

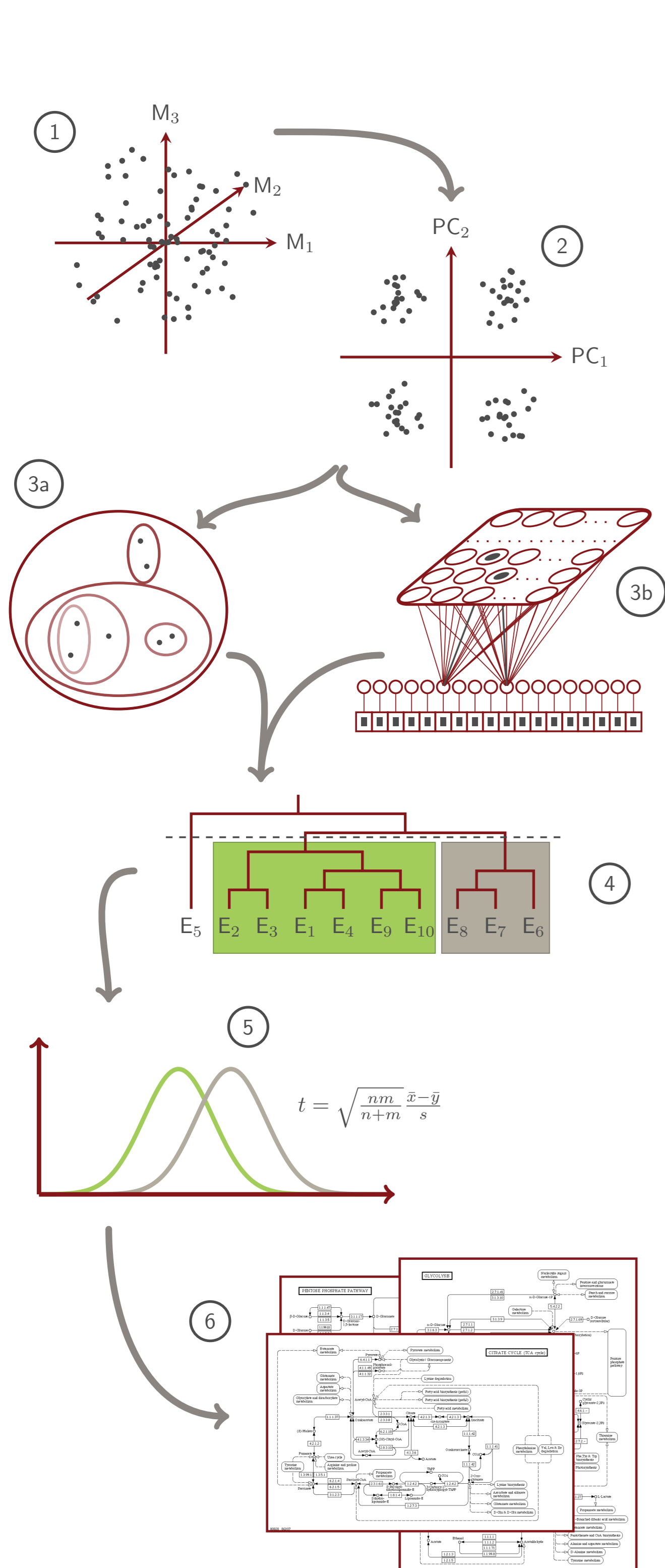
Abstract

Objective: Modern high-throughput techniques like GC-MS facilitate the identification and quantification of hundreds of metabolites of a biological system, thus covering large parts of the metabolome. Due to the amount and complexity of obtained data, there is an increasing demand for the development of appropriate computational analysis methods. We present a novel analysis pipeline specifically designed for high-throughput based metabolomics data, which enables the detection of hierarchical relationships within different metabolic patterns measured under various conditions.

Results: eSOMet is a new software that implements established algorithms like hierarchical cluster analysis (HCA) along with modern methods like emergent self-organizing maps (ESOM). This makes it highly reliable for the purpose of deducing underlying relationships within a series of metabolomics data. The functionality of the tool covers the ability of biomarker discovery, detection of statistical outliers and the automatic mapping of detected metabolic differences onto KEGG metabolic pathway maps. In order to validate the described methods we analyzed a metabolomics time-series data set containing in total 126 metabolic patterns of *Corynebacterium glutamicum* cells grown on different carbon sources. Although the hierarchical overall structure of the metabolic patterns was similarly detected by HCA and ESOMs, obvious differences concerning the highly resolved relations were observable and investigated.

Conclusions: The developed analysis pipeline could be rendered as a valuable tool for the reliable detection of hidden structures within GC-MS based metabolome data. Especially the deployment of emergent self-organizing maps is an indispensable extension to the spectrum of metabolome data analysis methods. eSOMet is freely available as a Java Web Start application at <http://esomet.tu-bs.de>.

Analysis pipeline



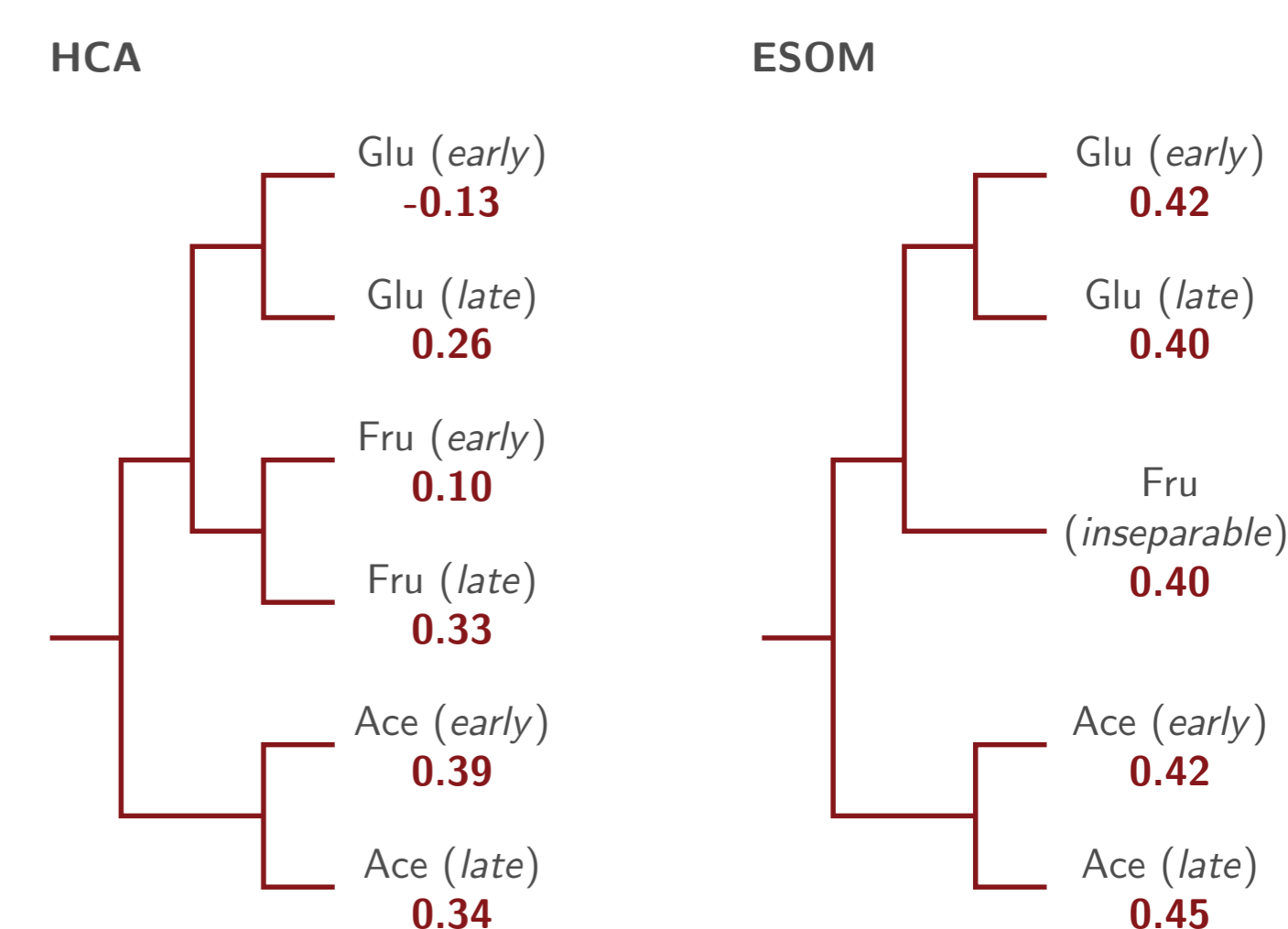
- Acquisition of the multivariate GC-MS data. The data set is organized such that the metabolite concentrations are described as statistical variables, where each experimental trial denotes one statistical object.
- Normalization and reduction of the data set by application of a principle component analysis. The PCA rearranges the data in the multidimensional space such that the first dimensions represent the largest variances within the data set, whereas the dimensions of higher order cover the noise only.
- Clustering of the experimental trials according to their similarity, using either hierarchical cluster analysis (HCA) or emergent self-organizing maps (ESOMs).
- In both cases, a dendrogram representing the relationships of the metabolic patterns of each experimental trial is calculated.
- Different branches from the dendrograms representing distinct environmental conditions can manually be selected and the metabolic changes statistically evaluated.
- Finally, the fold change ratios of the metabolite concentration changes and their statistical significance are displayed on the KEGG metabolic pathway maps (Kanehisa & Goto 2000) using color codes, hence, presenting the results in a more biological fashion.

Evaluation of the results

In order to evaluate the performance of HCA and ESOMs, hourly sampled metabolome data of three different fermentations of *C. glutamicum* (grown on glucose, acetate or fructose) were acquired. The data were clustered using both approaches and their accuracy measured by the silhouette width (Rousseeuw 1987).

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

Depicted on the left side are the main branches of the calculated dendrograms. *early* and *late* denote the point in time during the exponential growth phase when the samples were taken. The green numerical values are the average silhouette widths for all samples comprised by the according cluster. Values close to 1 denote that the samples were correctly clustered to this cluster. Contrary, values close to -1 denote that the samples are more similar to those in a sibling cluster, and thus, were clustered wrong.



ESOMs tend to produce results of higher accuracy.

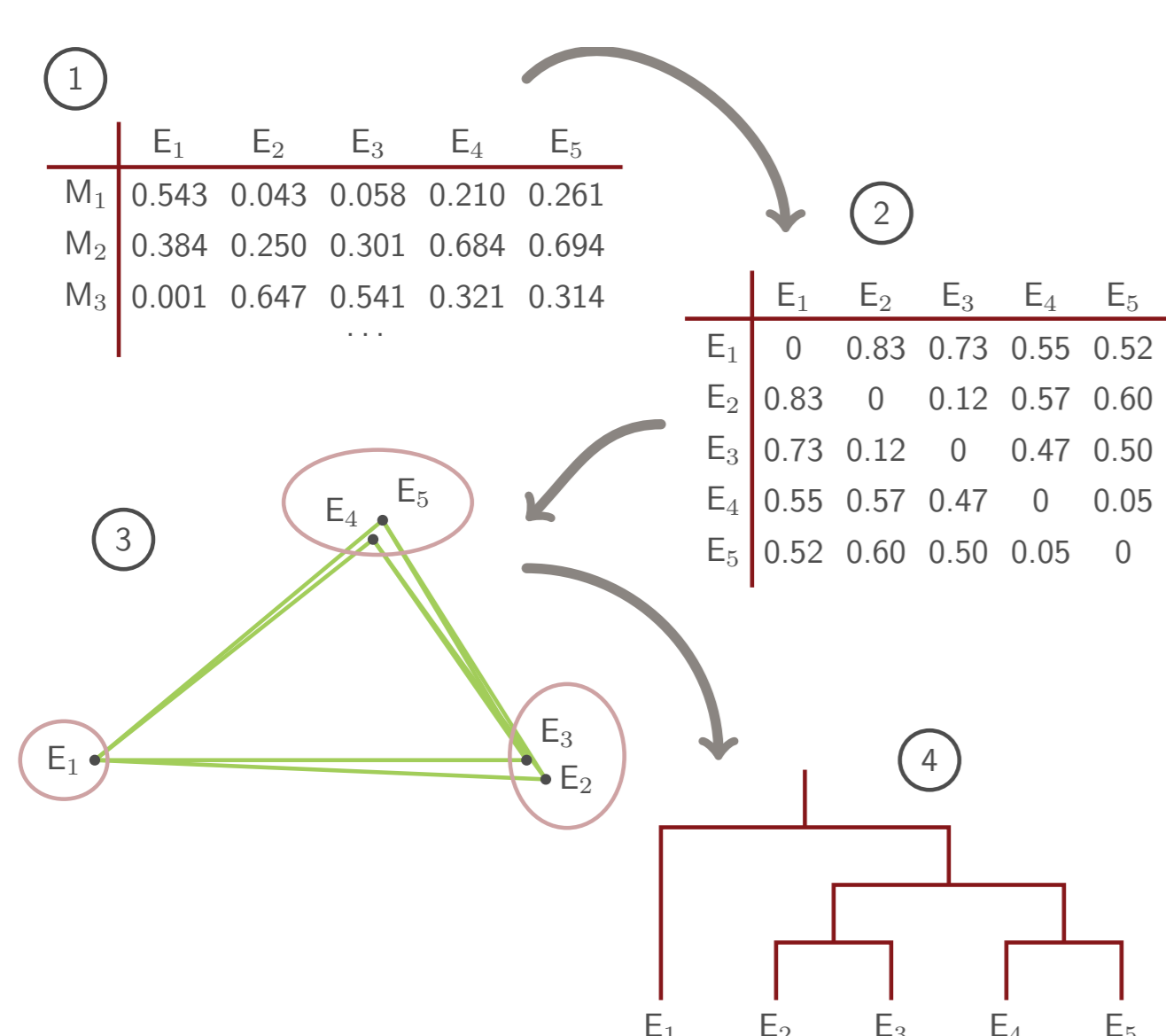
HCA	ESOM
Based on geometric assumptions	Based on a neural network
⊕ Fast and determining algorithm	⊖ Slow and with varying results
⊖ Arbitrary linkage method	⊕ Self-organization

References

- Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 30 (1), pp. 56–58.
- Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, pp. 53–65.
- Ultsch, A. (1999). *Kohonen Maps* chapter Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series, pp. 33–45. Elsevier.

Two different cluster approaches

Hierarchical Cluster Analysis



- Derivation of the multivariate data matrix.
- Calculation of the distance matrix using an arbitrary metric.
- Clustering of the statistical objects using an arbitrary linkage method. The linkage method defines a rule how the distance between clusters, containing more than one object is calculated. Commonly used linkage methods are *single linkage*, *complete linkage* or *average linkage*.
- Generation of the dendrogram from the clustering is straight forward.

Emergent Self-Organizing Maps

- Derivation of the multivariate data matrix.
- Training of the neural network. The two dimensional output layer is embedded in a borderless torus space.
- After the training, the output layer is transformed into the so called U-Matrix, where similarities between the output nodes are represented as heights in a three dimensional landscape (Ultsch 1999).
- The dendrogram is generated by application of a watershed algorithm on the U-Matrix. Emerging branches are merged on that level where two valleys in the U-Matrix are flooded.

