t-1)]



s1=>s2, p=0
s2=>s1, p=1

Directionality deducable!

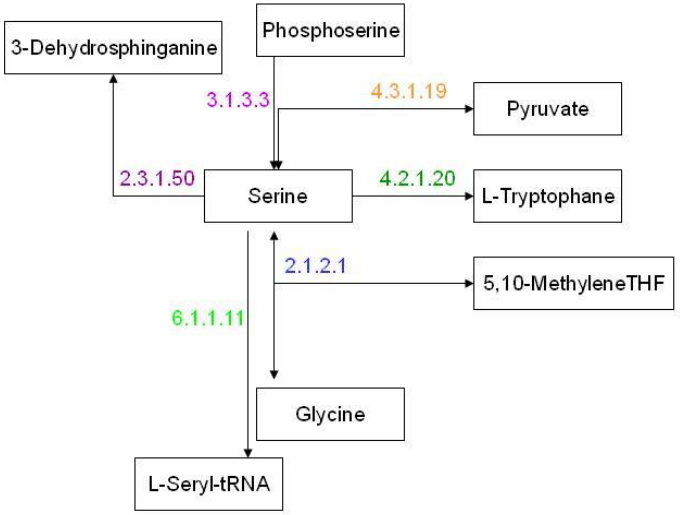
Causal relationships using the concept of Granger causality

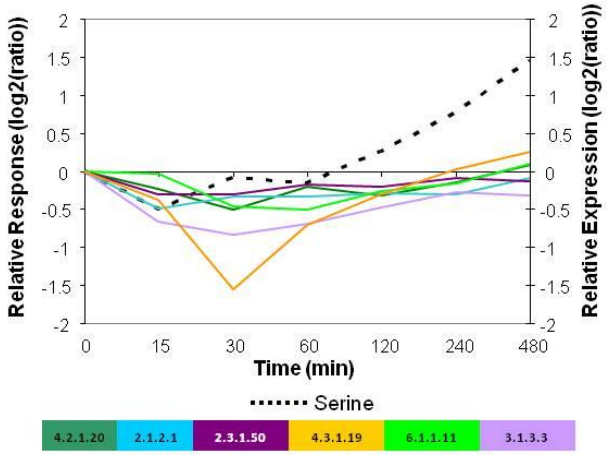
Cause		Effect	p-value	BH-corrected p-value
Heat				
Phenylalanine	->	YPR047W, 1773228_at	3.38E-04	7.88E-02
Adenosine-5-monophosphate	->	YBR115C, 1775044_at	1.46E-03	1.71E-01
Serine	->	YGL026C, 1779478_at	2.37E-03	1.84E-01
YJR109C, 1777097_at	->	Glutamic	3.41E-03	1.99E-01
Cold				
YKL106W, 1776650_at	->	Phenylalanine	1.99E-03	1.51E-01
Fumaric	->	YKL148C, 1772202_at	2.42E-03	1.51E-01
YER073W, 1771943_at	->	Glyceric	2.75E-03	1.51E-01
YBR299W, 1774491_s_at	->	Fructose	3.23E-03	1.51E-01
Arginine	->	YDR341C, 1773272_at	3.38E-03	1.51E-01
YFL022C, 1779336_at	->	Phenylalanine	3.95E-03	1.51E-01

Multiple Testing Correction problem: ~450 tests (metabolites-cognate transcript)
Solution: Take only direction of greater significance (cuts number of tests in half)

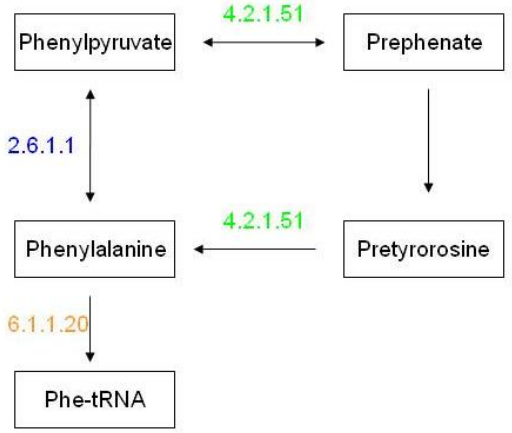
Cause/Effect Metabolites: Examples

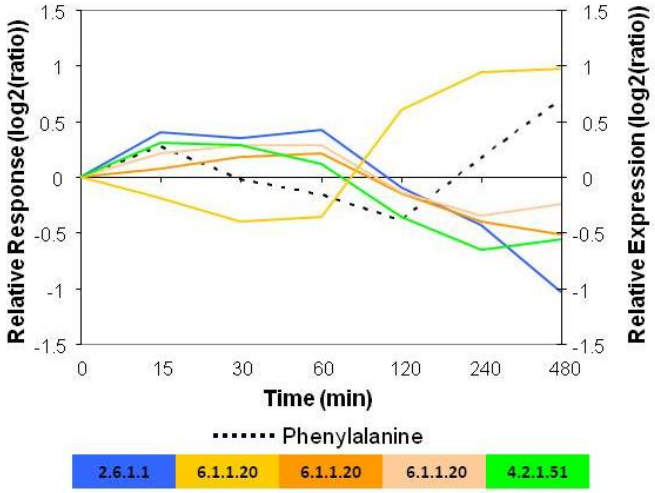
Serine= Cause metabolite





Phenylalanine= Effect metabolite





Walther, Strassburg, Kopka, submitted

Identification of central Cause-Metabolites

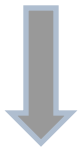
Glutamic acid and **serine** identified as central **cause metabolites**

Metabolite	is Cause	is Effect	p-value	BH-corrected p-value
Glutamic acid	59	19	4.19E-06	1.30E-04
Serine	23	3	4.92E-05	7.63E-04
Fructose-6-phosphate	10	1	6.35E-03	5.12E-02
Adenosine-5-monophosphate	18	5	6.61E-03	5.12E-02
Threonine	10	2	2.25E-02	1.39E-01
Glucose-6-phosphate	8	2	6.54E-02	3.38E-01
Glucose	16	8	1.08E-01	3.87E-01
Aspartic acid	17	9	1.22E-01	3.87E-01
Lysine	5	1	1.25E-01	3.87E-01
Valine	5	1	1.25E-01	3.87E-01
Proline	6	2	1.80E-01	4.28E-01
Isoleucine	6	2	1.80E-01	4.28E-01
Leucine	6	2	1.80E-01	4.28E-01
Alanine	7	3	2.27E-01	5.02E-01
Phenylalanine	5	9	2.50E-01	5.17E-01
Glyceric acid-3-phosphate	2	0	3.02E-01	5.85E-01
Glutamine	9	6	4.53E-01	6.71E-01
Citric acid	2	4	4.53E-01	6.71E-01
Glyceric acid	4	2	4.53E-01	6.71E-01
Glycerol	2	4	4.53E-01	6.71E-01
Homoserine	4	2	4.54E-01	6.71E-01
Trehalose	5	3	5.08E-01	7.09E-01
Asparagine	4	6	5.49E-01	7.09E-01
Malic acid	6	4	5.49E-01	7.09E-01
Fumaric acid	11	13	6.88E-01	7.92E-01
Ribose-5-phosphate	11	13	6.90E-01	7.92E-01
Tyrosine	2	3	6.90E-01	7.92E-01
Glycine	7	7	1.00E+00	1.00E+00

is Cause
is Effect

From Metabolites and Pathways to Genomes

GC/MS Data



- Metabolomics databases: The Golm Metabolome Database (GMD)
- Automated GC/MS-spectra interpretation

Metabolites

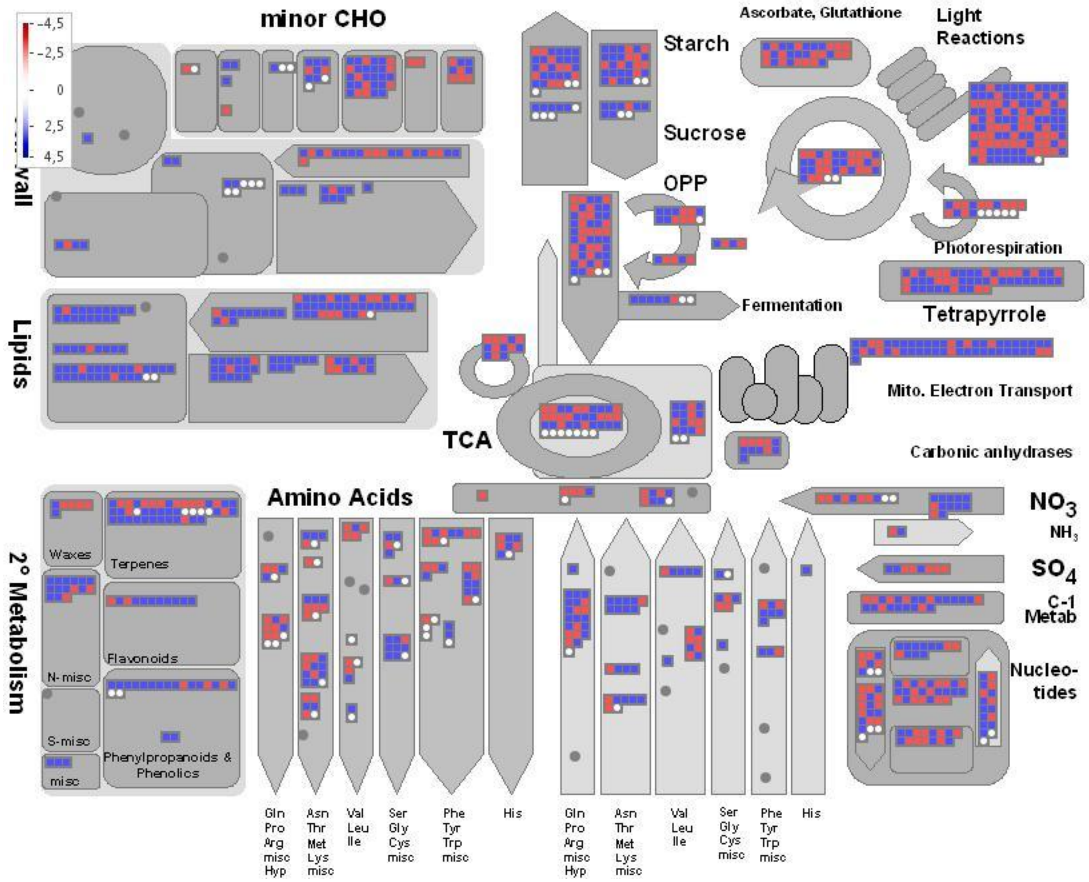
- Metabolite-transcript causal relationships from time course data

Pathways

- **Metabolomics-assisted genome annotation in *Chlamydomonas reinhardtii***

Genomes

Metabolomics & Proteomics-assisted genome annotation in *Chlamydomonas reinhardtii*



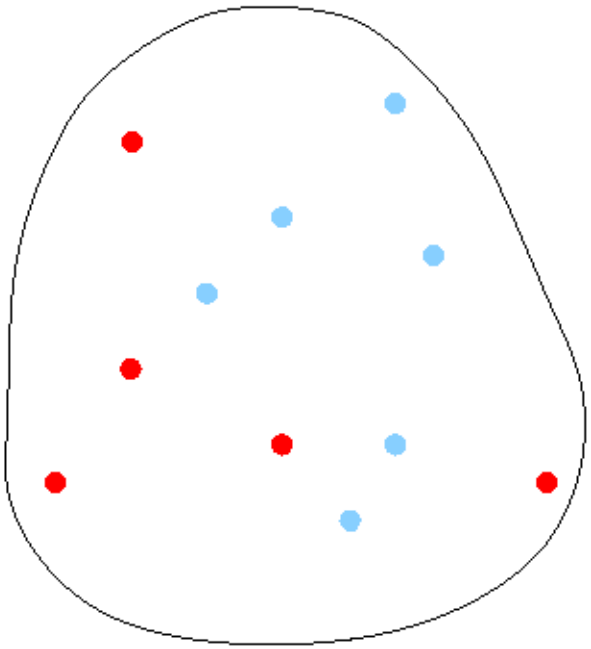
- 159 metabolites
- 3,483 proteins identified by MS
- proteins with no peptide support

MapMan view of Central Metabolism

May et al. *Genetics* 2008

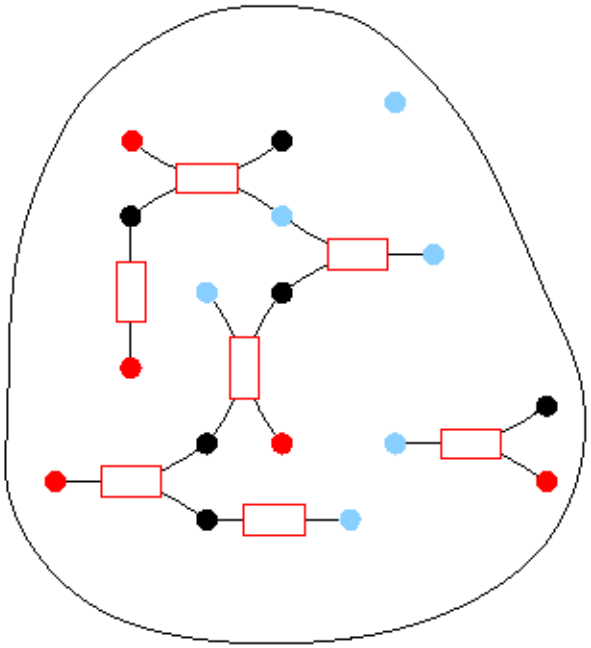
Björn Usadel, Patrick May, Stefanie Wienkoop, Stefan Kempa

Metabolic Expansion Algorithm: From metabolites to genes



- in medium
- detected by GC/MS

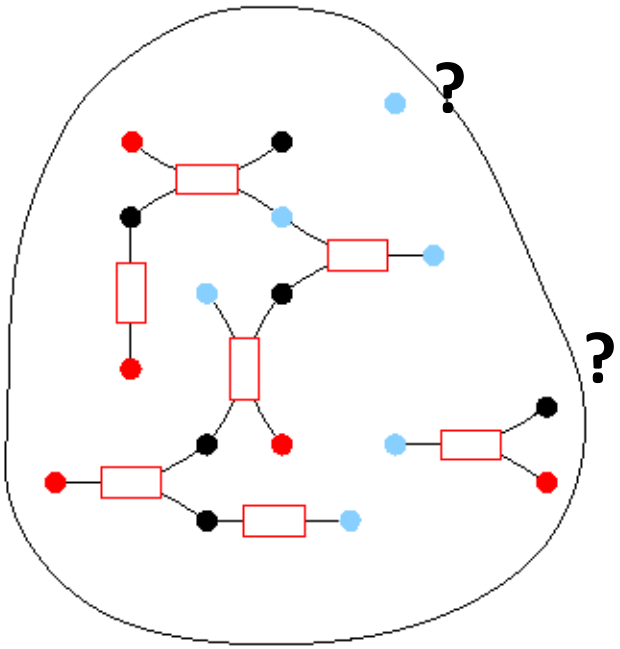
Metabolic Expansion Algorithm: From metabolites to genes



- in medium
- detected by GC/MS
- from KEGG

▭ : annotated in genome

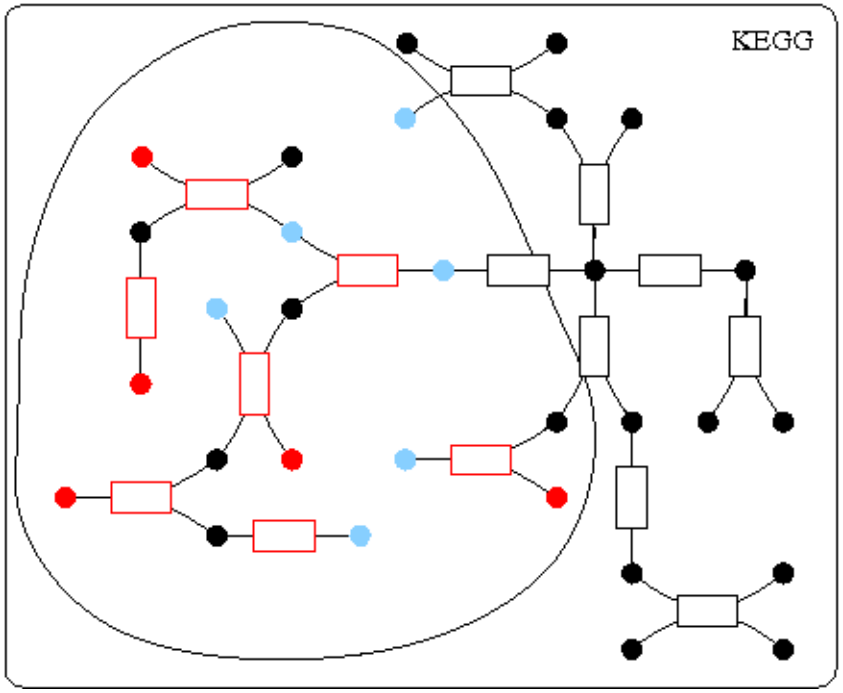
Metabolic Expansion Algorithm: From metabolites to genes



- in medium
- detected by GC/MS
- from KEGG

▭ : annotated in genome

Metabolic Expansion Algorithm: From metabolites to genes

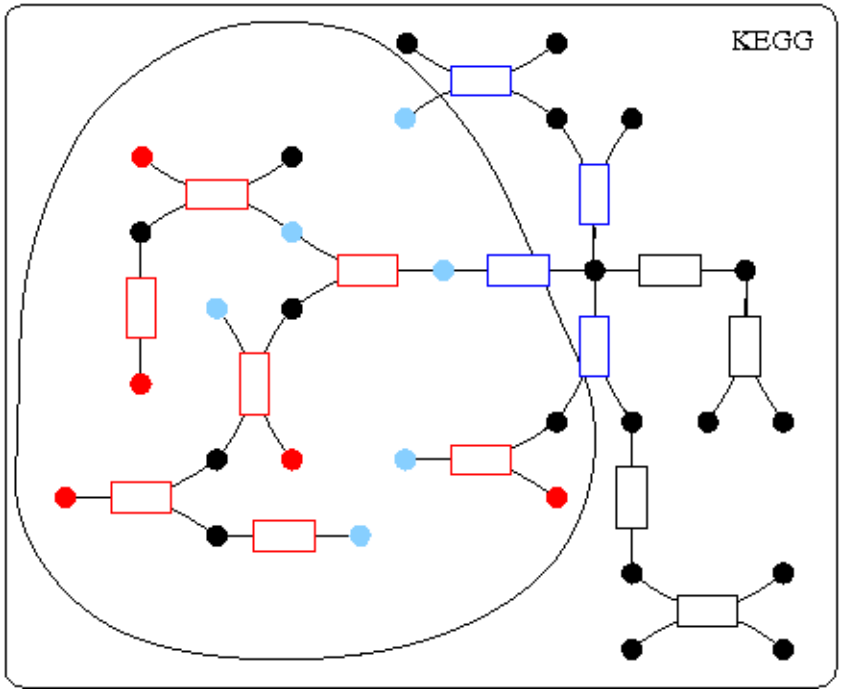


- in medium
- detected by GC/MS
- from KEGG

: annotated in genome

: inferred from KEGG

Metabolic Expansion Algorithm: From metabolites to genes



- in medium
- detected by GC/MS
- from KEGG

: annotation

: inferred from KEGG

: presence predicted = minimal expansion

Nils Christian, Oliver Ebenhöh (MPIMP, U Potsdam)

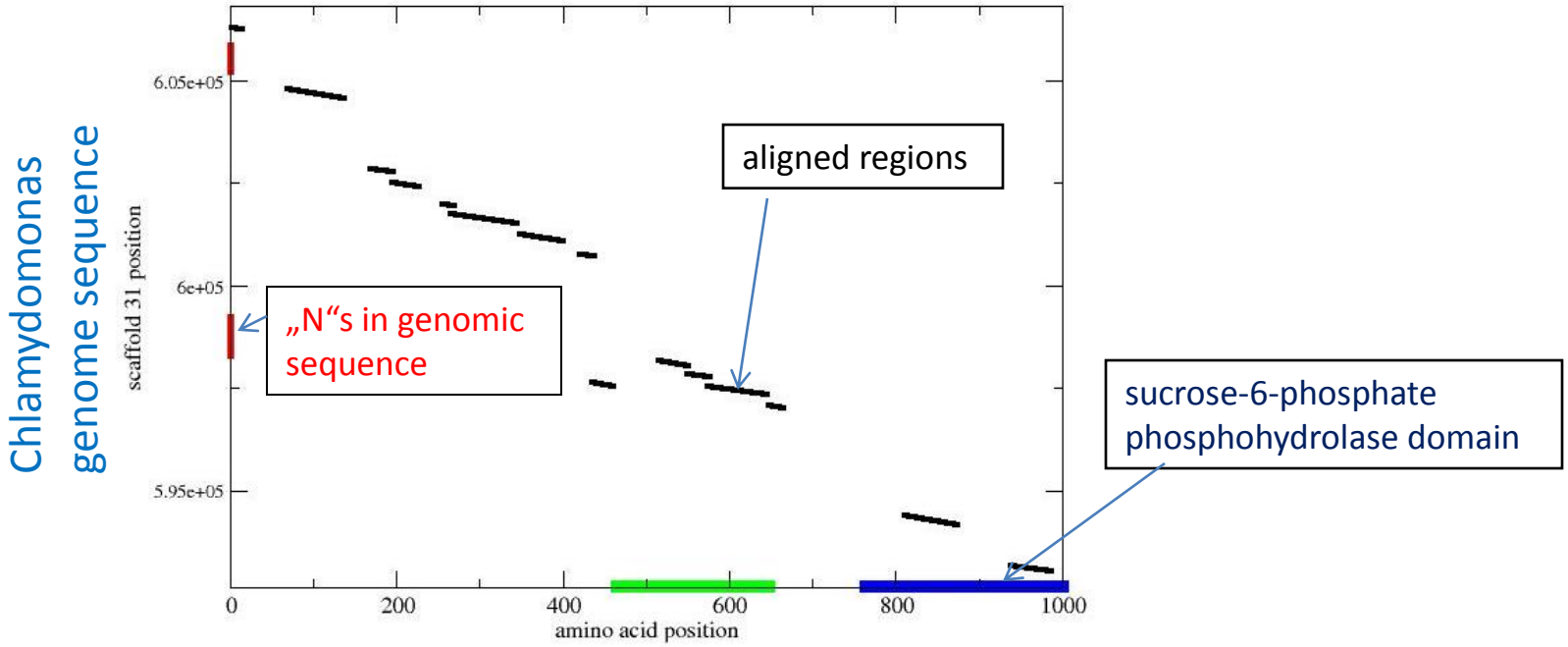
Omics-assisted metabolic pathway analysis in Chlamydomonas reinhardtii

- **Building a Chlamydomonas draft network from KEGG**
  using reciprocal best hit analysis of JGI 3.1 against KEGG
  - 3365 *nc, cp, mt* proteins mapped on
    - 198 KEGG pathways (JGI: 114)
    - 7330 KEGG reactions
    - 713 EC classifications (JGI: 552)

- ***network expansion*** **used to determine which of the 159 metabolites are producible by the draft network**
  - 127 metabolites are represented in KEGG
  - 70 are in the *scope*
  - 57 are **NOT** in the *scope*

Metabolic Expansion Algorithm: From metabolites to genes

- Sucrose detected in Chlamydomonas
- Sucrose pahtway genes including SPS (Sucrose Phosphate Synthase) must be encoded in the genome, but were not annotated in the draft genome
- Bioinformatic analysis of the Chlamydomonas genome identified candidate SPS gene



Arabidopsis SPS sequence

Summary

- Classification tree based automated GC/MS-spectra interpretation is possible
- Metabolite-transcript causal relationships from time course data by using Granger causality
- metabolites appear to be leading and are more pathway-informative
- Metabolomics measurements help complete genome annotations

Acknowledgements

AG Walther (MPIMP)
Jan Hummel
Patrick May
Pawel Durek

AG Kopka (MPIMP)
Katrin Straßburg
Nadine Strehmel
Joachim Kopka

AG Usadel/Stitt (MPIMP)
Björn Usadel

AG Selbig (MPIMP, U Potsdam)
Joachim Selbig
Zoran Nikoloski

AG Weckwerth (GoFORSYS)
Stefanie Wienkoop
Stefan Kempa
Julia Weiss
Luis Recuenco-Munoz
Wolfram Weckwerth (now U Vienna)

AG Ebenhöh (GoFORSYS)
Nils Christian
Oliver Ebenhöh (now U Aberdeeen)