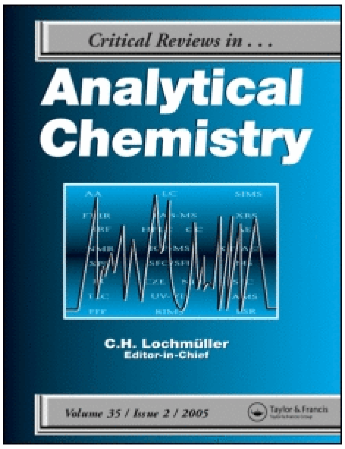

Critical Reviews in Analytical Chemistry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713400837

Bagged K-Means Clustering of Metabolome Data

J. A. Hageman[a]; R. A. van den Berg[b]; J. A. Westerhuis[a]; H. C. J. Hoefsloot[a]; A. K. Smilde[a]
[a] Biosystems Data Analysis, Swammerdam Institute for Life Sciences (SILS), Universiteit van Amsterdam, Amsterdam, The Netherlands [b] TNO Quality of Life, AJ Zeist, The Netherlands

To cite this Article Hageman, J. A. , Berg, R. A. van den , Westerhuis, J. A. , Hoefsloot, H. C. J. and Smilde, A. K.(2006) 'Bagged K-Means Clustering of Metabolome Data', Critical Reviews in Analytical Chemistry, 36: 3, 211 — 220
To link to this Article: DOI: 10.1080/10408340600969916
URL: http://dx.doi.org/10.1080/10408340600969916

Critical Reviews in Analytical Chemistry, 36:211–220, 2006
Copyright © Taylor and Francis Group, LLC
ISSN: 1040-8347 print / 1547-6510 online
DOI: 10.1080/10408340600969916

Bagged K-Means Clustering of Metabolome Data

J. A. Hageman
Biosystems Data Analysis, Swammerdam Institute for Life Sciences (SILS), Universiteit van Amsterdam, Amsterdam, The Netherlands

R. A. van den Berg
TNO Quality of Life, AJ Zeist, The Netherlands

J. A. Westerhuis, H. C. J. Hoefsloot, and A. K. Smilde
Biosystems Data Analysis, Swammerdam Institute for Life Sciences (SILS), Universiteit van Amsterdam, Amsterdam, The Netherlands

Clustering of metabolomics data can be hampered by noise originating from biological variation, physical sampling error and analytical error. Using data analysis methods which are not specially suited for dealing with noisy data will yield sub optimal solutions. Bootstrap aggregating (bagging) is a resampling technique that can deal with noise and improves accuracy. This paper demonstrates the possibilities for bagged clustering applied to metabolomics data. The metabolomics data used in this paper is computer-generated with the human red blood cell model. Perturbing this model can be done in several ways. In this paper, inhibition experiments are mimicked inhibiting enzyme activity to 10% of its original value. Comparing bagged K-means clustering to ordinary K-means, the number of metabolites switching clusters under the influence of heteroscedastic noise is lower if bagging is used. This favors bagged K-means above ordinary K-means clustering when dealing with noisy metabolomics data. A special validation scheme, independent of the addition of noise, has been devised to demonstrate the positive effects of bagging on clustering.

Keywords clustering, metabolomics, correlation, resampling, bootstrap aggregating, bagging, perturbing metabolome models, validation

INTRODUCTION

The goal in metabolomics (1–3) is to investigate the (functional) relationship between metabolites in an organism. The interest in metabolomics has grown considerably, as the metabolome is the most direct reflection of the cell state, in contrast to proteomics or transcriptomics in which regulatory effects hamper clear interpretation of the results. In the ideal situation the underlying metabolic network can be elucidated by studying the metabolome. Usually, a complete elucidation is not obtained and the resulting information is cruder in nature.

After the identification and quantification of metabolites, the next step is to estimate the functional relationship(s) between

Address correspondence to J. A. Westerhuis, Biosystems Data Analysis, Swammerdam Institute for Life Sciences, Universiteit van Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands. E-mail: westerhuis@science.uva.nl

Abbreviations: used: 1,3-BPG (B13PG), 1,3-bisphosphoglycerate; 2,3-BPG (B23PG), 2,3-bisphosphoglycerate; BPGP, 2,3-BPG phosphatase; BPGS, 2,3-BPG synthase; BPGS/PP, 2,3-BPG synthase/phosphatase; 2-PGA (P2GA), 2-phosphoglycerate; 3-PGA (P3GA), 3-phosphoglycerate; 6-PG (P6G), 6-phosphogluconate; 6-PGDH, 6-phosphogluconate dehydrogenase; 6-PGL (P6GL), 6-phosphogluconolactone; AK, adenylate kinase; Ald, aldolase; ENO, enolase, Ery4P, erythrose 4-phosphate; Fru(1,6)P2, fructose 1,6-bisphosphate; Fru6P, fructose 6-phosphate; GDH, glucose dehydrogenase; G6PDH, glucose-6-phosphate dehydrogenase; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; Glc(1,6)P2, glucose 1,6-bisphosphate; Glc, glucose; Glc6P, glucose 6-phosphate; GPI, glucosephosphate isomerase; GSSGR, glutathione reductase, lactonase, δ-gluconolactonase; GraP, glyceraldehyde 3-phosphate; GrnP, dihydroxyacetone phosphate; Hct, haematocrit; HK, hexokinase; kATPase, non-glycolytic energy consumption; kox, reduction processes consuming GSH; koxNADH, reducing processes requiring NADH; Lac, lactate; Lace, extracellular lactate; lactonase, dgluconolactonase; LDH(P), NADPH-dependent lactate dehydrogenase; LDH, lactate dehydrogenase; PEP, phosphoenolpyruvate; PFK, phosphofructokinase; PGK, phosphoglycerate kinase; PGM, phosphoglycerate mutase; Pie, extracellular inorganic phosphate; PK, pyruvate kinase; PPP, pentose phosphate pathway; Pyr, pyruvate; Pyre, extracellular pyruvate; R5PI, ribose-5-phosphate isomerase; Rib5P, ribose 5-phosphate; Ru5E, ribulose-5-phosphate epimerase; Ru5P, ribulose 5-phosphate; Ru5PE, ribulose 5-phosphate epimerase; Sed7P, sedoheptulose 7-phosphate; TA, transaldolase; TK, transketolase; TPI, triosephosphate isomerase; Xu5P, xylulose 5-phosphate.

the different metabolites. By discovering which metabolites can chemically be converted to each other, metabolic networks can be elucidated. The long term goal is to answer questions such as which metabolites are able to regulate which reactions, or genes, or gene products. Answers to these questions are important, as they can for instance, help to elucidate disease mechanisms, or provide leads to increase the yield in microbial production processes by identifying bottlenecks (1). The ultimate goal of systems biology is to combine the data from all x-omics fields to create a complete picture of the response of an organism to environmental conditions.

There are still some problems associated with metabolome data. To mention a few; First, the biological variation is rather large. Second, at the moment it is not possible to measure cell compartments separately. If compartments are present, they are broken and no spatial metabolic information can be obtained. Third, metabolite concentrations are usually determined when the biological system is (supposedly) in a steady state, while localized time resolved concentration measurements can also provide valuable information.

One of the most basic comparisons that can be made between different metabolite concentrations measured at varying conditions, is the Pearson product moment correlation coefficient (in short correlation coefficient) (4) and this has been used as a basis in several methods. Steuer et al. (5) devised a method to construct a metabolic correlation network. Using the correlation between metabolites as distances between metabolites, each metabolite is assigned coordinates in a two-dimensional plane by applying multidimensional scaling. Steuer et al. drew a map in which metabolites were connected by a line if their correlation value exceeded a certain threshold level (5, 6). A different method for visualizing metabolite interaction is a method using cliques by Kose et al. (7). Again, correlations between metabolites are taken as the starting point. Another method for reconstructing metabolic networks was presented by Arkin et al. (8) called correlation metric construction (CMC). This method applies small random perturbations to a glycolytic model. Using the resulting data, reaction pathways are estimated. Clish et al. (9) also used the Pearson product moment correlation for visualizing possible associations between two entities (in their case a protein or metabolite peak or gene). When the correlation coefficient was higher than 0.7, both entities were connected to each other by a line, resulting in a network of highly correlated species.

In this paper metabolites are clustered using K-means and bootstrap aggregating (bagging) (10). Bagging is a special bootstrapping technique. Normally, bootstrap samples are used for estimating confidence intervals while bagging is used to improve results by decreasing the variance. When applying bagging to clustering, the results of many clustered bootstrap samples are combined and visualized, relating metabolites to each other. The power of the method is demonstrated by using simulated data that has been generated with the red blood cell model from Kuchel and Mulquiney (11–13). The influence of noise on the bagged clustering method will be checked by comparing the total number of metabolites that switch clusters when using different levels of noise. The goal of this paper is to demonstrate when taking special considerations when dealing with noisy metabolomics data, clustering results improve. The use of algorithms that are especially well suited for dealing with high levels of noise can help to improve the results of data analysis. To demonstrate any effects of bagging in the presence of high levels of noise, a special validation scheme is used.

THEORY

A typical result of a metabolomics experiment is an $M \times N$ data matrix $\mathbf{X}$ that contains the measured concentrations of the M metabolites from N experiments. It is possible to look for clusters within the metabolite concentrations. Unfortunately, the absolute concentrations of many metabolites are very low and changes in their concentrations are minimal in different experiments. These metabolites are all clustered together since their concentrations and changes therein show many similarities. However, this does not reveal any information on their functional relationship. It merely indicates that these metabolites have similar concentrations. It is possible that scaling can improve such an approach, however, this is not pursued here.

To investigate a possible connectivity between metabolites, the correlation coefficient is much more informative. The correlation coefficient indicates, on an absolute scale, how much two variables co-vary and it is straightforward to calculate. The correlation coefficient r is shown in Eq. [1].

$$r(y_1, y_2) = \frac{\sum (y_{1i} - \bar{y}_1)(y_{2i} - \bar{y}_2)}{\sqrt{\sum (y_{1i} - \bar{y}_1)^2 \sum (y_{2i} - \bar{y}_2)^2}} \qquad [1]$$

Here, $\bar{y}_1$ and $\bar{y}_2$ indicate the average intensity of metabolites y_1 and y_2, i indicates the i^{th} measurement. The term in the denominator is a scaling factor, to ensure that the correlation coefficient r is a dimensionless number between -1 and $+1$, indicating a perfect (negative-)correlation between two profiles y_1 and y_2. A correlation of 0 indicates no correlation. For all pairs of metabolites, the correlation coefficient is calculated and stored in correlation matrix $\mathbf{P}$. Matrix $\mathbf{P}$ contains all the information on correlation between metabolites and is the basis of the proposed method. Matrix $\mathbf{P}$ can be very large, depending on the number of metabolites being analyzed. Typical numbers are in the range of a few hundred to a few thousand. Visualization of this matrix is very important, especially for large matrices.

Instead of using metabolite concentrations as a starting point for clustering algorithms, correlation coefficients can also be used. In that case, metabolites that behave similarly (and thus have high correlation coefficients) will be put in the same cluster. Similar behavior in this context means that metabolites profiles change coherently during each experiment indicating a functional relationship.

A problem usually present in metabolomics data are the high noise levels. Due to the sampling errors and the biological and analytical variation present in metabolomics experiments, the noise levels are usually quite high. The high noise levels make it hard to obtain correct cluster information since metabolites can switch to other clusters much easier. Therefore it is important for any method dealing with metabolomics data to deal accurately with large noise levels, since any clustering method will be influenced by the noise present.

To overcome chance correlations due to noise, a different approach based on bootstrap aggregating (bagging) has been developed. Bagging was introduced by Breiman in 1996 to improve the performance of prediction models (10). Since then, a number of papers have appeared in which the algorithms are modified and applied (some examples can be found in Refs. (14–18)). In general, bagging creates *B* bootstrap samples by drawing experiments with replacement from the total data set. Each bootstrap sample is fitted to a given model and each model is aggregated by computing the mean (in case of regression) or by majority voting (in case of clustering). By averaging the information resulting from each model, bagging is able to improve the accuracy of the global model.

The method proposed in this paper is based on bagging. It consists of the following steps (depicted in Figure 1):

- Instead of performing one clustering on the complete data as described above, the data is resampled b times (b = 1... B). During each resampling, a number of experiments are randomly chosen with replacement (step 1 in Figure 1).
- After each resampling, a K-means clustering on the correlation matrix $\mathbf{P_b}$ of bootstrap dataset $\mathbf{X_b}$ is performed (steps 2–3 in Figure 1).
- The clustering results are collected and added to a series of histograms. For each metabolite, a histogram is created that keeps track of how often that metabolite is clustered with what other metabolites (step 4 in Figure 1). Figure 3 shows an example of a histogram for 1,3-BPG collecting all clustering information.
- These histograms can be combined in a heatplot, in which the degree of clustering between metabolites is expressed with a color. Such a heatplot looks very chaotic (step 5 in Figure 1). An example of such a heatplot is shown in the top figure of Figure 4.

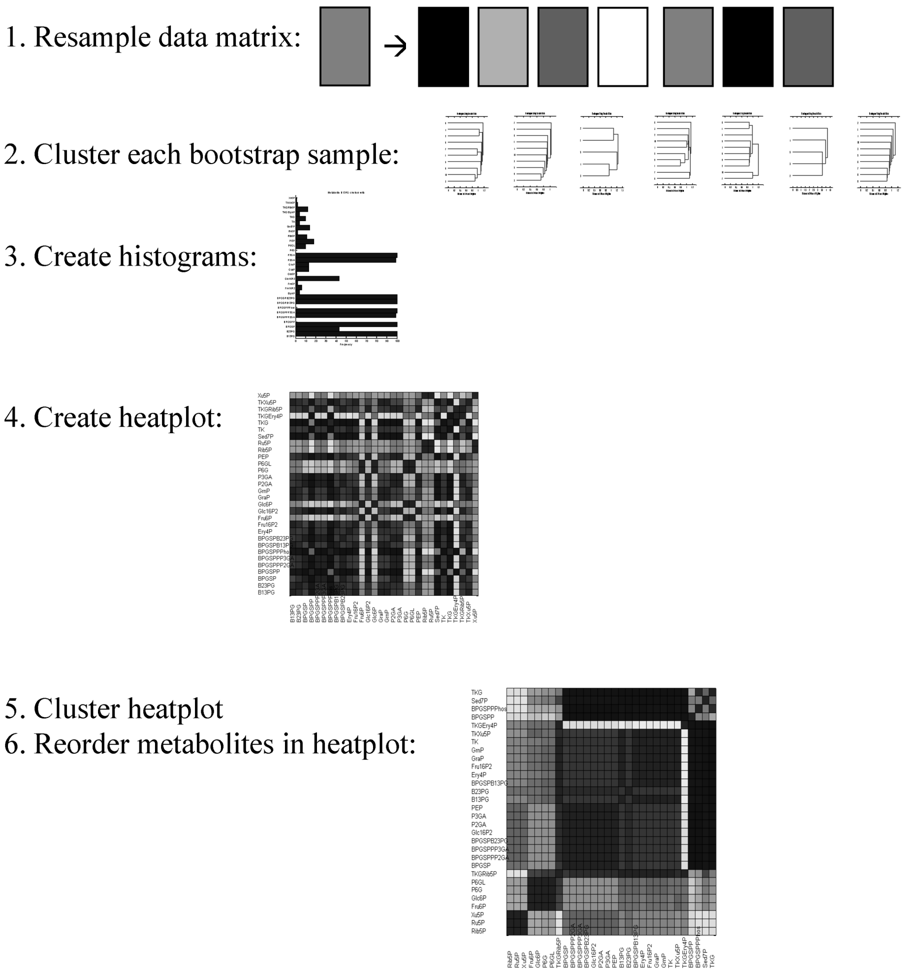

FIG. 1. Flowchart of the histogram clustering method.

- K-means clustering of the heatplot reveals which metabolites should be grouped together. Reordering the metabolites in the heatplot according to this clustering, gives information which metabolites are clustered together (step 6 and 7 in Figure 1). The bottom figure of Figure 4 shows an example of a reordered heatplot.

EXPERIMENTAL

Data

The data used in this paper is simulated data from the erythrocyte metabolism (11–13, 19). The erythrocyte metabolism is a well modeled biochemical system for a number of reasons. Red blood cells can be extracted easily which makes research relatively simple, and red blood cells have a relatively simple metabolism since they lack mitochondria and other organelles. Figure 2 shows an overview of erythrocyte metabolism. The best known physiological function of the erythrocyte is the transport of oxygen and CO_2 through the body. The erythrocyte metabolism consists of three pathways: the glycolytic, pentose phosphate and the 2,3-BPG or Rapoport-Luebering pathway. More details on the model can be found in the book of Mulquiney and Kuchel (19, 20).

Different approaches are available for the generation of metabolic profiles from models. Steuer et al. (6) used a time dependent stochastic variable as parameter for the external glucose concentration. At a given point in time, the concentrations are recorded. This is repeated n-times to obtain small variations within the same state (6). Camacho and Mendes have a different approach. They apply small random perturbations (90–110%) to the enzyme concentrations that are present in the model, mimicking biological variance and thus obtaining different metabolic profiles (21).

When applying the perturbing methods from Steuer and Camacho, the resulting metabolic profiles within each method are rather similar. To create profiles with larger differences, we have used a different approach. By lowering the V_{max} of the enzymes one at the time to 10% of the original activity, a series of enzyme inhibitory experiments is mimicked. In a total of 18 experiments, the steady state concentrations for 54 metabolites are calculated. In each of these experiments, the V_{max} value of a specific enzyme is lowered to 10% and the model is allowed to reach a steady state. The enzymes are HK(1), GPI(2), PFK(3), ALD(4), TPI(5), GDH(6), PGK(7), PGM(8), ENO(9), PK(10), LDH(11), G6PDH(13), lactonase(14), 6PGDH(15), GSSGR(16), Ru5PE(17), R5PI(18) and TA(25). The numbers

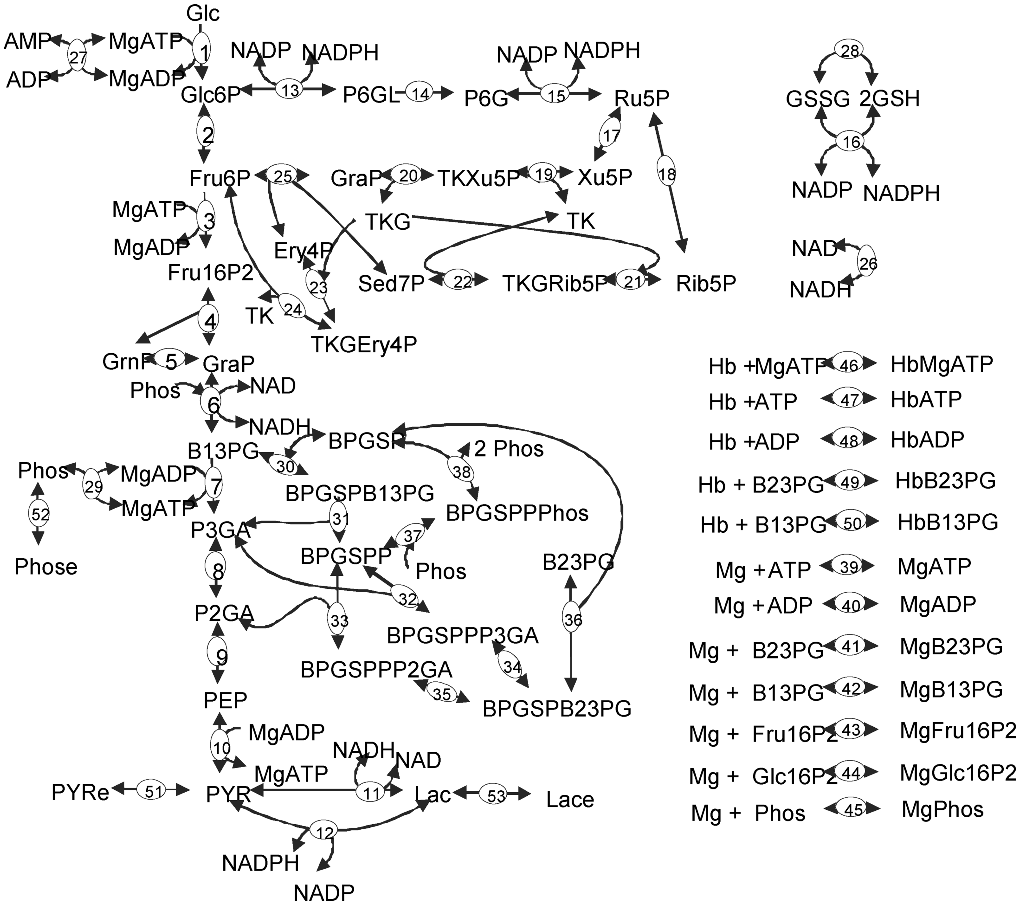

FIG. 2. Scheme of the metabolism of the red blood cell model, taken with permission from Ref. (19).

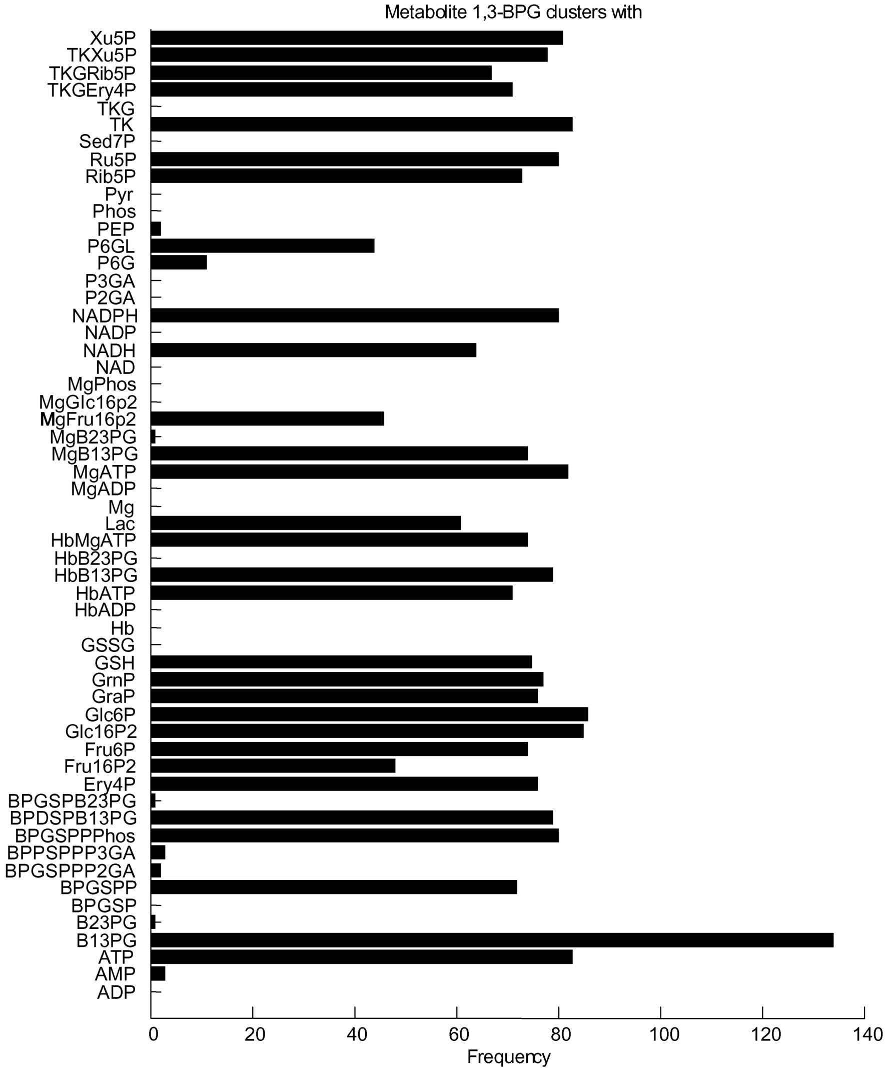

FIG. 3. Histogram of 1,3-BPG metabolite showing the relative degree of clustering with other metabolites.

behind the enzyme abbreviations indicate the reaction that particular enzyme is involved in according to the scheme in Figures 2 and 6. In this way, a total of 18 different metabolite profiles are obtained.

To test the influence of noise on the bagged clustering method, different levels (5, 10 and 20% of the signal intensity) of heteroscedastic, normal distributed noise has been added to the simulated data. Normal distributed, heteroscedastic noise has been chosen since this is a representative type of noise for these types of experiments. The added noise mimics errors origination from biological variance, sampling techniques and analytical error. The resulting data was clustered with bagged K-means clustering as well as ordinary K-means clustering. The result of the clustering was compared to the clustering of the noise free data. For both methods, metabolites that have switched to different clusters are counted. In this way, any changes in clustering can be ascribed to the presence of noise. Because of the random nature of the added noise, it is possible that sometimes the added noise favors one method above another. To counteract these effects, the complete procedure (including the addition of noise to the data) was repeated 100 times, to assess the variability of the procedure. The numbers of metabolites that were clustered differently were plotted in a histogram. Comparison of the histograms of the bagged K-means and the ordinary K-means will reveal any effect of the bagging approach.

Software

The erythrocyte model and the bagged clustering method were programmed in Matlab 7.0. K-means clustering was performed by calculating the Euclidean distance measure on the correlation coefficients between the metabolite concentrations. K-means uses a random starting partition, and depending on this partition the end solution might be different for multiple runs. To overcome ending up with a local minimum, each K-means clustering was repeated with 10 different starting solutions and the solution with the smallest within-cluster distance was kept.

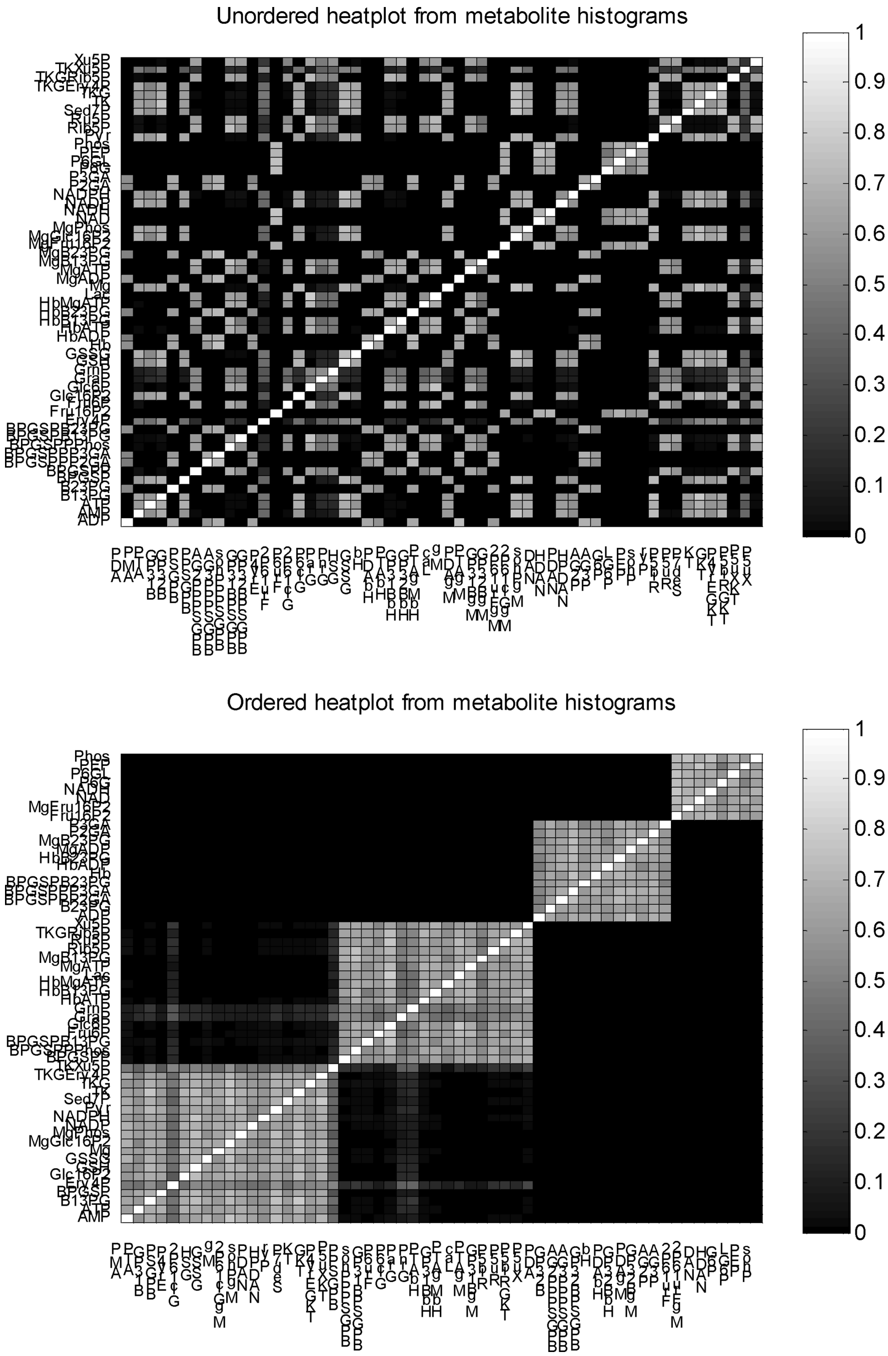

FIG. 4. Heatplot of all histograms. The top figure shows the unordered heatplot. The bottom figure shows the heatplot when it is reordered according to the clustering information of the heatplot.

The total number of bagging samples was 200. All calculations were performed on an AMD Athlon XP 2400 + 2.00 GHz 512 MB RAM personal computer running Windows XP. All Matlab routines in this paper are available at http://www.bdagroup.nl.

RESULTS AND DISCUSSION

When the interest is in the functional relationship between metabolites, clustering on absolute concentrations yields no information since metabolites with low and stable concentrations in all experiments are clustered together without showing any functional relationship. The disadvantages of absolute concentrations are overcome by using the correlation coefficient. Clustering the correlations between metabolites gives us the metabolites that co-vary together and these metabolites are likely to have some functional relationship. An additional advantage of using correlation coefficients is that correlations remain intact

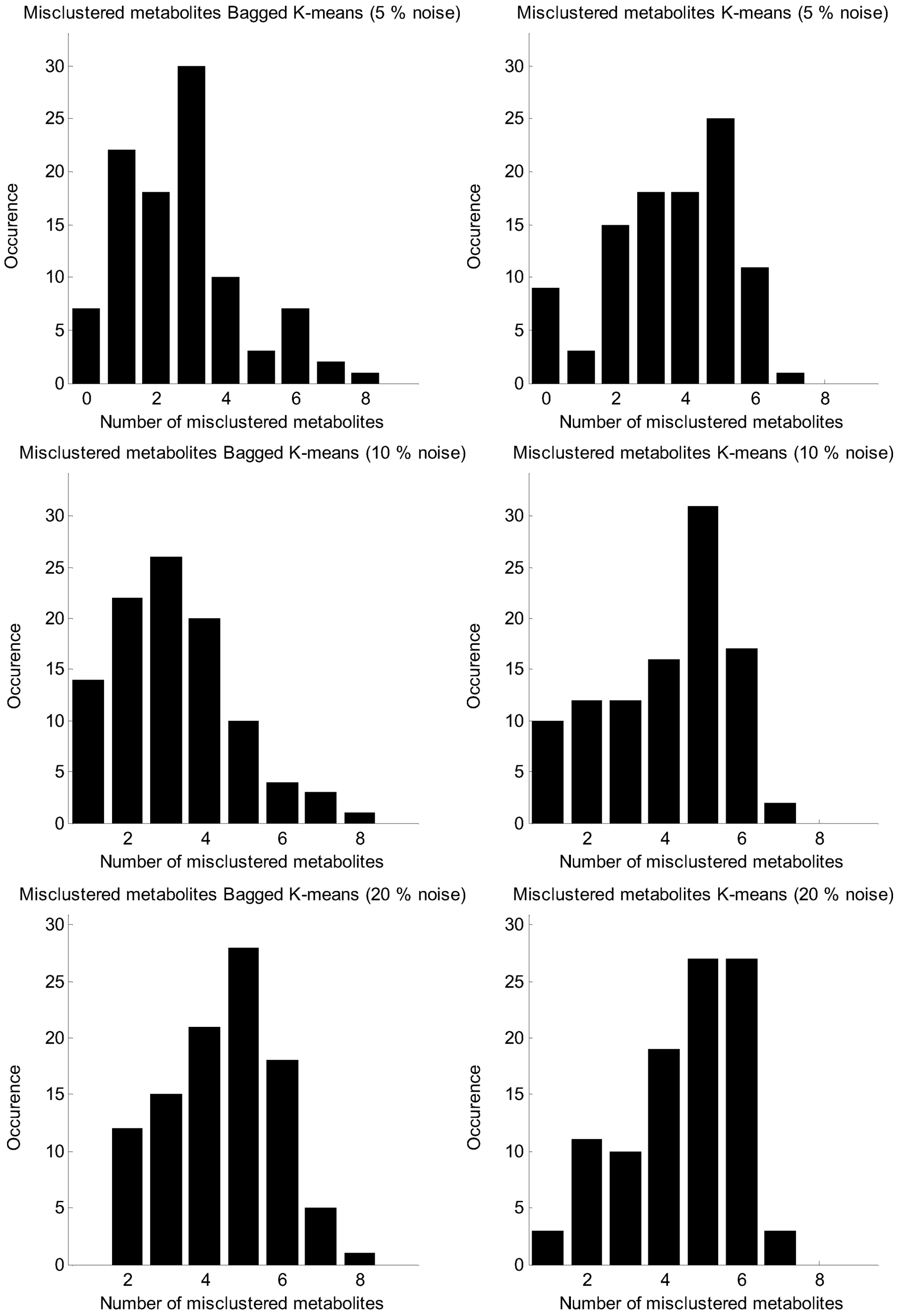

FIG. 5. Distributions of misclassifications for bagged K-means clusterings and ordinary K-means clustering with 5% (top), 10% (middle) and 20% (bottom) noise.

when metabolite profiles originating from different measuring instruments are fused into one large data set. Absolute concentrations can differ due to, for instance, different sensitivities to detectors.

Correlations can either be positive or negative. Camacho et al. state that at least two metabolites belonging to a moiety conserved cycle will at least have a negative correlation (21). To identify metabolites in such a moiety conserved cycle, using the correlation coefficient squared (R^2) can help to identify such metabolites. Also stated by Camacho et al., positive correlations are likely to result from metabolites which are in chemical equilibrium. Negative correlations are not in equilibrium (21). Depending on what is important to see, the correlation coefficient or R^2 will have the preference. In this work the plain correlation coefficient has been used.

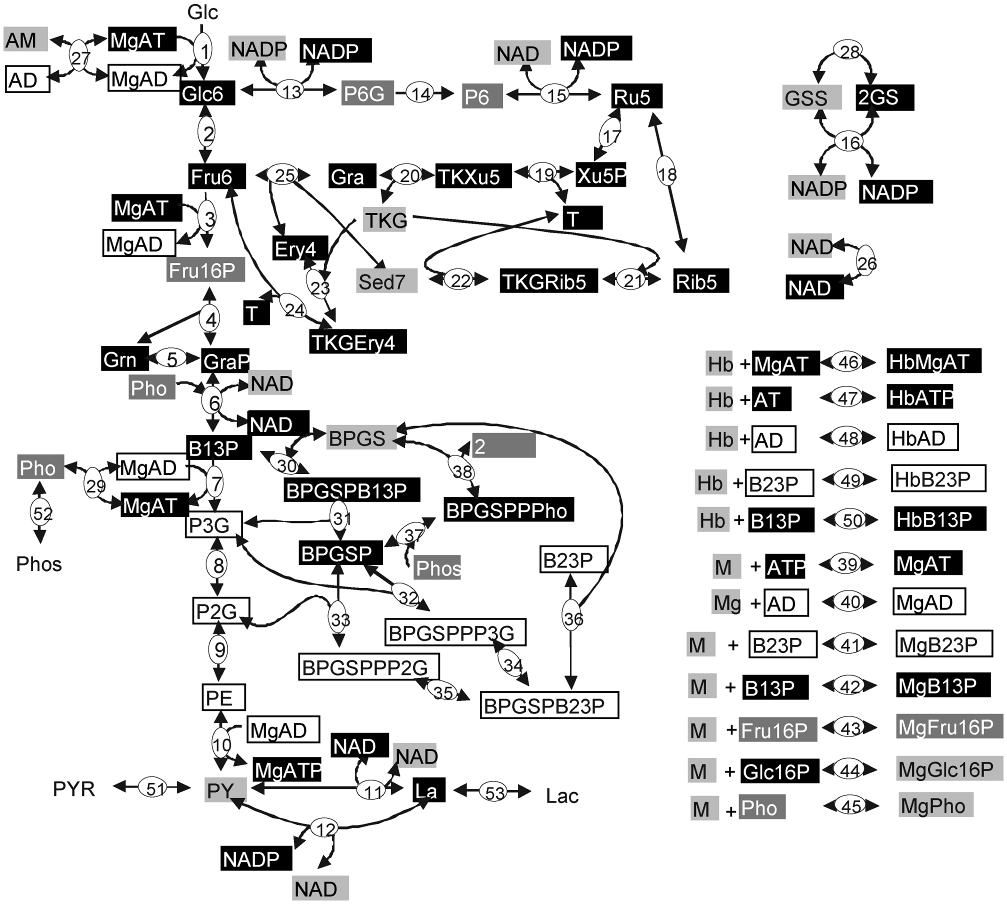

FIG. 6. Overview (taken with permission from Ref. (19)) of reactions used in the simulation together with clustering information. External metabolites (Glc, CO_2, Phose PYRe and Lace) have been kept constant during simulation and did not participate in the clustering.

K-means and bagged K-means clustering are both methods that require a predefined number of clusters. In this case, the number of clusters was set to 4. After 4 clusters the addition of another cluster did no longer decrease the average within-cluster-distance. This indicates that the optimal number of clusters is at 4 clusters (22).

The basis of the method that was described earlier in section 2 is the construction of histograms that collect how often metabolites are clustered together during resampling. All clustering information resulting from all bootstrap samples is collected in these histograms. Figure 3 shows, as an example, the histogram for metabolite 1,3-BPG. It shows the relative degree of clustering of 1,3-BPG with all other metabolites. From this figure, it becomes clear that there are three levels at which metabolites cluster with 1,3-BPG. Some metabolites cluster with a high degree (for instance BPGSPP), some to an intermediate degree (e.g., P6GL), while other metabolites do not, or only to a small extent cluster with 1,3-BPG. The histograms of all the metabolites are collected and displayed as a heatplot (example shown in the top figure of Figure 4). The clustering of the heatplot yields the final clustering. This clustering can also be used to reorder the heatplot to obtain a clearer figure (example shown in bottom figure of Figure 4).

Clustering of metabolites is unsupervised and it is not possible at forehand to know the correct clustering. Thus, comparing how noise affects any clustering can therefore only be done by comparing the ‘noisy’ clustering to the clustering of noise free data. It is important when using K-means clustering to repeat the procedure (with the same data) to be sure that any beneficial effect of the bagging procedure was due to the bagging and not due to random starting solutions.

When dealing with methods involved with (random) noise, it is important to repeat the procedure with different additions of noise. The noise has to be initialized with different random seeds. Because of the nature of the randomness, some noise can favor some methods above others. Figure 5 shows the distribution of the number of metabolites that have changed cluster due to different levels of noise. Clearly it can be seen, that this number is not constant but has some (not normal) distribution. However, comparing the mean value shows the beneficial effects of using bagged K-means clustering in comparison with ordinary K-means clustering. At the 5% noise level, bagged K-means clustering has on average 3 misplaced metabolites, while the ordinary K-means clustering has 5 changed items. This is the same for the 10% noise level. At 20%, the number of misplaced metabolites for bagged K-means clustering is 5, while the ordinary K-means clustering has 5–6 misplaced metabolites. In all cases, the bagged approach yields less misplaced metabolites. When striving for the best results, bagged K-means clustering is able to outperform K-means clustering. At larger noise levels, the beneficial effects of the bagged approach diminish somewhat indicating an upper noise level from where the structure of the data is too much degraded and bagged clustering can no longer cope with the noise levels and improve clustering results.

The result of the clustering of the metabolites (noise free) is shown in Figure 6. Metabolites that are placed in boxes with the same color belong to the same cluster. The structure of the three pathways present in the red blood cell is not reflected in the clusters. Obviously, not all metabolites from one pathway behave identically. This can be caused by participation of metabolites in more that one reaction, or by different types of chemical reactions (e.g., equilibrium reactions). Others have also noted before that the observed pattern of correlations is complex and has no clear connection to the underlying pathway (6). The approach in this paper is likely to find metabolites which are directly connected to each other and form chemical equilibriums, an observation supported by Camacho et al. (21). A clear example of such an equilibrium is the sequence P3GA, P2GA and PEP.

CONCLUSIONS

Clustering of metabolites is an often used tool for analyzing metabolomics data. Correlation coefficients are a useful similarity measure for clustering since they identify co-varying metabolites, which indicates some functional relationship. Clustering results can be deteriorated by the noisy nature of metabolomics data. The use of methods that are not specially suited for noisy data will yield sub-optimal results as is demonstrated in this paper with the computer-generated metabolism of the human red blood cell model. For perturbing this model, a special perturbation scheme was devised in which every enzyme was inhibited to 10% of its original activity. In this way, profiles with very different metabolic concentrations are obtained.

Bagged K-means clustering is a method that, by using resampling techniques, can deal with noise and is capable of more accurately clustering metabolites than ordinary K-means clustering. Using bagged K-means clustering gives a lower misclassification rate compared to ordinary K-means. Validating clustering methods in the presence of different noise levels can be troublesome since the random nature of noise can favor some methods over other. However by repeating the procedure a number of times with different additions of noise this drawback can be overcome.

ACKNOWLEDGMENTS

The authors like to thank Dr. U. Thissen (TNO, Quality of Life, The Netherlands) for contributing to the Matlab implementation of the red blood cell model and Dr. J. Snoep (Vrije Universiteit, Department of Molecular Cell Physiology, The Netherlands) for sharing the scheme used in Figures 2 and 6.

REFERENCES

1. M. J. van der Werf, R. H. Jellema, and T. Hankemeier, Microbial metabolomics: Replacing trial-and-error by the unbiased selection and ranking of targets. *Journal of Industrial Microbiology & Biotechnology* 32 (2005):234–252.
2. D. B. Kell, Metabolomics and systems biology: Making sense of the soup. *Current Opinion in Microbiology* 7 (2004):296–307.
3. O. Fiehn, and W. Weckwerth, Deciphering metabolic networks. *European Journal of Biochemistry* 270 (2003):579–588.
4. B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. d. Jong, P. J. Lewi, and J. Smeyers-Verbeke, *Handbook of chemometrics*, vol. 20B, (Elsevier, Amsterdam, 1998).
5. R. Steuer, J. Kurths, O. Fiehn, and W. Weckwerth, Interpreting correlations in metabolomic networks. *Biochemical Society Transactions* 31 (2003):1476–1478.
6. R. Steuer, J. Kurths, O. Fiehn, and W. Weckwerth, Observing and interpreting correlations in metabolomic networks. *Bioinformatics* 19 (2003):1019–1026.
7. F. Kose, W. Weckwerth, T. Linke, and O. Fiehn, Visualizing plant metabolimic correlation networks using clique-metabolite matrices. *Bioinformatics* 17 (2001):1198–1208.
8. A. Arkin, P. D. Shen, and J. Ross, A test case of correlation metric construction of a reaction pathway from measurements. *Science* 277 (1997):1275–1279.
9. C. B. Clish, E. Davidov, M. Oresic, T. N. Plasterer, G. Lavine, T. Londo, M. Meys, P. Snell, W. Stochaj, A. Adourian, X. Zhang, N. Morel, E. Neumann, E. Verheij, J. Vogels, L. M. Havekes, N. Afeyan, F. Regnier, J. Van Der Greef, and S. Naylor, Integrative biological analysis of the APOE*3-leiden transgenic mouse. *Omics-a Journal of Integrative Biology* 8 (2004):3–13.
10. L. Breiman, Bagging predictors. *Machine Learning*, 24 (1996):123–140.
11. P. J. Mulquiney, W. A. Bubb, and P. W. Kuchel, Model of 2,3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations in vivo kinetic characterization of 2,3-bisphosphoglycerate synthase/phosphatase using c-13 and p-31 nmr. *Biochemical Journal* 342 (1999):567–580.
12. P. J. Mulquiney, and P. W. Kuchel, Model of 2,3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations: Equations and parameter refinement. *Biochemical Journal* 342 (1999):581–596.
13. P. J. Mulquiney, and P. W. Kuchel, Model of 2,3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations: Computer simulation and metabolic control analysis. *Biochemical Journal*, 342 (1999):597–604.
14. S. Dudoit, and J. Fridlyand, Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* 19 (2003):1090–1099.
15. M. K. Kerr, and G. A. Churchill, Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments.

Proceedings of the National Academy of Sciences of the United States of America, 98 (2001):8961–8965.

16. F. Leisch, Bagged clustering. *Technical report, SFB Adaptive Information Systems and Modelling in Economics and Management Science. Vienna University of Economics and Business Administration, http://www.ci.tuwien.ac.at/∼leisch/papers/,* (1999).
17. B. H. Mevik, V. H. Segtnan, and T. Naes, Ensemble methods and partial least squares regression. *Journal of Chemometrics*, 18 (2004):498–507.
18. Y. Raviv and N. Intrator, Bootstrapping with noise: An effective regularization technique. *Connection Science* 8 (1996):355–373.
19. B. G. Olivier, and J. L. Snoep, Web-based kinetic modelling using jws online. *Bioinformatics* 20 (2004):2143–2144.
20. P. J. Mulquiney, and P. W. Kuchel, *Modeling Metabolism with Mathematica: Detailed Examples Including Erythrocyte Metabolism.*, (CRC Press, Boca Raton, FL., 2003).
21. D. Camacho, A. de. la. Fuente, and P. Mendez, The origin of correlations in metabolomics data. *Metabolomics* 1 (2005):53–63.
22. S. Salvador and P. Chan, *Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms*, Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004) 2004.