

An Emergent Self-Organizing Map Based Analysis Pipeline for Comparative Metabolome Studies

Isam Haddad^{a,1}, Karsten Hiller^{b,1,*}, Eliane Frimmersdorf^c, Beatrice Benkert^a,
Dietmar Schomburg^c and Dieter Jahn^a

^a*Technische Universität Braunschweig, Institute of Microbiology, Braunschweig, Germany*

^b*Massachusetts Institute of Technology, Department of Chemical Engineering, Cambridge, MA 02140, USA*

^c*Technische Universität Braunschweig, Department of Bioinformatics and Biochemistry, Braunschweig, Germany*

Edited by H. Michael; received 6 February 2009; revised 23 March 2009; accepted 30 March 2009; published 16 May 2009

ABSTRACT: Modern high-throughput techniques allow for the identification and quantification of hundreds of metabolites of a biological system which cover central parts of the metabolome. Due to the amount and complexity of obtained data there is an increasing need for the development of appropriate computational interpretation methods.

A novel data analysis pipeline designed for high-throughput determined metabolomic data is presented. The combination of principal component analysis (PCA) with emergent self-organizing maps (ESOM) and hierarchical cluster analysis (HCA) algorithms is used to unravel the structure underlying metabolomic data sets, including the detection of outliers. Observed differences between various analyzed metabolomes are automatically mapped and visualized using KEGG metabolic pathway maps. This way typical metabolic biomarker for data sets from various analyzed growth conditions and genetic backgrounds become visible. In order to validate the described methods we analyzed time resolved metabolomic datasets obtained for *Corynebacterium glutamicum* cells grown on various carbon sources consisting of 126 different metabolic patterns.

The analysis pipeline was implemented in the user-friendly Java software eSOMet. The software was successfully used for the clustering of the metabolome data mentioned above. Metabolic biomarkers typical for the utilized carbon sources and analyzed growth phases were identified.

KEYWORDS: Emergent self-organizing maps, cluster analysis, comparative metabolome data analysis

INTRODUCTION

Currently, more than 800 microbial and approximately 100 eukaryotic genome projects have been finished. These genomic data provide the basis for modern high-throughput methods like transcriptomics [Hoheisel, 2006] and proteomics [Patterson and Aebersold, 2003]. Additionally, there are currently strong efforts to establish analogous high-throughput methods for the simultaneous concentration determination of hundreds of metabolites occurring in a living cell. These techniques are known as

¹Both authors contributed equally to this work.

*Corresponding author: Karsten Hiller, Massachusetts Institute of Technology, Department of Chemical Engineering, 77 Massachusetts Ave., 56-439, Cambridge, MA 02140, USA. Tel.: +1 617 2580 349; Fax: +1 617 258 687; E-mail: khiller@mit.edu.

metabolomics [Fiehn *et al.*, 2000]. In contrast to transcriptomics and proteomics, metabolomics techniques focus on the end products of regulatory processes and allow the detection of the final response of the living system to environmental or genetic perturbations. Therefore, high-throughput metabolomics is currently one of the most challenging techniques in the context of systems biology [Oliver *et al.*, 1998; Kopka *et al.*, 2004].

The technical fundamentals are usually provided by gas or liquid chromatography coupled with subsequent mass spectrometry (GC-MS/LC-MS). Depending on the used instrument and the applied analysis method several hundred substances can be detected. Assuming that the detector is not saturated the measured signals are proportional to the metabolite concentrations thus providing the basis for their further statistical analyses [Bunk *et al.*, 2006]. Modern automated GC-MS based methods are able to perform accurate measurements in less than 20 minutes thereby covering main parts of the metabolome [Börner *et al.*, 2007]. Once applied in a high-throughput manner these techniques produce a large amount of high dimensional data.

Typically, the aim of a metabolome analysis is to identify differences or similarities in the metabolic constitutions of an organism between different conditions. These are different environmental set-ups or introduced genetic modifications. For each condition, the quantities of the detectable metabolites are measured. Their values are recorded in a mathematical vector. We call this vector a metabolic pattern. Assuming that each metabolite represents one coordinate of a high-dimensional space, the data vector describes the position of a specific point in this coordinate system. Hence, a data set consisting of an arbitrary number of metabolic patterns, can be expressed as a scatterplot in the high-dimensional metabolome space. Such a data set is called multivariate, because within each dimension a high variability in the values is assumed. This variability results from different parameters [van den Berg *et al.*, 2006]:

Induced biological variation is explained by the differences in the chosen environmental or genetic conditions, which cause a different mode of metabolic operation.

Uninduced biological variation is explained by fluctuations of metabolic concentrations under identical experimental conditions.

Statistical noise is explained by systematic and random errors, due to the experimental setup.

In general one can assume that induced biological variation has a stronger influence on the overall structure of the data than uninduced biological variation, whereas the latter has a much stronger influence than statistical noise. The high dimensionality of the data and different sources of variation render the demand for efficient and accurate methods of mathematical analysis. Appropriate multivariate methods include Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA) and unsupervised artificial neural network based methods such as Self-Organizing Maps (SOM).

PCA performs a linear transformation of the input data into a new coordinate system such that the dimension with the greatest variance is represented by the first coordinate (first principal component), the dimension with the second greatest variance by the second coordinate and so on. Only the first principal components (PCs) are supposed to cover a sufficient amount of variance in order to describe the data. The remaining PCs are assumed to capture only the noise contained in the dataset. Hence, the PCA is used in two ways. It enables a dimensionality reduction of the high dimensional input data with a minimal loss of information. At the same time it is able to detect those metabolites which contribute most significantly to the variance within the dataset. PCA was successfully applied by various groups for the interpretation of metabolomics datasets [Roessner *et al.*, 2001; Fiehn *et al.*, 2002; Askenazi *et al.*, 2003; Hirai *et al.*, 2004; Jonsson *et al.*, 2004].

Another frequently used method for metabolomics data interpretation is HCA [Roessner *et al.*, 2001]. According to the degree of similarity, e.g. the spacial distance of the data points in the high-dimensional

space, the data points are successively combined to clusters and these clusters are grouped to greater groups. Finally, the result of clustering can be visualized using a dendrogram [Eisen *et al.*, 1998; Herrero *et al.*, 2001].

Self-Organizing Maps (SOMs) or Kohonen Feature Maps belong to a class of artificial neural networks [Kohonen, 1982, Kohonen and Mäkisara, 1989], which are capable of projecting high-dimensional input data on a two-dimensional map. The projection is topology preserving, meaning that input patterns, that are located geometrically close to each other in the high-dimensional space, will be mapped to close neurons on the resulting feature map. The map is an arrangement of nodes on a grid, where each node represents initially a metabolic pattern with arbitrary records for each metabolite. During an unsupervised training procedure, each metabolic pattern of the input data is connected to the specific node, which shares the highest similarity to it. Afterwards, this specific node adjusts itself to the assigned pattern. Furthermore, it adjusts the patterns of its neighbouring nodes gradually to its own one, whereas its influence attenuates with the distance to its neighbours. This process of self-organization is done for several training cycles, resulting in a map, that is able to locate any metabolic pattern of the input data with high reliability in its corresponding area on the map.

Meanwhile, the concept of the SOMs was extended in several respects and applied to various bioinformatical problems [Tamayo *et al.*, 1999; Kanaya *et al.*, 2001; Abe *et al.*, 2003; Hirai *et al.*, 2004; Meinicke *et al.*, 2008]. However, most of these implementations provide for a low number of nodes on the output map, which are usually adjusted to the number of expected clusters. Hence, the patterns of the input data distribute over these few clusters, even though they do not support this structure. Furthermore, it is not possible to detect outliers using these strategy. It was shown that application of SOMs with a low number of nodes produces similar results like *k*-means clustering [Ultsch, 1999]. Real emergence, on the other hand, is only expected to occur in SOMs with a number of nodes being at least a magnitude larger, than the number of input patterns. These SOMs are called Emergent Self-Organizing Maps (ESOMs) [Ultsch, 1999]. After the training, most of the output nodes represent either only a few or none input data points, preserving the topology. Finally, the U-Matrix method allows the detection of data cluster by examining the overall structure of the trained ESOM and transformation of this structure into a dendrogram [Ultsch, 1999].

After the elucidation of the structure underlying the experimental metabolomics data either by an ESOM analysis or HCA, detected clusters require further evaluation in order to identify incorrectly assigned data points. The silhouette width representation is a useful tool for this task [Rousseeuw, 1987]. It is based on a comparison of the intra-cluster distances with the inter-cluster distances. This way a quality level for each assigned data point and for each derived cluster is defined. Subsequently, typical data points for a certain cluster and though for a specific environmental or genetic condition became detectable via statistical hypothesis testing like Student's *t*-test or via analysis of variance (ANOVA). Corresponding metabolites conserve as biomarkers for the analyzed condition. In addition, solid knowledge exists about biochemical processes and the genomic organization of cells. The pathways for metabolite interconversion and corresponding macromolecular functions are annotated in databases like the Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa and Goto, 2000], EcoCyc [Karp *et al.*, 2002b], MetaCyc [Karp *et al.*, 2002a] or BRENDA [Schomburg *et al.*, 2002]. This information in combination with the results of described multivariate data analysis and statistical processing allows for the detection of underlying regulatory events.

For several of the required computational tasks necessary for the outlined analyses exist accessible implementations, including Matlab or the R statistical package. Nevertheless, each single interpretation step requires multiple parameter settings in order to make the approach applicable to metabolomics data

analysis. Furthermore, the combination and integration of results from different analysis steps constitutes another problem. To overcome these obvious problems we developed an integrated analysis pipeline for GC-MS based metabolomics data. In our application, PCA was used for an initial analysis of the recorded data, thus, providing an information preserving dimensionality reduction for the efficient application of the subsequent analysis methods. Subsequently, HCA was applied as a fast and efficient clustering method for a recognition of similarities within the metabolome data, providing a solid base for further analysis of the data. Finally, ESOM analyses were used for the accurate deduction of the underlying data structure and hence allowing investigations with high reliability. All described methods and algorithms are integrated in a Java based software package and adjusted to the specific needs of metabolomics data analysis. Additionally, the mapping and visualization of obtained results onto KEGG pathway maps was made feasible. Our novel program is especially directed to the needs of high-throughput metabolomics data analysis hence performing a nearly automatic analysis with only a few manual interventions.

METHODS

Multivariate data matrix

The raw records of the GC-MS analyses were organized in a multivariate data matrix $\mathbf{X} \in Q^{p \times q}$, where p is the number of variables (chemical compounds) and q is the number of observations (e.g. samples, experimental trial). The records of one sample j were denoted by the corresponding column vector of \mathbf{X} : $\mathbf{s}_j = (x_{1,j}, x_{2,j}, \dots, x_{p,j})^T$ and the records for one compound i by the corresponding transposed row vector of \mathbf{X} : $\mathbf{c}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,q})^T$.

Data normalization

The measured metabolome data were semi-quantitative. This means, that the concentrations of a certain metabolite within all samples were proportional to the captured signal strengths. However, as long as no external quantitative calibration of the signals has been performed, the signals themselves were used as input data. This procedure exhibited the one major drawback, that employed data were not equally scaled across various chemical compounds and hence were not directly comparable. Therefore, it was necessary to normalize the data for each compound to the same scale. Furthermore, normalization methods are necessary to adjust fold changes within the data towards an improved interpretability from a biological perspective [van den Berg *et al.*, 2006]. Six different normalizations were implemented and partially applied, depending on the subsequent analysis:

Z-score scaling: lead to the normalized data matrix $\mathbf{N}_Z \in Q^{p \times q}$ with coefficients $n_{i,j} = \frac{x_{i,j} - \bar{x}_i}{s_i}$, x_i as the mean of \mathbf{c}_i and s_i as the standard deviation of \mathbf{c}_i .

Range scaling: lead to the normalized data matrix $\mathbf{N}_R \in Q^{p \times q}$ with coefficients $n_{i,j} = \frac{x_{i,j} - \bar{x}_i}{x_{i \max} - x_{i \min}}$, $x_{i \max}$ as the maximum value of \mathbf{c}_i and $x_{i \min}$ as the minimal value of \mathbf{c}_i .

Pareto scaling: lead to the normalized data matrix $\mathbf{N}_P \in Q^{p \times q}$ with coefficients $n_{i,j} = \frac{x_{i,j} - \bar{x}_i}{\sqrt{S_i}}$.

Vast scaling: lead to the normalized data matrix $\mathbf{N}_V \in Q^{p \times q}$ with coefficients $n_{i,j} = \frac{x_{i,j} - \bar{x}_i}{s_i} \cdot \frac{\bar{x}_i}{s_i}$.

Level scaling: lead to the normalized data matrix $\mathbf{N}_L \in Q^{p \times q}$ with coefficients $n_{i,j} = \frac{x_{i,j} - \bar{x}_i}{\bar{x}_i}$.

Total-signal scaling: scaled the data of one sample to the relative cumulative signal of the sample [Villas-Bôas *et al.*, 2005]. It lead to the normalized data matrix $\mathbf{N}_S \in Q^{p \times q}$ with coefficients

$$n_{i,j} = \frac{x_{i,j}}{\lambda_j}, \text{ with } \lambda_j = \frac{\sum_{i=1}^p x_{i,j}}{\sum_{l=1}^q \sum_{k=1}^p x_{k,l}}.$$

Principal component analysis

PCA aims to find a new basis in a multivariate data space which is meaningful. For this purpose it filters out the noise and reveals underlying structures of the data set. This is achieved by basis vectors, that are orthonormal to each other and still cover the largest variance of the data.

This new basis is represented by the matrix $\mathbf{A} \in Q^{p \times p}$. Each column of \mathbf{A} is one Eigenvector of the corresponding covariance matrix of \mathbf{X} and usually the columns are ordered in descending order to the corresponding Eigenvalues. The projection \mathbf{X}^{PCA} of the normalized data onto the new basis was derived by:

$$\mathbf{X}^{PCA} = \mathbf{A}^T \mathbf{N}_Z$$

For details we refer to [Villas-Bôas *et al.*, 2007].

Hierarchical clustering

HCA is a *bottom-to-top* classification approach, that partitions the samples of the data set into disjoint subsets (or clusters). The goal of this method is to group those samples in one cluster, which are relatively similar and separate those samples in different clusters which are relatively dissimilar. As a measure of similarity of two samples \mathbf{s}_k and \mathbf{s}_l , the distance of the vectors was chosen:

$$d_{k,l} = \|\mathbf{s}_l - \mathbf{s}_k\|^n$$

where $\|\cdot\|^n$, $n \in [1, 2, \dots, \infty]$ denotes an n -norm.

For two disjoint clusters α and β containing the sample vectors $(\mathbf{s}_t, \dots, \mathbf{s}_u) \in \alpha$ and $(\mathbf{s}_v, \dots, \mathbf{s}_w) \in \beta$ a linkage method was deployed, which defined the distance of the two clusters. Our implementations considered three different linkage methods.

- (1) The single linkage was defined as the smallest distance of any two samples of each cluster: $d(\alpha, \beta) = \min (d_{i,j})$ with $\mathbf{s}_i \in \alpha$ and $\mathbf{s}_j \in \beta$.
- (2) The complete linkage defined the distance as the largest distance of any two samples of each cluster: $d(\alpha, \beta)$ again with $\mathbf{s}_i \in \alpha$ and $\mathbf{s}_j \in \beta$.
- (3) The average linkage defined the average of the distances of all pairs of samples from each cluster: $d(\alpha, \beta) = \frac{1}{|\alpha| \cdot |\beta|} \sum_{\mathbf{s}_i \in \alpha, \mathbf{s}_j \in \beta} d_{i,j}$.

For the general clustering algorithm, it is referred to [Villas-Bôas *et al.*, 2007] again.

ESOMs

The employed ESOM maps the sample vectors \mathbf{s}_j , also referred to as input vectors, onto a two dimensional grid \mathbf{M} of $X \times Y$ neurons ($M_{x,y}$) also known as feature map. To avoid unfavorable border effects, the feature map was embedded into a finite but borderless torus space. In order to fulfill the requirement of the ESOMs and allow true emergence, the number of neurons was chosen between two and three magnitudes larger than the number of sample vectors. All nodes of the input layer were connected to all nodes of the output layer by the weight vectors (or feature vectors) $\mathbf{w}_{x,y} \in Q^p$.

During the unsupervised training procedure of SOMs, the feature vectors were adjusted in order to reflect the underlying structure of the high dimensional training data in a lower dimensional space. Finally, the SOM mapped those input vectors that were closely located to each other in the input space to neighboring output nodes of the feature map. Although the core algorithm is described elsewhere [Kohonen, 1982], we briefly summarize it here for better understanding.

1. Z -score normalization \mathbf{N}_Z of the original data.
2. Random initialization of all feature vectors $\mathbf{w}_{x,y} \in Q^p$.
3. For each training cycle $t = 1, 2, \dots, t_{\max}$:
 - (a) Random selection of one normalized sample vector \mathbf{s}_j from \mathbf{N}_Z .
 - (b) Identification of the node (winning node) $M_{x,y}$ whose weight vector $\mathbf{w}_{x,y}$ had the closest Euclidean distance to the selected sample vector:

$$\min(\|\mathbf{s}_j - \mathbf{w}_{x,y}\|^2) \forall M_{k,l} \in \mathbf{M}$$

- (c) Adjustment of the weight vector $\mathbf{w}_{x,y}$ according to:

$$\mathbf{w}_{x,y}(t+1) = \mathbf{w}_{x,y}(t) + \tau(t)[\mathbf{s}_j - \mathbf{w}_{x,y}]$$

- (d) where τ was the training rate that depended on the initial training rate $\alpha \in [0 \dots 1]$ and decreased linear with the number of performed training cycles to zero for t_{\max} :

$$\tau(t) = \alpha \cdot \left(1 - \frac{t-1}{t_{\max}}\right)$$

- (e) In order to preserve the topology on the p -dimensional input space, the weight vectors of the neighboring nodes $M_{x+\Delta x, y+\Delta y}$ were adjusted as follows:

$$\mathbf{w}_{x+\Delta x, y+\Delta y}(t+1) = \mathbf{w}_{x+\Delta x, y+\Delta y}(t) + \tau(t) \cdot \eta(r, t)[\mathbf{s}_j - \mathbf{w}_{x+\Delta x, y+\Delta y}(t)]$$

- (f) $\eta(r, t)$ is a gaussian function that depended on the Euclidean distance $r = \sqrt{\Delta X^2 + \Delta y^2}$ of the output nodes $M_{x,y}$ and $M_{x+\Delta x, y+\Delta y}$, v as initial training radius and $\sigma(t) = \tau(t) \cdot v \cdot \frac{1}{2} \cdot \sqrt{X^2 + Y^2}$ as training cycle dependent training radius:

$$\eta(t, r) = e^{-\frac{1}{2} \left(\frac{r}{\sigma(t)}\right)^2}$$

- (g) The greater the Euclidean distance between the winning and neighboring node the smaller is the influence of the training on the corresponding feature vector.

4. Steps 3a–3d were repeated iteratively for all sample vectors \mathbf{s}_j .

Due to the high number of neurons used for the ESOM construction, the metabolomics input vectors were sparsely distributed over the feature map. In order to detect structures in the feature map of the trained ESOM, a U-Matrix was calculated [Ultsch, 1999]. The U-Matrix is an expansion of the two-dimensional map on a three-dimensional landscape. For this purpose, each node $M_{x,y}$ is assigned by scalar value $u_{x,y}$ representing the ‘height’ of each node in the U-Matrix. It is calculated as the sum of the Euclidean distances between the corresponding weight vector $\mathbf{w}_{x,y}$ and the weight vectors of $M_{x,y}$ immediate four neighbors:

$$u_{x,y} = \|\mathbf{w}_{x,y} - \mathbf{w}_{x+1, y+1}\|^2 + \|\mathbf{w}_{x,y} - \mathbf{w}_{x-1, y+1}\|^2 \\ + \|\mathbf{w}_{x,y} - \mathbf{w}_{x+1, y-1}\|^2 + \|\mathbf{w}_{x,y} - \mathbf{w}_{x-1, y-1}\|^2$$

Visualization of the U-Matrix was accomplished by representing the values of $u_{x,y}$ as color codes (Fig. 1). The U-Matrix divides the feature map into several sections: Input vectors closely related in the high dimensional input space were located in ‘valleys’. In contrast, unrelated input vectors were divided by ‘hills’ in the U-Matrix visualization. This demonstrates the actual power of the ESOMs. Nodes that were frequently selected as winning node during the training, had a strong adjusting influence on their near neighborhood. The accumulated Euclidian distance is in these cases low for these nodes, resulting in the emergence of the valleys. Contrary, nodes that are only subject to weak adjustment by several far-placed winning nodes, cause the emergence of ‘hills’.

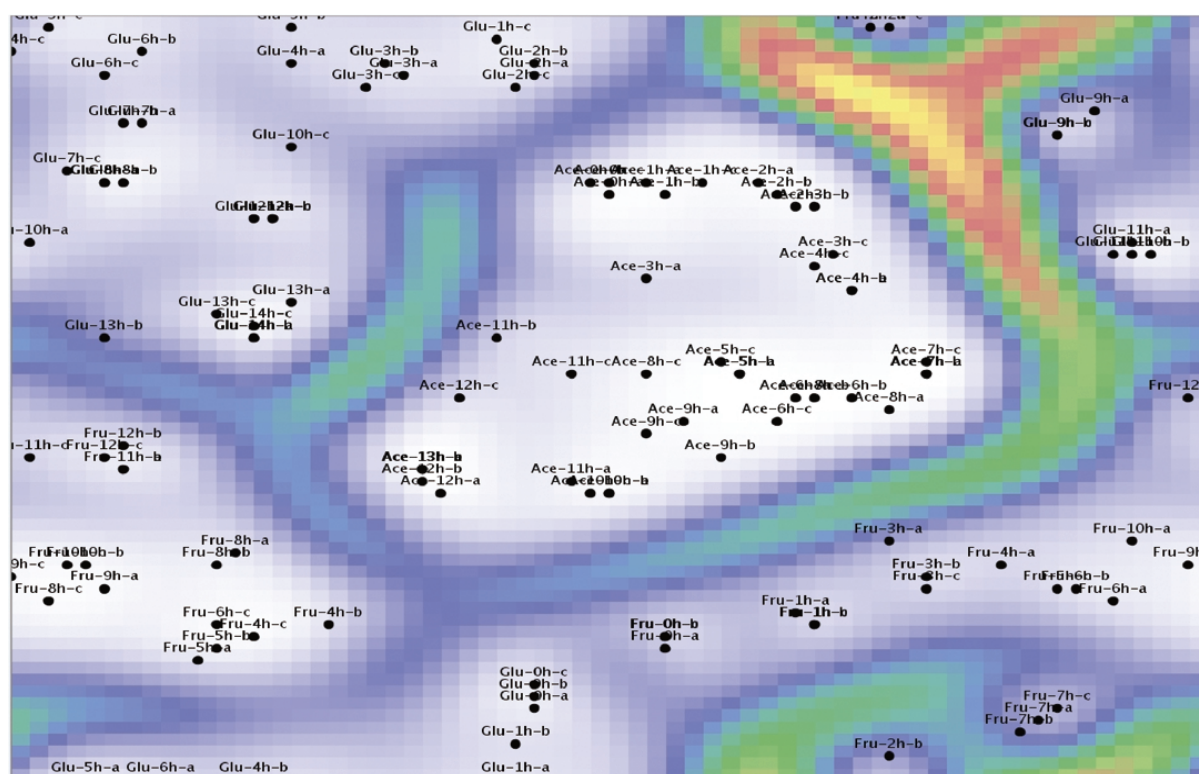


Fig. 1. A two-dimensional representation of the U-Matrix trained with the metabolome data from three different fermentations of *C. glutamicum*. The color code represents the accumulated Euclidean distances between each output node of the U-Matrix to its four immediate neighbours and can be interpreted as a three-dimensional landscape. The input vectors of the metabolic fingerprints are embedded in this landscape. Those of similar patterns are located in *valleys* (white – blue regions), whereas dissimilar patterns are separated by *hills* (green – yellow regions). The labels have the following format: <substrate> – <hours after inoculation> – <replicate>. As substrates Ace stands for acetate, Fru for fructose and Glu for glucose. The U-Matrix is embedded in a torus space, meaning that the right side of the map is continued on the left and the bottom side at the top.

Dendrograms

The results of the hierarchical clustering as well as the applied ESOMs were visualized in dendrograms. In case of the HCA, the levels of the branch points were determined by the distance of two clusters according to the chosen linkage method. In case of the ESOMs we raised a virtual ‘water level’ inside the U-Matrix. Always when two ‘valleys’ were connected by the raised water level and both valleys were flooded we connected two branches in the emerging dendrogram at the according height.

Silhouette width

The quality of the determined cluster was measured by the calculation of the silhouette width [Rousseeuw, 1987] for each sample and the average silhouette width for each cluster. The silhouette width is based on a comparison of the intra- with the inter-cluster distances. For the distance calculation we used Euclidean distance. The silhouettes were calculated by the following formula:

$$S(j) = \frac{b(j) - a(j)}{\max(a(j), b(j))}$$

$S(j)$ was the silhouette width for sample j located in a chosen cluster α , $b(j)$ was the average Euclidean distance of \mathbf{s}_j to all samples located in cluster β and $a(j)$ was the average Euclidean distance of \mathbf{s}_j to all samples located in α . It is obvious that $S(j) \in [-1;1]$. A silhouette width of 1 means that $a(j)$ is very small compared to $b(j)$ and therefore sample j was assigned to the correct cluster. If the silhouette width is -1 , $b(j)$ was very small compared to $a(j)$ and consequently sample j was misclassified. However, for $S(j) \approx 0$ no clear assignment of sample j either to α or β could be made.

Statistical tests

Metabolites whose concentration were statistical relevant changed among the different clusters, were detected by the application of Student's t -tests [Press *et al.*, 1992] on the normalized data. Based on the estimated p -values metabolites with p -values lower than 0.05 were selected as significantly changed metabolites. In combination with the pathway information provided by the KEGG database [Kanehisa and Goto, 2000] a contingency table for each defined pathway was set up. Based on this contingency table a Fisher's exact test was used to calculate the probability whether this observation was obtained by chance.

Experimental protocol for metabolome analysis

The *Corynebacterium glutamicum* wild type strain ATCC 13032 was used for batch cultivations in a bioreactor (Minifors, Infors GmbH, Einsbach, Germany) with an operating volume of 3 l at 30°C. Bacteria were grown in minimal medium containing either glucose (20 g/l), fructose (20 g/l) or acetate (16.4 g/l) as sole carbon source. Additionally, 1 l medium consists of 5 g of $(\text{NH}_4)_2\text{SO}_4$, 5 g urea, 2 g KH_2PO_4 , 2 g K_2HPO_4 , 3 H_2O , 0.25 g $\text{MgSO}_4 \cdot 7 \text{H}_2\text{O}$, 10 mg CaCl_2 , 0.2 mg biotin, 28.5 mg $\text{FeSO}_4 \cdot 7 \text{H}_2\text{O}$, 16.5 mg $\text{MnSO}_4 \cdot 1 \text{H}_2\text{O}$, 6.4 mg $\text{ZnSO}_4 \cdot 7 \text{H}_2\text{O}$, 0.764 mg $\text{CuSO}_4 \cdot 5 \text{H}_2\text{O}$, 0.128 mg $\text{CoCl}_2 \cdot 6 \text{H}_2\text{O}$, 0.044 mg $\text{NiCl}_2 \cdot 6 \text{H}_2\text{O}$, 0.064 mg $\text{Na}_2\text{MO}_4 \cdot 2 \text{H}_2\text{O}$, 0.048 mg H_3BO_3 , 0.05 mg $\text{SrCl}_2 \cdot 6 \text{H}_2\text{O}$, 0.05 mg $\text{BaCl}_2 \cdot 2 \text{H}_2\text{O}$, 0.028 mg $\text{KAl}(\text{SO}_4)_2 \cdot 12 \text{H}_2\text{O}$. The pH was adjusted to 7.0 with 5 M KOH.

Approximately $5 \cdot 10^{10}$ cells were hourly sampled in triplicate. Sampled cells were pelleted by centrifugation ($3940 \times g$, 4°C, 3 min) and washed twice with 20 ml of a NaCl solution (0.9%). Finally, the cells were resuspended in 1.5 ml methanol containing ribitol (40 μl ribitol/ml methanol) as internal standard. The cells were disrupted by ultrasonification (15 min, 70°C). Subsequently, lysed cells were incubated on ice (2 min), 1.5 ml H_2O were added and every sample was thoroughly mixed for 30 sec. After addition of 1 ml chloroform samples were mixed again and centrifuged ($3940 \times g$, 4°C, 6 min) for separation of hydrophilic and hydrophobic phases. 1 ml of the polar phase was transferred to a new reaction tube and dried.

The dried samples were solved in 20 μl pyridine containing methoxyamine hydrochloride (20 mg/ml) and incubated at 180 rpm and 30°C for 90 min. After adding 32 μl N-methyl-N-tri-methylsilyltrifluoroacetamid (Chromatographie Service, Langerwehe, Germany) samples were incubated again at 180 rpm and 37°C for 30 min and for another 2 h at 20°C. Finally, 4 μl of a mix containing decane, dodecane, pentadecane, nonadecane, docosane, octacosane, dotriacontane and hexatriacontane (20 mg/ml in cyclohexane each) were added in order to perform a calculation of retention indices [Vandendool and Kratz, 1963]. 2 μl of the samples were injected in a gas chromatograph coupled with a mass spectrometer (Thermo Quest, Thermo Fisher Scientific GmbH, Dreieich, Germany). All parameters for sample injection, gas chromatography, mass spectrometry, metabolite identification and quantification were described before [Strelkov *et al.*, 2004].

RESULTS AND DISCUSSION

Combination of principal component analysis (PCA), hierarchical cluster analysis (HCA) and emergent self organizing maps (ESOMs) for metabolome analysis

For interpretation of complex metabolome data we developed a new data analysis strategy and implemented it in the novel Java based software eSOMet. A typical application of this tool comprises four steps:

1. The multivariate data matrix, describing the quantities of metabolite signals of the different experimental trials is imported. The tool allows for the investigation of typical statistical measures and application of a PCA in order to get a first impression of the data quality.
2. Afterwards an optional data preprocessing can be performed. At least a Z -score transformation of the data of interest should be applied, in order to comply with the requirements of the subsequent steps. For even more reliable data processing, it is recommended to reduce the data on a PCA projected subspace. In this context, previous studies proposed the application of data reducing steps with the sole consideration of the metabolites with highly variable concentrations, the so called descriptors [Kouskoumvekaki *et al.*, 2008]. However, with our approach it is neither necessary nor convenient to discard low deviating metabolites, since it does not necessarily require the separation of true variation and measurement noise. The PCA driven reduction allows for the intensification of the influence of the highly varying metabolite concentrations on the clustering and the consideration of low deviating metabolite concentrations.
3. Subsequently, an unsupervised clustering of the preprocessed data deploying either HCA or ESOMs follows. As a result of both methods a dendrogram describing the hierarchical relations between the different metabolic patterns is generated.
4. Distinct branches from the dendrogram can be chosen and the changes of metabolite concentrations between the selected classes as well as the level of significance are calculated. Finally, the results are mapped onto KEGG metabolic pathway maps allowing for an intuitive interpretation.

Several bioinformatics tools exist which combine statistical methods with an visualization using biological pathway maps. Examples for gene expression profiles are PathMAPA [Pan *et al.*, 2003], PathwayExplorer [Mlecnik *et al.*, 2005] or Pathway Processor [Grosu *et al.*, 2002] or for metabolome profiles MetNet [Wurtele *et al.*, 2003], KaPPA-View [Tokimatsu *et al.*, 2005] or MAPMAN [Thimm *et al.*, 2004]. However, all these tools have in common that they require already processed data and the statistical analysis is restricted to the identification of significantly affected pathways. The tool VANTED [Junker *et al.*, 2006] has been developed to fill this gap of missing analytical methods and provides several statistical procedures for a subsequent data analysis. However, employed procedures do not allow for the clustering of the metabolome data obtained for different conditions. Instead, the clustering of metabolites or genes according to there common type of regulation is possible. Clearly, in our case the consequences of employed environmental conditions on the changes between the corresponding metabolomes is in the focus of the investigation.

The necessary data interpretation can be achieved by the application of clustering methods such as HCA or ESOMs, which detect hierarchical relations between the sample patterns. In contrast to partitioning cluster procedures as k -means or ordinary SOMs, no particular knowledge about the processed data is required. It is not even necessary to define an initial number of clusters. Another important drawback of the partitioning cluster methods like k -means and ordinary SOMs is the influence of outliers which may affect the detected arrangement of the data. In contrast, hierarchical approaches, e.g. HCA or ESOMs,

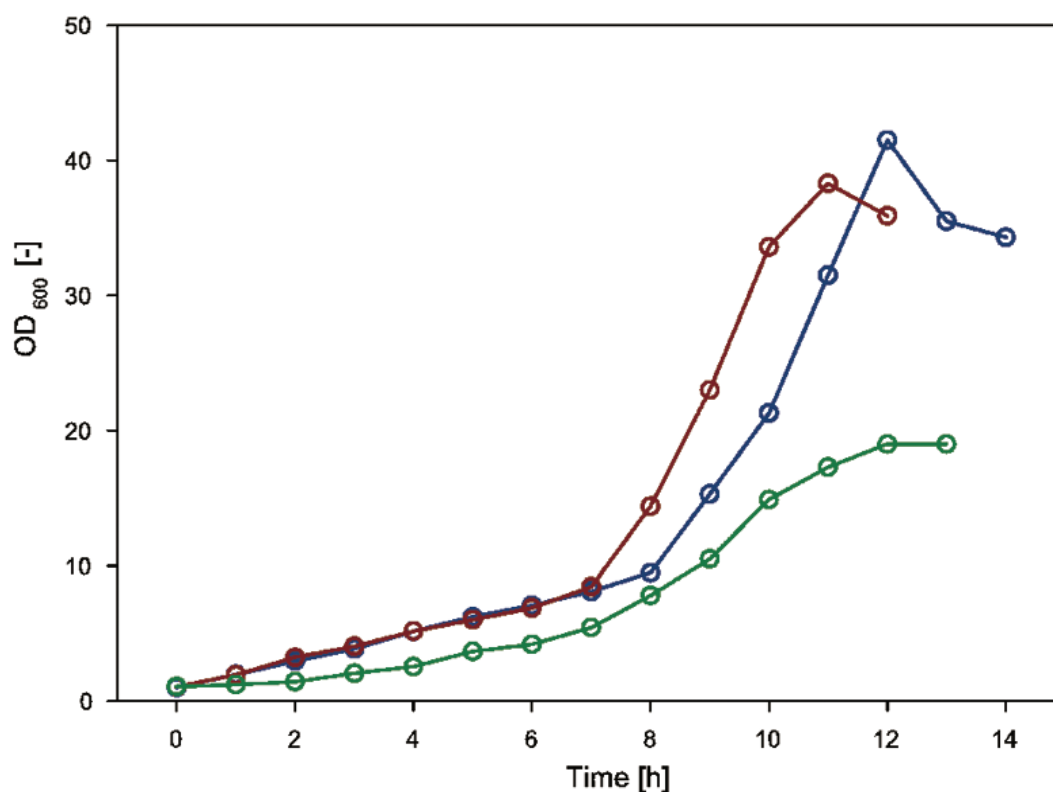


Fig. 2. Growth curves of the *C. glutamicum* fermentations with the three different substrates glucose (blue), fructose (red) and acetate (green).

do not force the data into functionally not supported classifications. They rather allow to unravel the high resolved topology of the underlying structure. Beyond these aspects, these methods offer the possibility to combine metabolic fingerprint data of already evaluated experiments with novel data sets. Hence, gaining detailed insights into unknown data based on previous analyses results becomes possible.

Analysis of C. glutamicum metabolome data using eSOMet

In order to demonstrate the power of the described clustering methods for metabolome interpretation we analyzed an experimental timeresolved metabolomic data set. For this purpose, growth experiments of *C. glutamicum* using three different carbon sources were performed (Fig. 2). Samples were taken hourly over a time period of 12 h–14 h and the signals for 209 metabolites were measured in triplicates yielding a total of 126 metabolomes. These data sets were highly valuable for this study since they comprise two variations, namely different carbon sources and different growth phases. Thus, an evaluation of the classification performance of a HCA and ESOM was possible at a low (carbon source) as well as a high resolved (growth phase) level. For this purpose the experimental data were processed as outlined above.

The evaluation of the obtained results was carried out in the following way: Firstly, the detected coarse-grained structures of the data sets were compared. For this purpose the largest cluster containing only samples grown on one particular carbon source each were selected. Both methods were able to clearly distinguish the metabolic patterns according to the utilized carbon sources (Fig. 3a, 3b). Additionally, the generated dendrograms confirmed the closer relation between the samples of the glucose and fructose

Table 1

Metabolites of significantly changed concentrations (p -value < 0.05) measured with a positive fold-change ratio for growth conditions utilizing one carbon source compared against the other two carbon sources

Glucose	Fructose	Acetate
Mannose	Oxalate	Glutamate
D-Xylulose-5-phosphate	D-Glucono-1,5-lactone	β -Alanine
Ribulose-5-phosphate	D-Mannitol	Succinate
Dihydroxyacetone phosphate	L-Leucine	Pyrrol-2-carboxylate
Glucose-6-phosphate	2-Oxoglutarate	Proline

fermentations compared to the samples using acetate as single carbon source. This result is in accordance to established biochemistry, since the glucose and fructose metabolism differs only in the sugar uptake and initial step of utilization. However, the major pathway of hexose breakdown via glycolysis is identical. In contrast the acetate catabolism operates different to a large extent [Wendisch *et al.*, 2000]. In order to measure the accuracy of the classification, the average silhouette width for each of the three clusters was calculated. Both methods were equally able to detect some samples of the glucose and fructose fermentation as outliers. Consequently, these samples were assigned outside the selected branches. However, the average silhouette widths for the glucose and fructose cluster clearly differed between the two methods (HCA: 0.05 and 0.25, ESOMs: 0.22 and 0.35, Fig. 3a, 3b). A closer investigation revealed that some samples that were classified as part of the main branches by the ESOM, were assigned as stronger related to the outlying patterns by the HCA, yielding to silhouette widths of less accuracy. Based on this observation it became obvious that HCA has partly misclassified these metabolomes.

During the next evaluation step the revealed fine-grained structures of both methods (HCA and ESOMs) were compared. For this purpose, those branches of the dendrograms were selected, which classified for a specific growth phase of a particular carbon source. The early and intermediate exponential growth phases were chosen as appropriate conditions (Fig. 3c, 3d). In the case of the glucose and acetate fermentations HCA distinguishes growth phases different compared to ESOMs. Although ESOMs in contrast to HCA failed to discriminate samples of different growth phases for the fructose fermentation, the calculated silhouette widths (HCA: -0.11, 0.22, 0.14, 0.21, 0.42, 0.34; ESOM: 0.36, 0.29, 0.39, 0.40, 0.47) provide the evidence that the ESOM classification is preferable. Summarizing these interpretations, ESOMs tend to generate cluster of a higher reliability than HCA.

All analyses are provided as detailed online material on the project's website.

Biomarker identification for carbon source utilization in C. glutamicum

The last step of the here described analysis pipeline was the identification of metabolites with significant cellular concentration changes between the three utilized carbon sources. For this purpose we statistically analyzed the intermediate growth phases of the three different fermentations (Fig. 3c, 3d). The following fold-change ratios were calculated: glucose/fructose, fructose/acetate and acetate/glucose. To measure the statistical significance of the ratios, Student's t -tests were performed on the data. For each carbon source, we exemplarily chose five metabolites, having a significant, positive fold-change ratio against the other two carbon sources (Table 1).

Additionally, the implemented eSOMet analysis allowed for the identification of the parts of the cell metabolism which exhibit significant differences in the metabolic profiles between the selected conditions. For this purpose we implemented functionalities to integrate data from the KEGG Ligand database as well as the KEGG pathway maps automatically. The software automatically downloads all required data and sets up a fully functional SQL-database. Combining these information with the calculated p -values

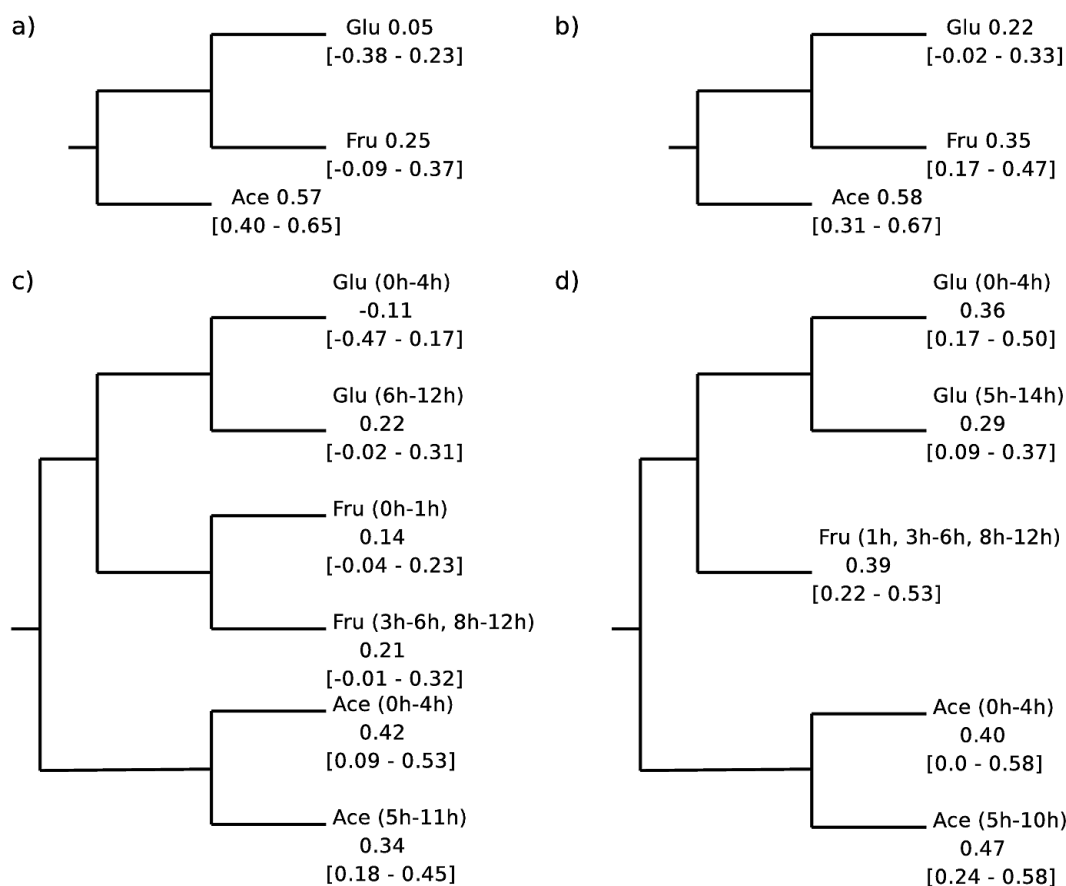


Fig. 3. A schematic representation of the dendrograms of metabolic fingerprints from *C. glutamicum*. Experimental metabolome data for the utilization of glucose (Glu), fructose (Fru) and acetate (Ace) by *C. glutamicum* were analyzed by the HCA (a) and ESOM-based (b) clustering methods. Analogously, corresponding early and intermediate growth phase experiment with these carbon sources were analyzed by HCA (c) and ESOM (d). The average silhouette widths as well as the maximum and minimum silhouette widths (squared brackets) were calculated as quality criterion of the clustering results. Metabolic fingerprints outlying of these main branches are not depicted. HCA was applied using average linkage and Euclidean distance. ESOMs were deployed on a lattice with 64×64 nodes, setting the initial training rate to 0.5 and the initial training radius to 0.6. The map was trained by 100 cycles. For both approaches the data underwent a preceding projection on a principal component space, covering 90% of the total variance. This way, the dimensionality of the input space could be reduced from 209 to 19 dimensions.

and fold-change ratios of the metabolites, the KEGG maps can be visualized accordingly, thus making an intuitive interpretation of the results feasible. Figure 4 presents such a visualization of the pentose phosphate pathway displaying the comparison of the intermediate growth phases between glucose and fructose fermentation. The discovered changes of the metabolite concentrations were due to the different mechanisms of sugar uptake for glucose and fructose. Fructose was imported by the fructose phosphotransferase system and successively converted into fructose-1-phosphate and fructose-1,6-bisphosphate. The lack of fructose-1,6-bisphosphatase activity in *C. glutamicum* under this cultivation conditions prevented the formation of fructose-6-phosphate [Eggeling and Bott, 2005]. Hence, it was observed that the fructose-6-phosphate concentration is approximately twice as high as for the glucose fermentation. Furthermore, the lack of fructose-1,6-bisphosphatase activity resulted in reduced fluxes through the pen-

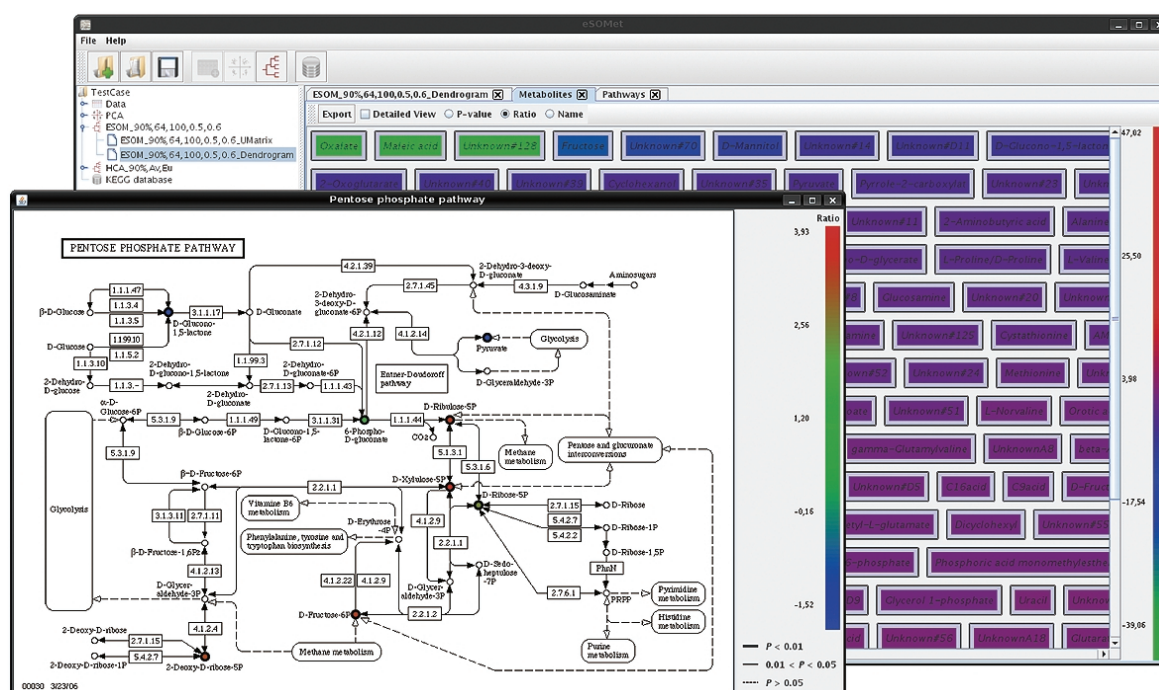


Fig. 4. A screenshot of the eSOMet user interface. The window in the front visualizes as an example the KEGG pentose phosphate pathway. The fold-changes as well as the depending significance levels of measured metabolite concentrations comparing the growth on glucose and fructose were automatically added by eSOMet. The other window shows the main eSOMet window, displaying all measured metabolites. The color code describes the fold-change ratios, the line width the calculated p -values.

tose phosphate pathway. This interpretation agreed nicely with the calculated concentration ratios for metabolites like xylulose-5-phosphate, ribose-5-phosphate or ribulose-5-phosphate, which were two to three folds higher in case of glucose utilization.

CONCLUSIONS

A novel nearly automatic analysis pipeline for the analysis of high-throughput based metabolomics data was introduced. For the first time algorithms based on emergent self organizing maps were applied for the analysis of a metabolic dataset of *C. glutamicum*. The results obtained by these methods were compared to the results of a traditional HCA of the same data. Although the overall structure of the data detected by HCA and ESOMs was similar, obvious differences in the determined fine-grained structures were observable. A well known drawback of HCA is the choice of an arbitrary linkage method. In contrast, ESOMs are very flexible since the preceding training process enables an adaption to the analyzed data. Further on, the training procedure of the ESOMs makes it useful for the detection of hidden underlying structures contained in the metabolomics data.

In the presented case-study, it was demonstrated that the deployed methods were able to confirm the different modes of metabolic activity of *C. glutamicum*. In this context, especially the visualization of the results as KEGG pathway maps turned out to be beneficial.

ACKNOWLEDGEMENTS

This work was funded by the German Federal Ministry of Education and Research (BMBF), Grant No. 031U110A/031U210A and Grant No. 0313980D.

We want to give sincere thanks to Dr. Jörn Pons-Kühnemann and Dr. Gabriel Schachtel (University Gießen) for fruitful discussions concerning the statistical foundations of this work.

APPENDIX

Software

Software name

eSOMet

Project home page

<http://esomet.tu-bs.de>

Operating system

Platform independent, tested on Linux 2.6_32bit, Windows XP_32bit, Mac OS-X 10.5

Other requirements

Java 1.6 or higher

License

GNU-GPL

Any restrictions to use by non-academics

For certain analytical steps of the software, you have to agree to the conditions of the KEGG license agreement.

Additional files

The following files are available from the project home page.

Raw metabolome data

A *.csv file, containing the multivariate metabolome data of the experiment. Column headers describe the experimental trials. The first row contains metabolite names, the second row corresponding KEGG identifier. The other fields contain the recorded signals.

Dendrograms

A *.zip file, containing 4 exported dendrograms of the test case analysis in *.png format.

Test case project file

A binary *.eso file exported from a session in eSOMet, containing all in this study described analytical steps. It can directly be loaded into eSOMet and makes all results interactively available.

REFERENCES

- Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. and Ikemura, T. (2003). Informatics for unveiling hidden genome signatures. *Genome Res.* **13**, 693-702.
- Askenazi, M., Driggers, E. M., Holtzman, D. A., Norman, T. C., Iverson, S., Zimmer, D. P., Boers, M.-E., Blomquist, P. R., Martinez, E. J., Monreal, A. W., Feibelman, T. P., Mayorga, M. E., Maxon, M. E., Sykes, K., Tobin, J. V., Cordero, E., Salama, S. R., Trueheart, J., Royer, J. C. and Madden, K. T. (2003). Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nat. Biotechnol.* **21**, 150-156.

- Bunk, B., Kucklick, M., Jonas, R., Münch, R., Schobert, M., Jahn, D. and Hiller, K. (2006). Metaquant: a tool for the automatic quantification of GC/MS-based metabolome data. *Bioinformatics* **22**, 2962-2965.
- Börner, J., Buchinger, S. and Schomburg, D. (2007). A high-throughput method for microbial metabolome analysis using gas chromatography/mass spectrometry. *Anal. Biochem.* **367**, 143-151.
- Eggeling, L. and Bott, M. (2005). *Handbook of Corynebacterium glutamicum*. CRC Press, Boca Raton, USA.
- Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863-14868.
- Fiehn, O. (2002). Metabolomics-the link between genotypes and phenotypes. *Plant Mol. Biol.* **48**, 155-171.
- Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R. N. and Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **18**, 1157-1161.
- Grosu, P., Townsend, J. P., Hartl, D. L. and Cavalieri, D. (2002). Pathway processor: a tool for integrating whole-genome expression results into metabolic networks. *Genome Res.* **12**, 1121-1126.
- Herrero, J., Valencia, A. and Dopazo, J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics* **17**, 126-136.
- Hirai, M. Y., Yano, M., Goodenowe, D. B., Kanaya, S., Kimura, T., Awazuhara, M., Arita, M., Fujiwara, T. and Saito, K. (2004). Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **101**, 10205-10210.
- Hoheisel, J. D. (2006). Microarray technology: beyond transcript profiling and genotype analysis. *Nat. Rev. Genet.* **7**, 200-210.
- Jonsson, P., Gullberg, J., Nordström, A., Kusano, M., Kowalczyk, M., Sjöström, M. and Moritz, T. (2004). A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. *Anal. Chem.* **76**, 1738-1745.
- Junker, B. H., Klukas, C. and Schreiber, F. (2006). Vanted: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics* **7**, 109.
- Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y., Yamada, Y., Nishi, T., Mori, H. and Ikemura, T. (2001). Analysis of codon usage diversity of bacterial genes with a self-organizing map (som): characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome. *Gene* **276**, 89-99.
- Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27-30.
- Karp, P. D., Riley, M., Paley, S. M. and Pellegrini-Toole, A. (2002a). The MetaCyc database. *Nucleic Acids Res.* **30**, 59-61.
- Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., Pellegrini-Toole, A., Bonavides, C. and Gama-Castro, S. (2002b). The EcoCyc database. *Nucleic Acids Res.* **30**, 56-58.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59-69.
- Kohonen, T. and Mäkisara, K. (1989). The self-organizing feature maps. *Phys. Scripta* **39**, 168-172.
- Kopka, J., Fernie, A., Weckwerth, W., Gibon, Y. and Stitt, M. (2004). Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.* **5**, 109.
- Kouskoumvekaki, I., Yang, Z., Jónsdóttir, S. O., Olsson, L. and Panagiotou, G. (2008). Identification of biomarkers for genotyping aspergilli using non-linear methods for clustering and classification. *BMC Bioinformatics* **9**, 59.
- Meinicke, P., Lingner, T., Kaever, A., Feussner, K., Göbel, C., Feussner, I., Karlovsky, P. and Morgenstern, B. (2008). Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps. *Algorithms Mol. Biol.* **3**, 9.
- Mlecnik, B., Scheideler, M., Hackl, H., Hartler, J., Sanchez-Cabo, F. and Trajanoski, Z. (2005). Pathwayexplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res.* **33** (Web Server issue), W633-W637.
- Oliver, S. G., Winson, M. K., Kell, D. B. and Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends Biotechnol.* **16**, 373-378.
- Pan, D., Sun, N., Cheung, K.-H., Guan, Z., Ma, L., Holford, M., Deng, X. and Zhao, H. (2003). Pathmapa: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for *Arabidopsis*. *BMC Bioinformatics* **4**, 56.
- Patterson, S. D. and Aebersold, R. H. (2003). Proteomics: the first decade and beyond. *Nat. Genet.* **33 Suppl**, 311-323.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press.
- Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L. and Fernie, A. R. (2001). Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13**, 11-29.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53-65.
- Schomburg, I., Chang, A. and Schomburg, D. (2002). Brenda, enzyme data and metabolic information. *Nucleic Acids Res.* **30**, 47-49.

- Strelkov, S., von Elstermann, M. and Schomburg, D. (2004). Comprehensive analysis of metabolites in *Corynebacterium glutamicum* by gas chromatography/mass spectrometry. *Biol. Chem.* **385**, 853-861.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907-2912.
- Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L. A., Rhee, S. Y. and Stitt, M. (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **37**, 914-939.
- Tokimatsu, T., Sakurai, N., Suzuki, H., Ohta, H., Nishitani, K., Koyama, T., Umezawa, T., Misawa, N., Saito, K. and Shibata, D. (2005). KaPPA-view: a web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps. *Plant Physiol.* **138**, 1289-1300.
- Ultsch, A. (1999). Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series. *In: Kohonen Maps*, Oja, E and Kaski, S. (eds.), Elsevier, pp. 33-45.
- van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K. and van der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **7**, 142.
- Vandendool, H. and Kratz, P. D. (1963). A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography. *J. Chromatogr.* **11**, 463-471.
- Villas-Bôas, S. G., Moxley, J. F., Akesson, M., Stephanopoulos, G. and Nielsen, J. (2005). High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. *Biochem. J.* **388**, 669-677.
- Villas-Bôas, S. G., Nielsen, J., Smedsgaard, J., Hansen, M. E. and Roessner-Tunali, U. (2007). *Metabolome Analysis: An Introduction*. Wiley-Interscience.
- Wendisch, V. F., de Graaf, A. A., Sahm, H. and Eikmanns, B. J. (2000). Quantitative determination of metabolic fluxes during cointilization of two carbon sources: comparative analyses with *Corynebacterium glutamicum* during growth on acetate and/or glucose. *J. Bacteriol.* **182**, 3088-3096.
- Wurtele, E., Li, J., Diao, L., Zhang, H., Foster, C., Fatland, B., Dickerson, J., Brown, A., Cox, Z., Cook, D., Lee, E.-K. and Hofmann, H. (2003). Metnet: software to build and model the biogenetic lattice of Arabidopsis. *Comp. Funct. Genom.* **4**, 239-245.