Metabolomics
DOI 10.1007/s11306-008-0105-7

OpenAccess

ORIGINAL ARTICLE

Genetic algorithm based two-mode clustering of metabolomics data

J. A. Hageman · R. A. van den Berg · J. A. Westerhuis · M. J. van der Werf · A. K. Smilde

Received: 28 September 2007 / Accepted: 11 January 2008
© The Author(s) 2008

Abstract Metabolomics and other omics tools are generally characterized by large data sets with many variables obtained under different environmental conditions. Clustering methods and more specifically two-mode clustering methods are excellent tools for analyzing this type of data. Two-mode clustering methods allow for analysis of the behavior of subsets of metabolites under different experimental conditions. In addition, the results are easily visualized. In this paper we introduce a two-mode clustering method based on a genetic algorithm that uses a criterion that searches for homogeneous clusters. Furthermore we introduce a cluster stability criterion to validate the clusters and we provide an extended knee plot to select the optimal number of clusters in both experimental and metabolite modes. The genetic algorithm-based two-mode clustering gave biological relevant results when it was applied to two real life metabolomics data sets. It was, for instance, able to identify a catabolic pathway for growth on several of the carbon sources.

Keywords Metabolomics · Two mode clustering · Biclustering · Genetic algorithms · Data analysis

1 Introduction

Functional genomics approaches have been applied in many different areas for the unraveling of complex biological questions. A functional genomics approach aims to obtain a complete overview of a certain biological response, for instance, gene expression levels or metabolite concentrations, in relation to the experimental conditions of interest. Obtaining a complete overview of the biological response enables the identification of interesting effects that would not be noticed if a subset of the genes or metabolites is analyzed.

Within functional genomics, metabolomics focuses on the analysis of the metabolome, the complete set of small organic molecules in, or outside, a cell. The metabolome is the most direct reflection of the phenotype of the organism under study, as regulatory effects, like post-transcriptional processing, or post-translational modification, do not hamper its interpretation (Fiehn 2002). In a metabolomics experiment, metabolome samples of an organism are generated under conditions that result in (large) variations of the metabolome composition.

The resulting variations are often analyzed with latent variable techniques or clustering methods. Latent variable techniques, such as PCA (Jolliffe 2002), PCDA (Hoogerbrugge et al. 1983), reduce the dimensions of the data to make interpretation easier. Clustering methods, on the other hand, order the data in groups that are similar

J. A. Hageman and R. A. van den Berg contributed equally to this paper.

Electronic supplementary material The online version of this article (doi:10.1007/s11306-008-0105-7) contains supplementary material, which is available to authorized users.

J. A. Hageman · J. A. Westerhuis (✉) · A. K. Smilde
Biosystems Data Analysis, Universiteit van Amsterdam, Nieuwe Achtergracht 166, Amsterdam 1018 WV, The Netherlands
e-mail: j.a.westerhuis@uva.nl

J. A. Hageman
ABC Metabolomics Centre, Wilhelmina Childrens Hospital, P.O. Box 85090, Utrecht 3508 AB, The Netherlands

R. A. van den Berg · M. J. van der Werf · A. K. Smilde
TNO Quality of Life, P.O. Box 360, Zeist 3700 AJ, The Netherlands

according to a particular similarity measure, such as the Euclidean distance, or the correlation coefficient (Vandeginste et al. 1998). The popularity of clustering methods results from their visualization and clear interpretation.

Clustering methods can be divided in two groups. The first group clusters the data set in either experiment or metabolite clusters; this is called one mode clustering. Here, the experiments or the metabolites are clustered based on the similarity of the behavior of all metabolite concentrations under an experimental condition or on the similarity of behavior of the concentration of a metabolite under all experimental conditions, respectively. The second group simultaneously creates experiment and metabolite clusters, which is called two-mode clustering or biclustering (Van Mechelen et al. 2004; Madeira and Oliveira 2004). Here the metabolites and experiments are clustered simultaneously to obtain groups of experiments and metabolites that behave as similar as possible. It is possible to apply a one-mode clustering method (e.g. hierarchical clustering, or k-means clustering) first to the metabolite mode and subsequently to the experiment mode, or vice versa. However, this will not result in identical results as by using two-mode clustering, as the clusters are not optimized for homogeneity in both the experimental and the metabolite mode. Therefore, two-mode clusters obtained by one-mode clustering methods are sub-optimal and the interpretation of these results will be hampered.

Two-mode clustering algorithms aim to find the best partitioning of the data in clusters. We define the best partitioning as the cluster assignment that results in the minimal difference between the model of the data and the original data. Different two-mode clustering algorithms exist, of which some algorithms are based on global optimization approaches, such as Simulated Annealing (SA) and Tabu Search (TS) (Prelic et al. 2006; Van Mechelen et al. 2004). The main advantage of global optimization methods is that they are able to find the global solution and not a locally optimal solution; something that is likely to happen with local optimization methods like steepest descent.

In this paper we introduce two-mode clustering of metabolomics data based on a Genetic Algorithm (GA). As GA's work on a group of solutions it can take large steps in the solution space and it is less likely to get stuck in local optima compared to SA and TS. The GA approach used in this paper is based on a cluster homogeneity criterion and not on distances between clusters. This means that clusters are based on metabolites that behave as similar as possible for a group of experimental conditions. Furthermore, quite some attention is paid to assess the cluster stability using a leave one out resampling of the two-mode clustering results. The selection of the number of clusters in both experimental and metabolite modes is performed using a generalized knee plot. Most two-mode clustering methods are specifically designed for gene expression data, but we apply our new two mode clustering approach to metabolomics data which improves their interpretation considerably. Two different metabolomics data sets with different complexity are analyzed to show the generality and usefulness of the new method.

2 Methods and materials

2.1 Data

The first data set *(P. putida* S12) is maintained at TNO (Zeist, the Netherlands). Cultures of *P. putida* S12 (Hartmans et al. 1990) were grown in batch fermentations at 30°C in a Bioflow II (New Brunswick Scientific) bioreactor as previously described (van der Werf et al. 2006). In short, samples were grown in triplicate on four carbon sources: D-fructose (sample F1, F2 and F3), D-glucose (sample G1, G2 and G3), gluconate (sample N1 and N2) and succinate (sample S1). Samples were analyzed by GC-MS and LC-MS. A detailed description is given elsewhere (Koek et al. 2006; van den Berg et al. 2006; Coulier et al. 2006). The GC-MS and LC-MS data set were fused together by concatenating the measurement tables (Smilde et al. 2005). The final data set was manually cleaned up, removing spurious and double entries and consisted of nine experiments and 162 metabolites.

The second data set (*E. coli* NST 74, a phenylalanine overproducing strain, and *E. coli* W3110, the wild-type strain) were grown at 30°C in a bioreactor containing 2 l of a medium with 30 g/l glucose as carbon source. A constant pH (pH 6.5) and oxygen tension (30%) was maintained. Samples were taken from the bioreactor after 16, 24, 40, 48 h, and immediately quenched. Variations in this standard fermentation protocol were introduced by changing one of the default conditions, resulting in a screening experiment. Samples were analyzed by GC-MS and LC-MS and fused together. A detailed description of this data set is given elsewhere (Smilde et al. 2005). The final data set was manually cleaned up, removing spurious and double entries and consisted of 28 experiments and 188 metabolites.

2.2 Genetic algorithms

GAs are a special class of global optimizers based on the theory of evolution. A GA minimizes a function F(x), where x represents a parameter vector, by searching the parameter space of x for the optimal solution. In the case of two-mode clustering, GAs will search for the optimal

partitioning of objects and variables by minimizing the residuals. The residuals are the difference between the model of the data and the original data matrix. Several steps in a GA are identical for all GAs and will be explained shortly (for a more detailed overview, see Supplementary Material Fig. 1) in the following:

(1) Initialization: GAs operate on a group of solutions, called a population. At the start of the GA, all solutions, also called strings or chromosomes, are set to random values.
(2) Evaluation: All strings in the population are evaluated by an evaluation function (see Sect. 2.3.1).
(3) Stop: A stop criterion is checked.
(4) Selection: A percentage of the best strings in a population is selected to form the next generation.
(5) Recombination: To form the new population, new solutions are created by combining two selected existing solutions (parents) to yield two different ones (children). This is called crossover.
(6) Mutation: Parts of a string in the new population are selected randomly and modified. To prevent the search from random behavior, the probability of mutation is usually chosen to be quite low.

Several parameters, such as the rate of crossover and mutation, regulate the performance of the GA. Each specific optimization problem has its own specific set of parameters for which the GA performs at its optimum. This so-called meta-optimization of the GA parameters can be tedious and can be considered a disadvantage of GAs in general. For more information regarding GAs, we refer to (Wehrens and Buydens 1998).

2.3 Two-mode clustering

2.3.1 The model

The goal of two-mode clustering is to simultaneously find the optimal partitioning between objects and variables of data matrix **X**, as depicted in Fig. 1. For two-mode clustering, data matrix **X** is approximated by

$$\mathbf{X} = \mathbf{UYV}^{\mathrm{T}} + \mathbf{E} \quad (1)$$

where

X (M × N): data matrix of *M* rows and *N* columns.
U (M × P): membership matrix for *M* rows (metabolites) of matrix **X** allowing for *P* row clusters. This matrix contains on each row (*P*-1) zeros and a single 1. The location of this 1 indicates the cluster membership.
Y (P × Q): matrix containing the clusters averages for *P* row and *Q* column clusters.

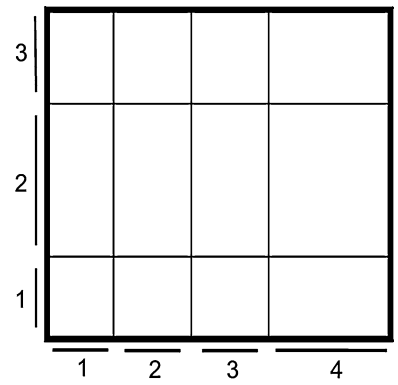


Fig. 1 Schematic representation of two-mode partitioning

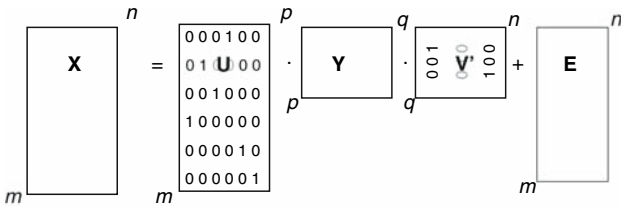


Fig. 2 Schematic representation of the decomposition of matrix X. See text for details

V (N × Q): membership matrix for *N* columns (experiments) of matrix **X** allowing *Q* column clusters. The structure of this matrix is similar to that of matrix **U**.
E (M × N): matrix containing the difference between each measurement and the average of the cluster it belongs to.

A schematic representation of this decomposition is given in Fig. 2.

Pretreatment of the data is an important aspect of data analysis that can dramatically influence the results of data analysis (van den Berg et al. 2006). In this paper, range scaling was applied to accentuate the biological information content of the metabolomics data set by converting the concentrations to values relative to the biological range of a metabolite. The biological range is defined as the difference between the minimum and maximum concentration measured for a metabolite in the data set. In this way, high or low metabolite concentrations and the way in which the concentrations of metabolites are affected by different environmental conditions are seen within the context of the natural variation of the concentration (dynamic range) of those metabolites.

2.3.2 Evaluation function

For the evaluation function, the partitioning information on the string is used to construct membership matrices U and

V. Matrix Y is obtained in two steps. In the first step, the sums of all metabolites in a cluster are obtained:

$$\tilde{\mathbf{Y}} = \mathbf{U}^T\mathbf{X}\mathbf{V} \tag{2a}$$

In the second step, all elements are divided by the number of members in that cluster to obtain cluster averages in Y.

$$y_{p,q} = \frac{\tilde{y}_{p,q}}{u_p \cdot v_q} \tag{2b}$$

Here $y_{p,q}$ is the average value of a two-mode cluster (p,q), u_p and v_q indicate the number of metabolites and experiments respectively for two-mode cluster (p,q).

The residual matrix **E** is then given by:

$$\mathbf{E} = \mathbf{X} - \mathbf{U}\mathbf{Y}\mathbf{V}^{\mathrm{T}} \tag{3}$$

Matrix $\mathbf{U}\mathbf{Y}\mathbf{V}^{\mathbf{T}}$ is the approximation of **X** and contains for each metabolite a value equal to its cluster average. For an optimal two-mode clustering result, the GA minimizes the sum of squares (SS) of the elements of **E**. The smaller the values in **E**, the tighter the corresponding clusters are.

2.3.3 Software

The two-mode genetic algorithm clustering method was programmed in Matlab 7.1 (The Mathworks Inc. 2005b) using the Genetic Algorithm and Direct Search (GADS) (The Mathworks Inc. 2005a). A special integer type coding scheme was written for use with this toolbox. This scheme encodes the cluster number for each M metabolites and N experiments, so each string in the GA population has length M + N. The cluster number is an integer between 1 and the maximum number of clusters. The mutation operator replaces, with a certain probability, a value from the string with a random number between 1 and the maximum number of clusters. The settings used for the GA are listed in the Supplementary Material Table 1.

All GA runs were executed in five-fold with different random seeds to exclude any (un)lucky starting positions. The results from the five runs should be similar, and the best solution is chosen. The evaluation function was optimized for speed using the profile function of Matlab, resulting in run-times of five minutes for five replicate runs for the *Pseudomona putida* S12 data set and run-times of ten minutes for the *Escherichia coli* data set. Since two-mode k-means is a local optimizer and is known to get easily stuck in local optima, the two-mode k-means was restarted 50 times for each solution and the best solution out a possible 50 was kept. All calculations were performed on an AMD Athlon XP 2400 + 2.00 GHz 512 MB RAM PC running Windows XP. The GA two-mode clustering routines applied in this paper are available at http://www.bdagroup.nl.

2.4 Number of clusters

Partitioning clustering algorithms require a predefined number of clusters. There are a number of methods for finding the most suited number of clusters in the data, such as, the Bayesian Information Criterion (BIC) (Raftery 1986), the GAP statistic (Tibshirani et al. 2001) and the knee or 'L' method (Salvador and Chan 2004).

We chose the knee method which finds the knee or 'L' in a plot of the number-of-clusters versus the SS of the residuals. The assumption of this method is that an additional cluster gives a sharp decrease in the SS of the residuals as long as the optimal number of clusters is not reached. When more than the optimal number of clusters is chosen, the decrease in SS of the residuals is less sharp and more or less equal for each additional cluster.

The knee method can be generalized to two-mode clustering. In this case, the curve of the number-of-clusters versus the sum of squared residuals plot is a contour plot. In this plot there is a combination of cluster numbers for the experiments and metabolites for which an additional cluster no longer sharply decreases the SS of the residuals.

2.5 Validation

The two mode clustering method was validated by leaving one experiment out (LOO) of the data set, clustering this data set again and comparing the obtained results with the clustering of the full data set. In this way, the dependence of the clustering on one single experiment can be assessed. A stable clustering will less likely be influenced by leaving one experiment out. For the *P. putida* S12 data set, at least one experiment per group remained in the data set to maintain the structure of the experimental design. All LOO-data sets were pretreated and clustered. When comparing the content of a cluster obtained with the LOO procedure, it was made sure that it was compared with the correct cluster obtained with the complete data by first establishing which clusters have to most overlap and linking them together. The LOO validating scheme only validates the effect of the experiments on the metabolite clustering. If desired, it is possible to validate the effect of the metabolites on the experiment clustering in a similar way.

In order to analyze the spread within clusters, the cluster variances are used as a diagnostic tool:

$$s_k^2 = \frac{\sum_{n_k=1}^{N_k} (x_{n_k} - y_k)^2}{N_k - 1} \tag{4}$$

Here, x_{n_k} indicates the cluster element n of cluster k for a total of N_k elements and y_k is the mean of the cluster k. The variances of the different clusters can be compared; a

relatively low variance indicates small and compact clusters. In contrast, a relatively high variance indicates large and/or heterogeneous clusters and this could be a sign of, for instance, outliers.

The cluster variances are a natural diagnostic of the cluster quality as they are directly linked to the evaluation function. The variances of each two-mode cluster can be combined to give the pooled variance:

$$s^2_{\text{pooled}} = \frac{\sum_{k=1}^{K} (N_k - 1) s_k^2}{\sum_{k=1}^{K} (N_k - 1)} \tag{5}$$

The evaluation function (Eq. 3) and the pooled variance are identical up to a scaling factor as is proven in the Supplement (Appendix A).

3 Results

3.1 Estimation of the number of clusters

3.1.1 *P. putida* data

The generalized knee method is used to obtain an estimate of the number of clusters in the partitioning. The rate of decrease for the residuals became smaller after four experimental clusters and four/five metabolite clusters (see Fig. 2 Supplementary Material). Obtaining four experiment clusters may seem trivial, however, it is possible that some of the experiments are rather similar and end up in the same cluster. For the metabolite clusters, both the four and five cluster solutions were analyzed and the five cluster choice was found to be more meaningful.

When comparing the results from the two-mode clustering with two single k-means clustering on the metabolites and the experiments (results not shown), the sum of squared residuals was 7.8% lower when applying two-mode clustering. The data set was also subjected to a classical non-GA based two-mode k-means method with the same evaluation function as the GA two mode algorithm (Vichi 2001; Baier et al. 1997). Figure 3 shows the comparison of the resulting sum-of-squares. For a larger number of clusters, GA tends to give better results than two-mode k-means. In the cases that two-mode k-means has a lower sum-of-squares it is usually only lower by a small amount, indicating that both algorithms have reached the same global minimum but with different precisions.

3.1.2 *E. coli* data

A similar analysis was performed for the *E. coli* data showing seven experimental clusters and six metabolite clusters was optimal. The performance of GA against k-means was again tested (see Fig. 3) and showed that relatively quickly the GA outperforms the two-mode k-means solution.

3.2 Two mode clustering

3.2.1 *P. putida* data

The two-mode cluster result is presented in Fig. 4. The different patterns of the metabolites under the different growth conditions are clearly visible. For example, the behavior of the metabolites in D-fructose and D-glucose grown cells differs the most for the metabolites in clusters II and V. The stability of the clustered metabolites was tested with a leave-one-out validation strategy (see Sect. 2.5). The gray scale shows how often metabolites switch to another cluster during LOO validation. Only a few metabolites switch often, so the results are stable.

The visualization of the two-mode clustering result allows for the instant detection of outliers, as the color of an outlying variable is different from the consensus color of a cluster. In

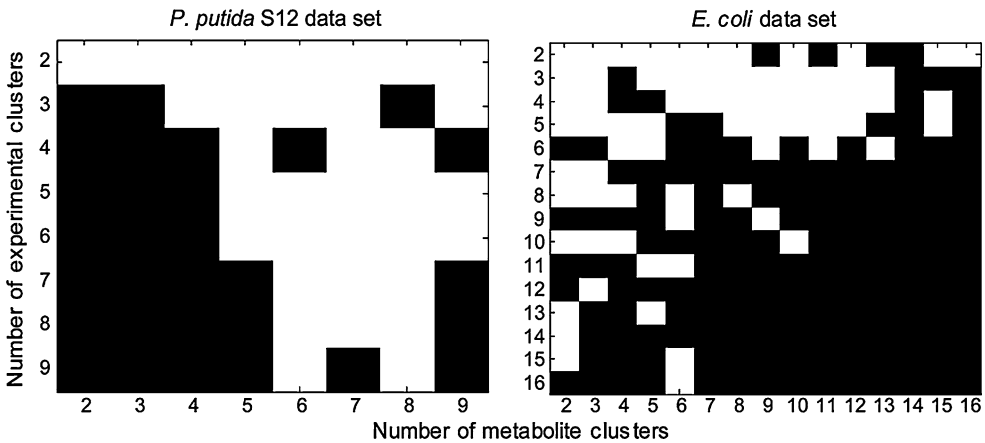


Fig. 3 Comparison of GA two-mode clustering and two-mode k-means clustering results for *P. putida* (left) and *E. coli* (right). The black area shows when GA two-mode clustering gave better results in terms of the evaluation criterion for a certain metabolite/experiment cluster combination. The white area shows when two mode k-means gave the best results

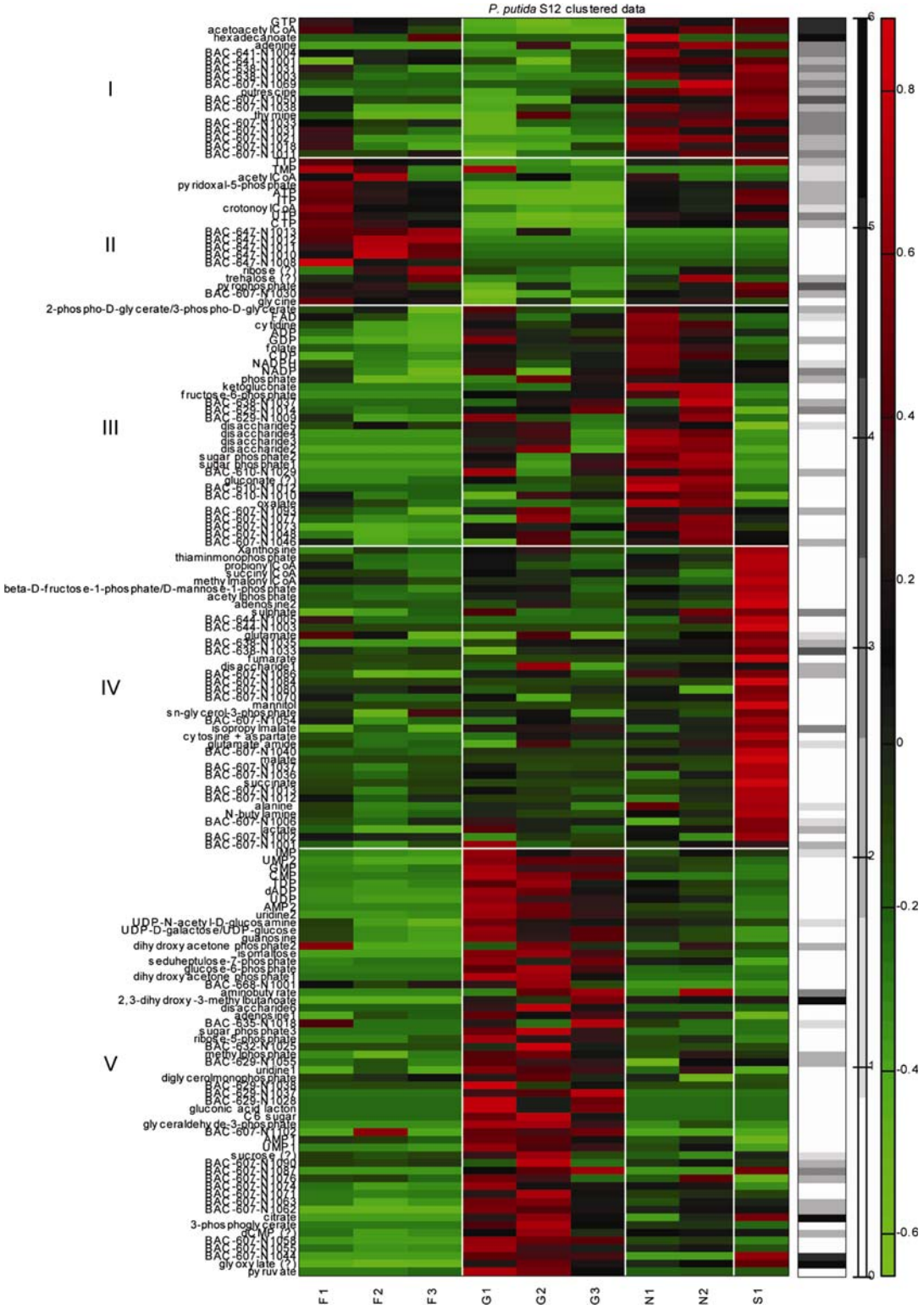


Fig. 4 Two-mode clustering results for the metabolome data set of *P. putida* S12 grown on four different carbon sources. The roman numerals I–V, and F, G, N and S, are used to refer to the corresponding clusters throughout the text. The black/white bar indicates the number of cluster swaps a certain metabolite has made during the loo validation. Some metabolites were analyzed by GC-MS and LC-MS but their identity is not known, or were only identified as part of a class of metabolites, e.g. disaccharides. These metabolites were given a number behind the metabolite name to be able to distinguish between them during validation. For some metabolites there is uncertainty about the identification. These metabolites were given a question mark

cluster FV, for instance, BAC-607-N1102 in experiment F2 is bright red, while most of the cluster is green, just as the results for F1 and F3 (Fig. 4). This indicates that BAC-607-N1102 is a deviating point in the result of F2.

It is important to know whether the estimated cluster average is a suitable estimate of a cluster. By calculating the variance of a cluster, a measure for the homogeneity of the cluster is obtained. The variances for the two-mode clustering are presented in Fig. 5. Most of the variances are comparable. FIII is the most homogeneous clusters found, while SII and GIII contain the most variance. Analysis of the cluster variance can thus be applied as a quick assessment of the capability of the cluster to summarize the containing data.

When the resulting clusters are studied in more detail, several clusters contain interesting information. For instance, cluster V contains dihydroxyacetonephosphate1 (DHAP), pyruvate, glucose-6-phosphate (G6P), 3-phosphoglycerate (3PGA), glyceraldehyde-3-phosphate (GAP) and gluconic-acid-lactone (GLN). These metabolites are catabolic intermediates of the degradation pathway of D-fructose, gluconate and D-glucose (Fig. 6).

On the other hand, fructose-6-phosphate (F6P) is member of cluster III, even though it is also an intermediate of the catabolic pathway of D-fructose, gluconate, and D-glucose. F6P connects the degradation pathways of D-fructose, gluconate, and D-glucose with the pentose phosphate pathway (PPP) (Fig. 6). It is possible that the switch between the PPP and the degradation pathway explains why F6P was assigned a different cluster. The lack of 6-phosphofructokinase in *Pseudomonas* (Lessie and Phibbs 1984) probably contributes to this behavior as well. This result shows that two-mode clustering can find clusters that are informative from a biological point of view.

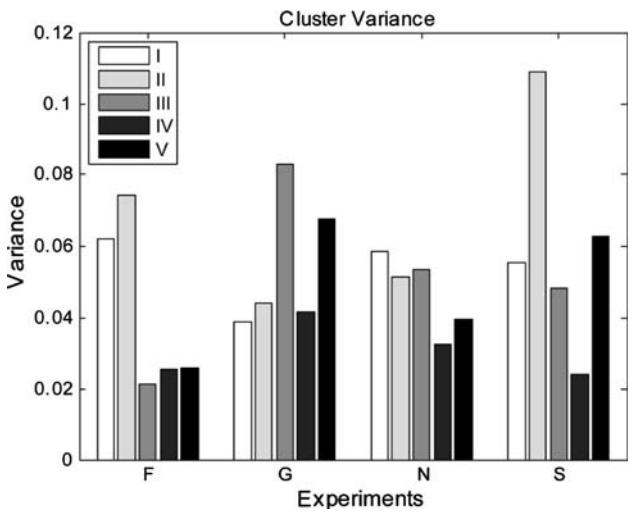


Fig. 5 Variances of the clusters in Fig. 4

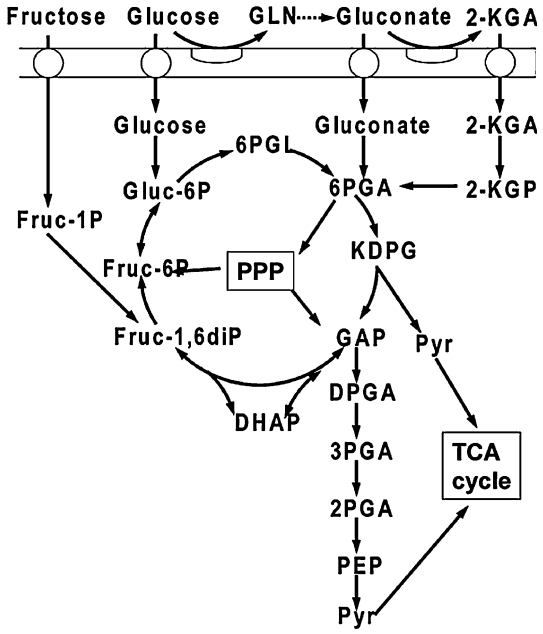


Fig. 6 Degradation of fructose, glucose and gluconate by the cyclic Entner-Doudoroff pathway in Pseudomonas. Taken with permission from SGM Microbiology (van der Werf et al. 2006)

3.2.2 E. coli *data*

The two-mode clustering results of the *E. coli* data is shown in Fig. 7. This data set is more complicated than the *P. putida* data because more perturbations were performed and longitudinal measurements were analyzed. The leave-one-out results are again shown as a gray scale bar (see Fig. 7). The complexity of the data set is reflected in these results since the clustering is less stable compared to the *P. putida* data. Yet, biological meaningful results were obtained with respect to both the clustering of the metabolites and the samples. For instance, most nucleotides cluster together (cluster III) and the ketoglutarate/glutamate and malate/fumarate/aspartate pairs that are converted into each other by one enzymatic reaction, cluster together. On the other hand, with the clustering of the samples it was observed that the samples taken at the earlier time points cluster together, but also the samples of the wild-type strain, and samples collected from fermentations using succinate as the carbon source, cluster together.

4 Concluding remarks

Genetic algorithm based two-mode clustering is a valuable tool for the identification of biologically meaningful clusters in metabolomics data. Furthermore, it visualizes which subset of metabolites responds to which experimental condition. The results are validated by the use of a leave-

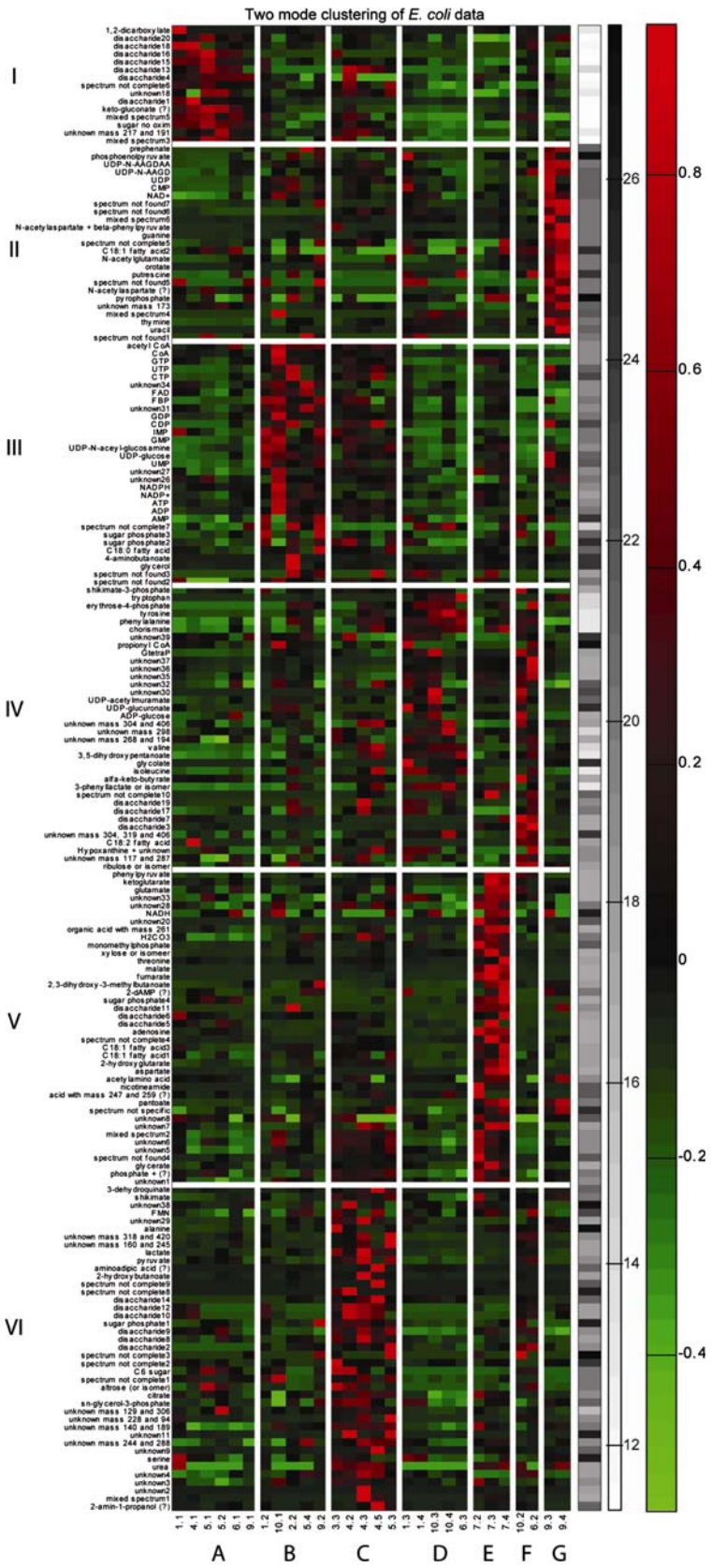


Fig. 7 Two-mode clustering results for the metabolome data set of *E. coli*. The numerals I–VI, and characters A–G are used to refer to the corresponding clusters throughout the text. The black/white bar indicates the number of cluster swaps a certain metabolite has made during the loo validation. Some metabolites were analyzed by GC-MS and LC-MS but their identity is not known, or were only identified as part of a class of metabolites, e.g. disaccharides. These metabolites were given a number behind the metabolite name to be able to distinguish between them during validation. For some metabolites there is uncertainty about the identification. These metabolites were given a question mark

one-out validation scheme that allows for the identification of metabolites that have an unstable clustering. A second validation measure is the analysis of the cluster variance. This gives insight in the homogeneity of the clusters and thus how well the clusters fit the data. Application of the newly developed approach to metabolomics data results in the identification of biologically relevant clusters.

The algorithm compares favorably to other approaches (e.g. two-mode k-means and single one-mode clustering). Hence, the genetic algorithm based two mode clustering, together with an extensive validation of the results, is a valuable addition to the omics data analysis toolbox, as it provides a detailed overview of the data.

Acknowledgements The authors would like to thank Richard Bas and Leon Coulier for analyzing the samples by LC-MS. Joost van Rosmalen is kindly thanked for sharing his implementation of the two mode k-means algorithm. This research was funded by the Kluyver Centre for Genomics of Industrial Fermentation, which is supported by the Netherlands Genomics Initiative (NROG) and the Netherlands Bioinformatics Consortium (NBIC) and this work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

Baier, D., Gaul, W., & Schader, M. (1997). Two-mode overlapping clustering with applications to simultaneous benefit segmentation and market structuring. In R. Klar & O. Opitz (Eds.), *Classification and knowledge organization*. Heidelberg: Springer.

Coulier, L., et al. (2006). Simultaneous quantitative analysis of metabolites using ion-pair liquid chromatography-electrospray ionization mass spectrometry. *Analytical Chemistry, 78*, 6573–6582.

Fiehn, O. (2002). Metabolomics—the link between genotypes and phenotypes. *Plant Molecuar Biology, 48*, 151–171.

Hartmans, S., van der Werf, M. J., & de Bont, J. A. M. (1990). Bacterial degradation of styrene involving a novel flavin adenine dinucleotide-dependent styrene monooxygenase. *Applied and Environmental Microbiology, 56*, 1347–1351.

Hoogerbrugge, R., Willig, S. J., & Kistemaker, P. G. (1983). Discriminant analysis by double stage principal component analysis. *Analytical Chemistry, 55*, 1710–1712.

Jolliffe, I. T. (2002). *Principal component analysis*. New York: Springer-Verlag.

Koek, M., et al. (2006). Microbial metabolomics with gas chromatography mass spectrometry. *Analytical Chemistry, 78*, 1272–1281.

Lessie, T. G., & Phibbs, P. V. J. (1984). Alternative pathways of carbohydrate utilization in Pseudomonads. *Annual Review of Microbiology, 38*, 359–387.

Madeira, S. C., & Oliveira, A. L. (2004). Bicluster algorithms for biological data analysis: A survey. *IEEE Transactions on Computational Biology and Bioinformatics, 1*, 24–45.

Prelic, A., et al. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics, 22*, 1122–1129.

Raftery, A. E. (1986). Choosing models for cross-classifications. *American Sociological Review, 51*, 145–146.

Salvador, S., & Chan, P. (2004). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Proceedings of the 16th IEEE International Conference on Tools with Arificial Intelligence (ICTAI 2004)* (pp. 576–584).

Smilde, A. K., et al. (2005). Fusion of mass-spectrometry-based metabolomics data. *Analytical Chemistry, 77*, 6729–6736.

The Mathworks Inc. (2005a). Genetic Algorithm Direct Search Toolbox 2.0.

The Mathworks Inc. (2005b). Matlab 7.1 (R14).

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society B, 63*, 411–423.

van den Berg, R. A., et al. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*, 7, 142.

van der Werf, M. J., et al. (2006). Multivariate analysis of microarray data by principal component discriminant analysis: Prioritizing relevant transcripts linked to the degradation of different carbohydrates in *Pseudomonas putida* S12. *Microbiology, 152*, 257–272.

Van Mechelen, I., Bock, H.-H., & De Boeck, P. (2004). Two-mode clustering methods: A structured overview. *Statistical Methods in Medical Research, 13*, 363–394.

Vandeginste, B. G. M., et al. (1998). *Handbook of chemometrics*. Amsterdam: Elsevier.

Vichi, M. (2001). Double k-means clustering for simultaneous classification of objects and variables. In S. Borra et al., (Eds.), *Advances in classification and data analysis* (pp. 43–52). Heidelberg: Springer.

Wehrens, R., Buydens, L. M. C. (1998). Evolutionary optimisation: A tutorial. *Trends in Analytical Chemistry, 17*, 193–203.