Scoring pathway activity from gene expression data

Jörg Rahnenführer

Computational Biology and Applied Algorithmics
Max Planck Institute for Informatics
D-66123 Saarbrücken
Germany

NGFN - Courses in Practical DNA Microarray Analysis

Berlin, November 25, 2004

Analysis of gene groups

MAX-PLANCK-GESELLSCHAFT

General approach for group testing

- Significance scoring of a biologically defined gene set G (see previous talk on group testing by Ulrich Mansmann)
- *Gene set enrichment:* Sort genes according to some score and analyze positions of gene members of G in this list
- *Global test:* Test for association with a phenotype by summing up sum of (squared) covariances between genes and phenotype

Metabolic pathways

- Metabolic pathways as specific gene sets
- Scoring the statistical significance of *co-regulation* of genes that belong to a pathway
- Mathematical and biological evaluation

Similarity measures for gene pairs

- Every gene is coded by an n-dimensional vector of expression measurements (e.g. time series).
- Consider two genes g_1 and g_2 with measurements $\mathbf{x} = (x_1,\ldots x_n)$ and $\mathbf{y} = (y_1,\ldots y_n)$, respectively.

Definitions of scores for co-expression of genes g_1 and g_2:

$$s_{correlation}(g_1, g_2) = \frac{\langle \mathbf{x} - \bar{x}, \mathbf{y} - \bar{y} \rangle}{||\mathbf{x} - \bar{x}||\,||\mathbf{y} - \bar{y}||} \qquad (1)$$ Joint trend

$$s_{covariance}(g_1, g_2) = \langle \mathbf{x} - \bar{x}, \mathbf{y} - \bar{y} \rangle \qquad (2)$$ Joint trend (scaled with variance)

$$s_{cosine}(g_1, g_2) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{||\mathbf{x}||\,||\mathbf{y}||} \qquad (3)$$ Angle between **x** and **y**

$$s_{dotproduct}(g_1, g_2) = \langle \mathbf{x}, \mathbf{y} \rangle \qquad (4)$$ Angle between **x** and **y** (scaled with variance)

L_2-norm

Scores for co-expression of gene sets

MAX-PLANCK-GESELLSCHAFT

Definition: Score of a gene set:

Choose a fixed score *s* for gene similarity. The score of the gene set

G:={g_1,…,g_k} is given by $$s(G) := \frac{1}{\binom{k}{2}} \sum_{1 \le i_1 < i_2 \le k} s(g_{i_1}, g_{i_2}).$$

This is the average similarity between all pairs of genes in G.

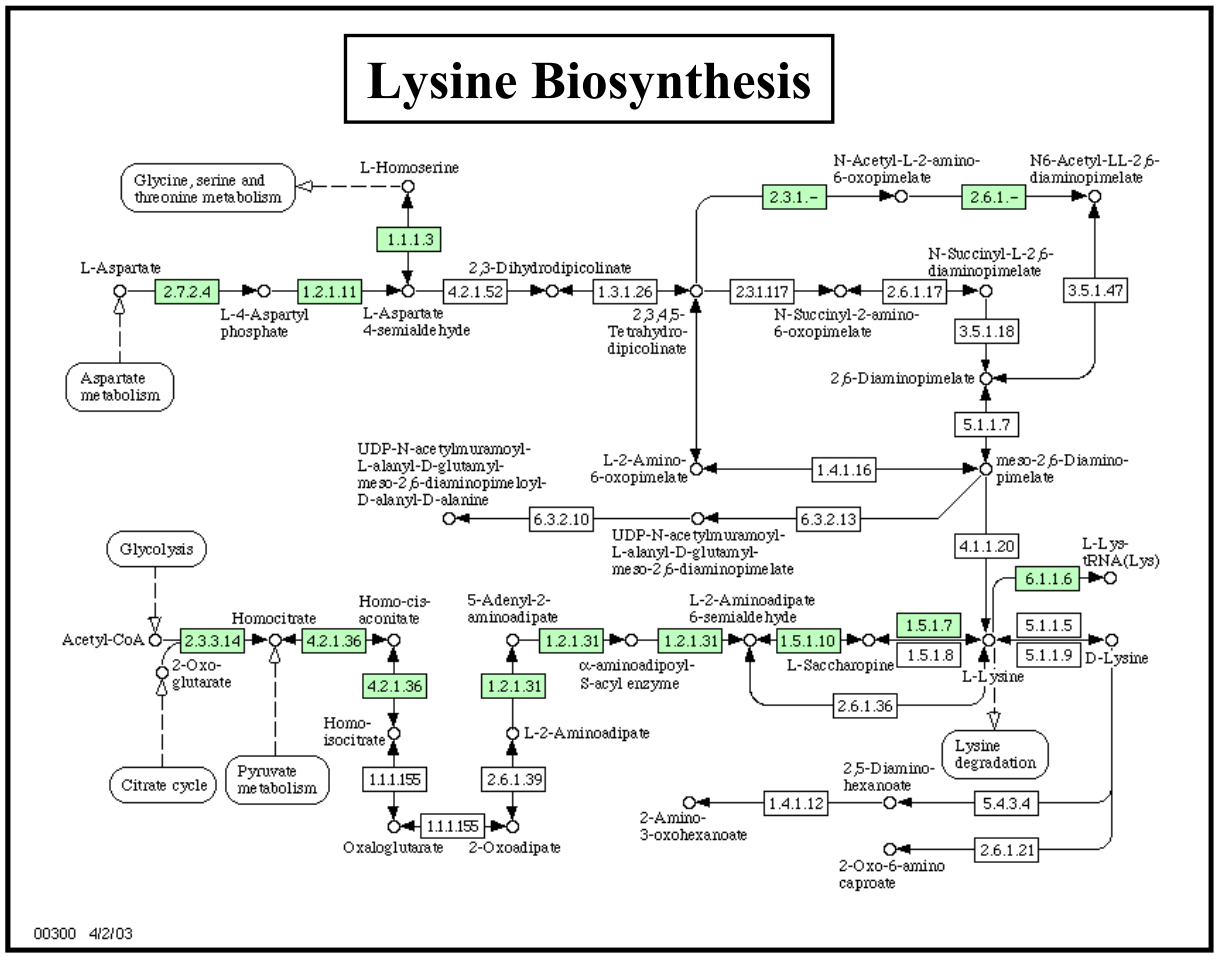
For example, if *s* is the correlation, the the score *s(G)* is the average correlation between two genes in the gene set

Specific gene sets: Metabolic pathways

MAX-PLANCK-GESELLSCHAFT

KEGG pathway database (http://www.genome.ad.jp/kegg/pathway)

Kyoto Encyclopedia of Genes and Genomes



MySQL database

Query implemented in R:

Input: Metabolic pathway.

Output: List of gene groups.

Every group contains gene identifiers matching one enzyme class of the pathway.

Specific gene sets: Metabolic pathways

KEGG pathway database (http://www.genome.ad.jp/kegg/pathway)

Kyoto Encyclopedia of Genes and Genomes

Lysine Biosynthesis

EC (Enzyme class)	Gene identifiers
1.1.1.3	YJR139C
1.2.1.11	YDR158W
1.2.1.31	YBR115C YGL154C
1.5.1.10	YNR050C
1.5.1.7	YIR034C
2.3.3.14	YDL131W YDL182W
2.7.2.4	YER052C
4.2.1.36	YDR234W
6.1.1.6	YDR037W YNL073W

Gene selection algorithms:

Random selection

Select randomly one gene per EC class.

Norm selection

Select gene with maximal Euclidean norm.

Optimal selection

Select combination of genes that maximizes pathway score.

Algorithms for gene selection

- Selection of single gene per enzyme necessary for statistical reasons: Some genes matching the same enzyme are always co-expressed!
- Selection of optimal combination of genes computationally not feasible:

Nr.	Combinations	Name of pathway
1	59719680000	Purine metabolism
2	1244160	Glycolysis / Gluconeogenesis
3	207360	Pyrimidine metabolism
4	41472	Starch and sucrose metabolism
5	6048	Citrate cycle (TCA cycle)
6	4608	Pyruvate metabolism
7	1134	Oxidative phosphorylation
8	1024	Glycine, serine and threonine metabolism
9	640	Pentose phosphate pathway
9	640	Alanine and aspartate metabolism
11	576	Glyoxylate and dicarboxylate metabolism
12	288	Glycerolipid metabolism
13	192	Galactose metabolism
14	144	Phenylalanine, tyrosine and tryptophan biosynthesis
14	144	Carbon fixation
16	72	Fatty acid metabolism

Algorithms for gene selection

- Selection of single gene per enzyme necessary for statistical reasons: Some genes matching the same enzyme are always co-expressed!
- Selection of optimal combination of genes computationally not feasible.

Random gene selection: For every enzyme, select a random gene.

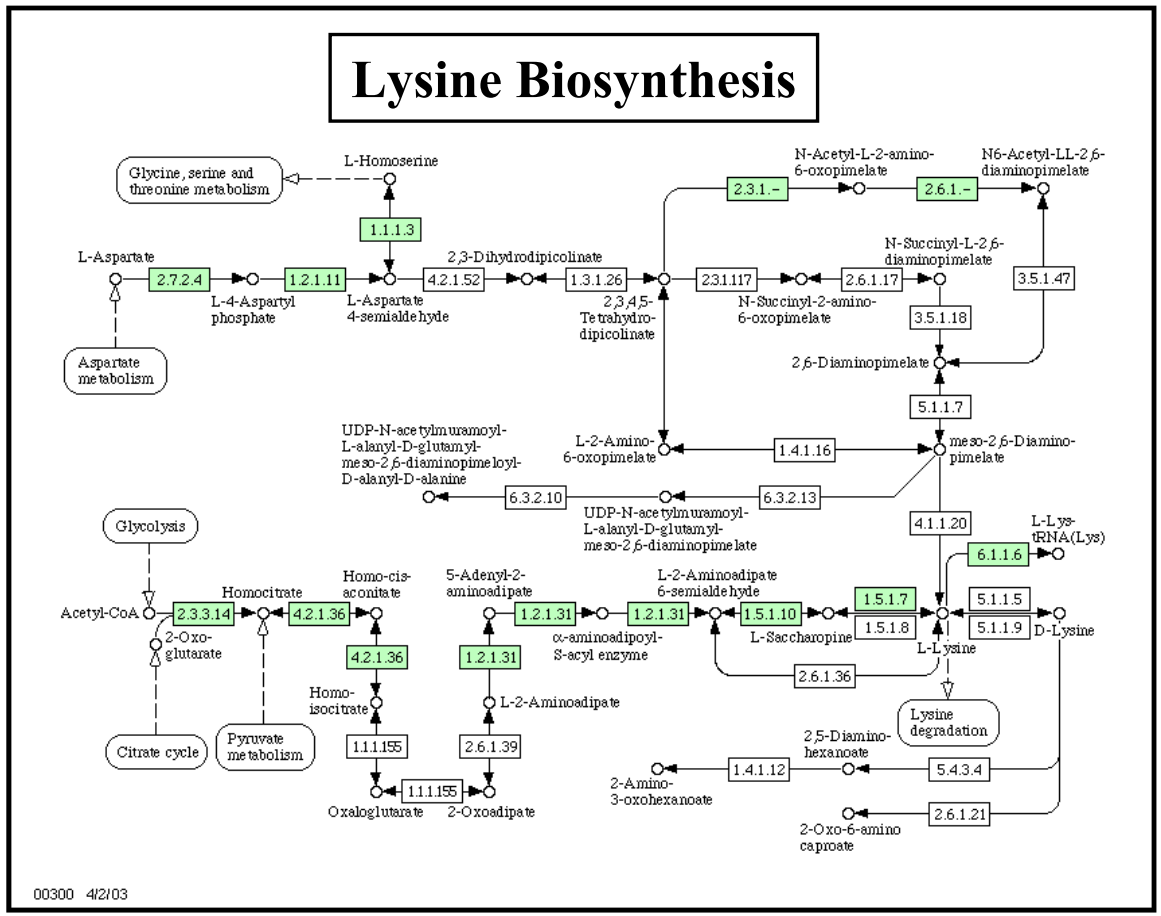
Norm gene selection: For every enzyme, select gene with maximal Euclidean norm of expression profile (maximal differential expression).

Greedy gene selection:

1. For every enzyme: Select gene that maximizes pathway score together with all matching genes of all other enzymes.
2. Iterate through enzymes until convergence: Select gene that maximizes pathway score given all other current gene selections.

Integrating pathway topology

- In the score calculation, introduce weights for gene pairs according to pathway distance



Integrating pathway topology

- In the score calculation, introduce weights for gene pairs according to pathway distance
- Let $d(g_1,g_2)$ denote the distance between two genes g_1 and g_2, i.e. the number of step needed to connect the respective enzymes
- Genes that are closer in pathway distance receive a larger weight in the score calculation
- To avoid singularities (when all genes are from different connected components) the maximal distance is set to a constant (here *10*)
- For similarity measure *s* the score of a gene set then is defined by

$$s(G) \quad := \quad \frac{1}{\binom{k}{2}} \sum_{1 \le i_1 < i_2 \le k} \frac{1}{\min\{d(g_{i_1}, g_{i_2}), 10\}} \, s(g_{i_1}, g_{i_2}).$$

ScorePAGE algorithm

ScorePAGE (**Scor**ing **P**athway **A**ctivity with **G**ene **E**xpression Data)

Input: Gene expression data and metabolic pathway data.

Output: List of active pathways with corresponding p-values.

1. Normalize gene expression data (for every experiment the mean or median expression should be close to 0).
2. Choose similarity measure ***s*** and gene selection algorithm ***sel***.
3. For every pathway ***pw***:
   - Choose $\boldsymbol{G_{sel}}$ according to ***sel*** and calculate pathway score $\boldsymbol{s(G_{sel})}$.
   - Calculate $\boldsymbol{n_{sim}=1000}$ pathway scores for randomly permuted gene identifiers (labels).
   - P-value of *pw:* Fraction of 'random' scores that are larger than $s(G_{sel})$.
4. Adjust vector of all p-values according to false discovery rate.

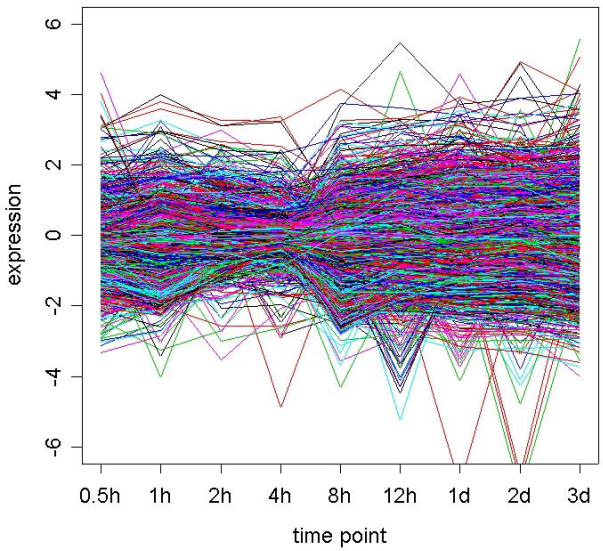
Evaluation: Gene expression data

MAX-PLANCK-GESELLSCHAFT

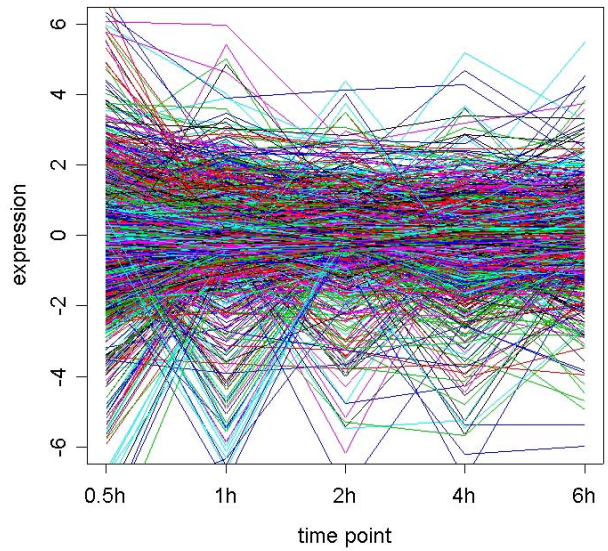
Time series measurements of yeast cell responses to stress:

- Gasch AP et al.: Genomic expression programs in the response of yeast cells to environmental changes, Mol. Biol. Cell. 11(12): 4241-57 (2000).
- Gasch AP et al.: Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p, Mol. Biol. Cell 12(10): 2987-3003 (2001).

Nitrogen depletion data (9 tp)



Amino acid starvation data (5 tp)



Mathematical evaluation

Application of ScorePAGE algorithm to yeast stress data.

Comparison of similarity measures and gene selection algorithms.

Numbers of significant pathways with p<0.05 (fdr-adjusted):

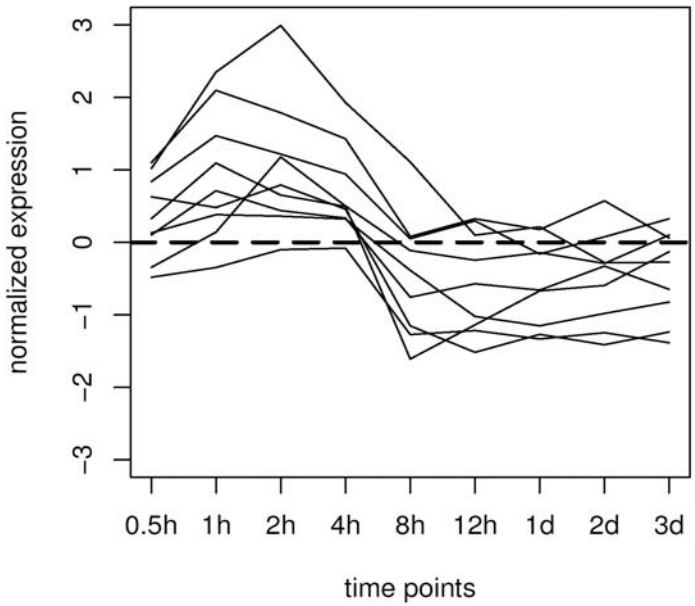
	Random	Norm	Optimal
Covariance	20 0	27 0	28 0
Dot-product	8 8	13 18	8 16
Adaptive	15 3	22 11	20 11

Blue: Nitrogen depletion data Red: Amino acid starvation data

Relevance of similarity measure

Example: Lysine biosynthesis pathway

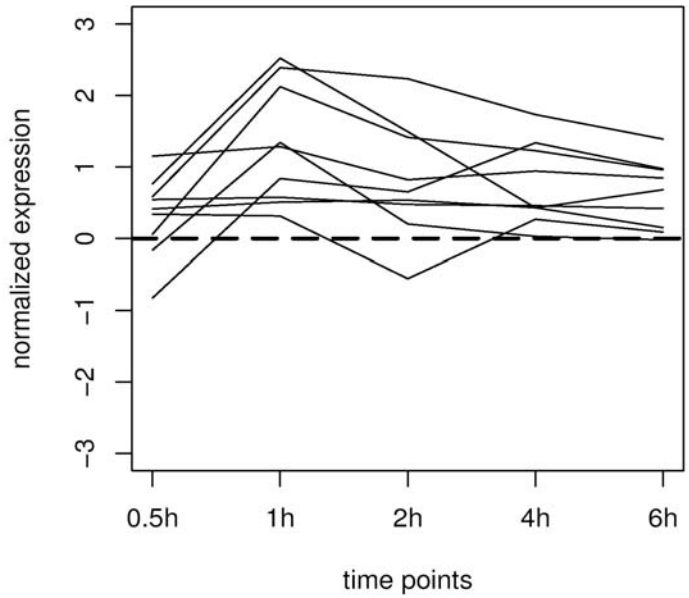
Nitrogen depletion data



Covariance: **Joint trend**

(*Correlation:* without variance*)*

Amino acid starvation data

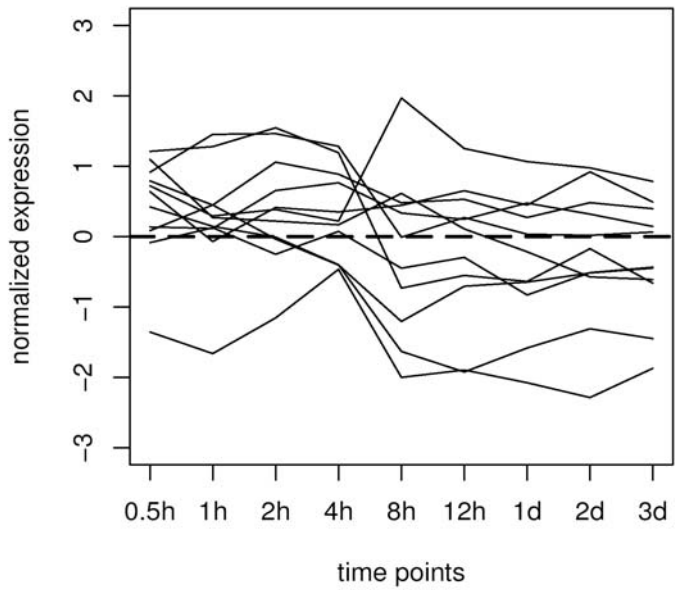


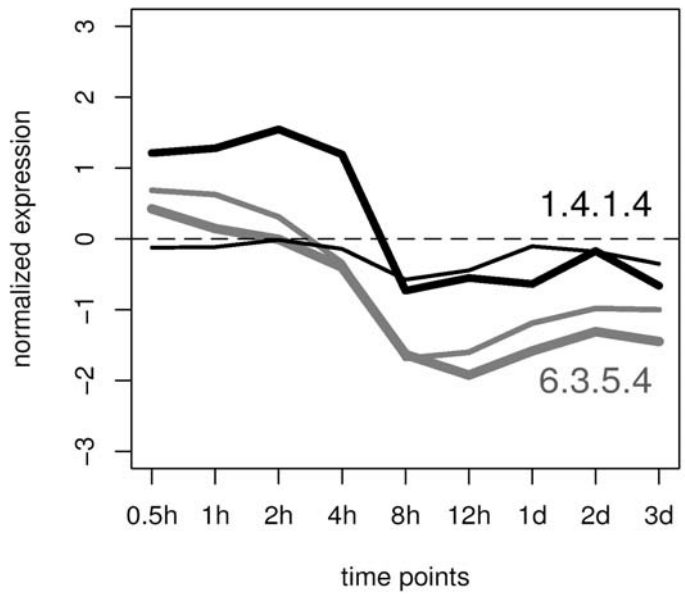
Dot-product: **Joint up-regulation**

(*Cosine*: without variance)

Relevance of gene selection

Nitrogen depletion data, nitrogen metabolism pathway





Pathway only significant with optimal gene selection (p=0.168 ⟶ p=0.022).

EC1.4.1.4 (NADP-Glutamate Dehydrogenase): Matching genes: GDH1, GDH3.

Only GDH3 repressed by glucose (included in media).

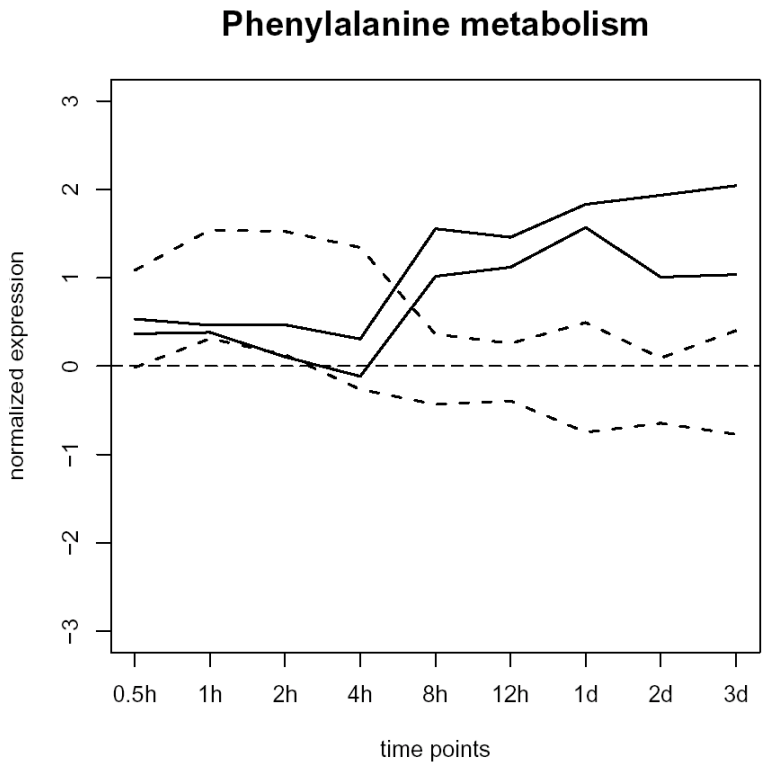
Relevance of pathway topology

Nitrogen depletion data, phenylalanine metabolism pathway

Distance matrix:

Enzyme	1.2.1.5		2.6.1.1		2.6.1.9		3.5.1.4	
1.2.1.5	0.308	(0)	-0.163	(10)	-0.273	(10)	0.363	(1)
2.6.1.1	-0.163	(10)	0.151	(0)	0.193	(1)	-0.250	(10)
2.6.1.9	-0.273	(10)	0.193	(1)	0.336	(0)	-0.388	(10)
3.5.1.4	0.363	(1)	-0.250	(10)	-0.388	(10)	0.520	(0)

- P-value decreases from 0.961 to 0.026 (random gene selection) resp. from 0.996 to 0.011 (greedy gene selection)
- Enzymes 1.2.1.5 and 3.5.1.4 are direct neighbors in the pathway (solid lines), enzymes 2.6.1.1 and 2.6.1.9 are also neighbors (dashed lines)
- No path exists between the two groups
- Distance-based score downweighs negative contributions



Biological evaluation

Nitrogen depletion data

ScorePAGE with covariance score and greedy gene selection

Top 19 pathways (p-value < 0.01) relate to metabolism of amino acids, purine and pyrimidine.

Folate/pantothenate pathways: Cofactors with important role in metab. of nitrogen compounds.

Lowest score for pathway not related to nitrogen metabolism is 0.015.

Nr.	Name of pathway	p-value (fdr)
1	Alanine and aspartate metabolism	< 0.001
1	Aminoacyl-tRNA biosynthesis	< 0.001
1	Aminosugars metabolism	< 0.001
1	Glycine, serine and threonine metabolism	< 0.001
1	Histidine metabolism	< 0.001
1	Lysine biosynthesis	< 0.001
1	Methionine metabolism	< 0.001
1	One carbon pool by folate	< 0.001
1	Purine metabolism	< 0.001
1	Selenoamino acid metabolism	< 0.001
1	Urea cycle and metabolism of amino groups	< 0.001
1	Valine, leucine and isoleucine biosynthesis	< 0.001
1	Sulfur metabolism	< 0.001
14	Lysine degradation	0.005
14	Phenylalanine, tyrosine and tryptophan biosyn.	0.005
14	Pyrimidine metabolism	0.005
17	Cysteine metabolism	0.008
17	Glutamate metabolism	0.008
17	Pantothenate and CoA biosynthesis	0.008
21	Taurine and hypotaurine metabolism	0.017
22	Arginine and proline metabolism	0.022
22	Nitrogen metabolism	0.022
24	Vitamin B6 metabolism	0.030
25	Glyoxylate and dicarboxylate metabolism	0.035
27	Nicotinate and nicotinamide metabolism	0.047
27	Valine, leucine and isoleucine degradation	0.047

Biological evaluation

Nitrogen depletion data

Comparison with competing approach: *Enrichment method*.

Define score for differential expression (here: Euclidean norm) and sort genes according to this score, select all genes with a score above a fixed cutoff and test for enrichment of gene set members within this set.

Here: We optimize cutoff w.r.t. number of significant findings (too optimistic).

Nr.	Name of pathway	p-value (fdr)
1	Alanine and aspartate metabolism	< 0.001
1	Aminoacyl-tRNA biosynthesis	< 0.001
1	Glycine, serine and threonine metabolism	< 0.001
1	Histidine metabolism	< 0.001
1	Lysine biosynthesis	< 0.001
1	One carbon pool by folate	< 0.001
1	Purine metabolism	< 0.001
1	Selenoamino acid metabolism	< 0.001
1	Urea cycle and metabolism of amino groups	< 0.001
1	Valine, leucine and isoleucine biosynthesis	< 0.001
14	Phenylalanine, tyrosine and tryptophan biosyn.	0.005
17	Cysteine metabolism	0.008
17	Glutamate metabolism	0.008

Conclusions

- Including correlations between genes in the score can increase the sensitivity compared with competing enrichment method
- Simple clustering in general not powerful enough to detect pathways (data not shown)
- Use of adaptive similarity measure is feasible
- Gene selection (per enzyme class) improves significance, for selected genes even functional assignment is possible (see also paper of Kharchenko et al., Bioinformatics 2004)
- Integrating pathway topology further improves sensitivity

Literature

- Kharchenko P, Vitkup D, Church GM. Filling gaps in a metabolic network using expression information. *Bioinformatics* 2004; 20 Suppl 1:I178-I185.
- Rahnenführer J, Domingues FS, Maydt J, Lengauer T. Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical Applications in Genetics and Molecular Biology* 2004; 3, No.1, Article 16.

Visit us in Saarbrücken!

Saarvoir vivre…