Analytica Chimica Acta 651 (2009) 173–181

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Metabolomics data exploration guided by prior knowledge

Robert A. van den Berg[a,b,*], Carina M. Rubingh[a], Johan A. Westerhuis[c], Mariët J. van der Werf[a], Age K. Smilde[c]

[a] TNO Quality of Life, P.O. Box 360, 3700 AJ Zeist, The Netherlands
[b] SymBioSys, Katholieke Universiteit Leuven, Tiensestraat 102, 3000 Leuven, Belgium
[c] Biosystems Data Analysis, Universiteit van Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, the Netherlands

ARTICLE INFO

Article history:
Received 7 May 2009
Received in revised form 19 August 2009
Accepted 20 August 2009
Available online 25 August 2009

Keywords:
Metabolomics
Microbiology
Multiblock analysis
Principal component analysis
Canonical correlation analysis

ABSTRACT

In metabolomics research, it is often important to focus the data analysis to specific areas of interest within the metabolome. In this paper, we describe the application of consensus principal component analysis (CPCA) and canonical correlation analysis (CCA) as a means to explore the relation between metabolome data and (i) biochemically related metabolites and (ii) an amino acid biosynthesis pathway. CPCA searches for major trends in the behavior of metabolite concentrations that are in common for the metabolites of interest and the remainder of the metabolome. CCA identifies the strongest correlations between the metabolites of interest and the remainder of the metabolome.

CPCA and CCA were applied to two different microbial metabolomics data sets. The first data set, derived from *Pseudomonas putida* S12, was relatively simple as it contained metabolomes obtained under four environmental conditions only. The second data set, obtained from *Escherichia coli*, was much more complex as it consisted of metabolomes obtained under 28 different environmental conditions. In case of the simple and coherent *P. putida* S12 data set, CCA and CPCA gave similar results as the variation in the subset of the selected metabolites and the remainder of the metabolome was similar.

In contrast, CCA and CPCA yielded different results in case of the *E. coli* data set. With CPCA the trends in the selected subset – the phenylalanine biosynthesis pathway – dominated the results. The main trends were related to high and low phenylalanine productivity, and the metabolites showing a similar behavior in concentration were metabolites regulating the phenylalanine biosynthesis route in the subset and metabolites related to general amino acid metabolism in the remainder of the metabolome. With CCA, neither subset truly dominated the data analysis. CCA described the differences between the wild type and the overproducing strain and the differences between the succinate and glucose grown cells. For the difference between the wild type and the overproducing strain, metabolites from the beginning and the end of aromatic amino acid pathways like erythrose-4-phosphate, tryptophan, and phenylalanine were important for the selected metabolites.

CCA and CPCA proved to be complementary data analysis tools that enable the focusing of the data analysis on groups of metabolites that are of specific interest in relation to the remainder of the metabolome. Compared to an ordinary PCA, focusing the data analysis on biologically relevant metabolites lead especially for the complex *E. coli* data to a better biological interpretation of the data.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Metabolomics research often requires statistical methods to extract information from the large data sets generated. The statistical methods that are presently used vary from unsupervised methods, such as, principal component analysis (PCA) [1,2], or hierarchical clustering [3,4] to supervised approaches like partial least squares (PLS) [5,6] or principal component discriminant analysis (PCDA) [7]. The difference between supervised and unsupervised methods is that for supervised approaches some form of prior knowledge is used to focus or emphasize a specific biological effect of interest. For instance, class information is applied for discriminating between two groups like treated and untreated patients; and the measurements of a phenotype parameter of interest, e.g. productivity, are modeled in regression analysis. Ideally, these analyses reveal which metabolites are the most relevant for the differences between the two classes, or for the behavior of the phenotype parameter.

Currently, data analysis methods that single out groups of metabolites and explore the relation between the behavior of these singled out metabolites and the other metabolites in the data set

* Corresponding author at: SymBioSys, Katholieke Universiteit Leuven, Tiensestraat 102, 3000 Leuven, Belgium. Tel.: +32 16 326256; fax: +32 16 325993.
E-mail address: robert.vandenberg@psy.kuleuven.be (R.A. van den Berg).

0003-2670/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2009.08.029

have not been applied. For example, a specific group of metabolites could contain the measurements of the glycolysis intermediates, or of metabolites that belong to the same class, e.g. amino acids. Consensus PCA-W (CPCA) [8], and canonical correlation analysis (CCA) [9] are methods that can focus the data analysis on the relation between the specified metabolites and the remaining metabolites. Both data analysis methods require the remaining and the specific metabolites to be separated in two data blocks which are subsequently analyzed. Furthermore, the analysis of the two data blocks is symmetric, unlike methods like PLS. That is, both data blocks are treated the same and the order in which the data blocks are treated does not influence the outcome of the analysis.

While CPCA and CCA both allow for the analysis of two data blocks, the principles underlying both methods are rather different. CPCA searches for the largest common trends between behavior of the concentrations of the metabolites of interest and the remaining metabolites. Stated differently, CPCA searches for trends that explain as much of the variation as possible. In contrast, CCA determines the strongest correlation between the behavior of the metabolites in the two data blocks and does not take explained variation into consideration. Both methods thus provide information on the relation between the metabolites of interest and the remaining metabolites; however, they give different views on the underlying biology as will be explained in the next section.

The discussed methods will be illustrated by their application on real life metabolomics data sets. Furthermore the CPCA results are compared with the results of an ordinary PCA in which the metabolites of special interest are not separated from the generic metabolites as a control case to illustrate the differences between the supervised and unsupervised variations of a very similar method. The CCA results are compared with the CPCA results to illustrate the differences between methods that focus on correlation and consensus, respectively.

1.1. Theory

We will first discuss different properties of CPCA and CCA. The following notations will be used: $\mathbf{X}_1$ ($I \times J_1$) a data block that contains the generic metabolome information (i.e. metabolites that are not of special interest), the matrix consists of I experiments and J_1 concentrations of measured metabolites; $\mathbf{X}_2$ ($I \times J_2$) a data block that contains the concentration measurements on the metabolites of specific interest, these metabolites are not in $\mathbf{X}_1$; $\mathbf{X}$ ($I \times (J_1 + J_2)$) is the concatenated matrix of the two data blocks $\mathbf{X}_1$ and $\mathbf{X}_2$, i.e. $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2]$.

1.1.1. CPCA

CPCA [8] searches for the largest trends in the two data sets (Fig. 1) analogous to PCA. It differs from PCA in that the two data blocks are treated differently in order to balance the contributions of both data blocks to the analysis. In a PCA the concatenated data matrix $\mathbf{X}$ is decomposed as follows:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} = \boldsymbol{T}\begin{bmatrix}\boldsymbol{P}_1\\ \boldsymbol{P}_2\end{bmatrix}^T + \boldsymbol{E}. \quad (1)$$

Here, $\mathbf{T}$ ($I \times R$), $\mathbf{P_1}$ ($J_1 \times R$), and $\mathbf{P_2}$ ($J_2 \times R$) are the PCA scores and loadings of a rank R PCA decomposition of $\mathbf{X}$. The scores $\mathbf{T}$ are orthonormal and the loadings $\mathbf{P}$ are orthogonal. This is a minor deviation from common practice in which the scores $\mathbf{T}$ are orthogonal and the loadings $\mathbf{P}$ are orthonormal. By choosing the scores $\mathbf{T}$ to be orthonormal, the explained variation of the sub blocks is fully contained in the block specific loadings. The scores $\mathbf{T}$ and the loadings $\mathbf{P}$ together are referred to as principal components (PCs). The PCA decomposition used in this paper is according to this decomposition.

CPCA ensures that both data blocks are considered equally important in the decomposition by weighting both data blocks.

$$[\, w_1\boldsymbol{X}_1 \quad w_2\boldsymbol{X}_2\,] = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} = \boldsymbol{T}\begin{bmatrix}\boldsymbol{P}_1\\ \boldsymbol{P}_2\end{bmatrix}^T + \boldsymbol{E}. \quad (2)$$

Here, $\mathbf{T}$ ($I \times R$), $\mathbf{P}_1$ ($J_1 \times R$), $\mathbf{P}_2$ ($J_2 \times R$) are the CPCA scores and loadings of a rank R CPCA decomposition of $\mathbf{X}$. The scores $\mathbf{T}$ are orthonormal and the loadings $\mathbf{P}$ are orthogonal. In this study, the weights for a specific block are chosen to be equal to the square root of the sum of squares of that specific data block. Consequently, the total variation in each block becomes '1' after the weighting and both blocks are a priori equally important in terms of explaining variation in the data decomposition. It is also possible to weight the data blocks differently by choosing other block weights [10], for instance, to correct for redundant information in the data blocks [11].

Correcting for differences in total variation in the data blocks is especially important when one data block contains more variables than the other data matrix. In this case, when it is assumed that every metabolite has on average the same variation, it is likely that the data matrix with the most metabolites will be dominant in the data analysis compared to the smallest data matrix. When the data matrices contain a similar amount of metabolites, the effect of block scaling is likely to be minimal (Fig. 1A).

As a consequence of equalizing the sum of squares (SS) for the two data matrices, the SS per metabolite is changed. If the total SS of $\mathbf{X}$ is 100% then after block scaling $\mathbf{X}_1$ and $\mathbf{X}_2$ both contain 50% of the SS. If $\mathbf{X}_2$ is smaller than $\mathbf{X}_1$, the SS is divided over less metabolites and consequently these metabolites individually become more important. As a result, the behavior of the concentrations of the metabolites smallest block can dominate the search for common behavior of metabolite concentrations in $\mathbf{X}_1$ and $\mathbf{X}_2$ (Fig. 1B). In this paper we will show the importance of block weighting when focusing the analysis on a particular data block.

There is also a second aspect that can increase the influence of the block containing the selected metabolites. The concentrations of the selected metabolites can be more correlated than the concentrations of the metabolites in the remainder of the data set. This effect follows from the idea that the selected metabolites share a common biological background. For instance, they are chemically related or share the same regulation. This will make it easier for CPCA to identify main effects based on the data block containing the chosen metabolites. As for a normal PCA, other data pretreatment [12] steps can be taken before block scaling to emphasize different aspects of the data.

1.1.2. CCA

CCA searches for the largest correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$ (Fig. 1). It does this by maximizing:

$$r = \mathrm{corr}(\boldsymbol{X}_1\boldsymbol{a}, \boldsymbol{X}_2\boldsymbol{b}) \quad (3)$$

The vectors $\mathbf{u} = \mathbf{X}_1\mathbf{a}$ and $\mathbf{v} = \mathbf{X}_2\mathbf{b}$ are the so-called canonical variates, and describe the nature of the correlation. The vectors $\mathbf{a}$ and $\mathbf{b}$ are the weights of the contributions of the different metabolites to the found correlation.

Searching for the largest correlation between two data matrices can result in trivial results. First, the largest correlation could be based on the correlation between only one metabolite in each set. While the correlation is very strong, it could be only a minor effect in comparison to the total variation in both matrices. Second, when the data sets consist of more metabolites than experimental conditions it is always possible to find perfect correlations ($r = 1$ or -1), and therefore the solutions will be trivial. It is possible to circumvent these effects by using a dimension reduction technique such as PCA. By reducing the data sets to their main effects, the

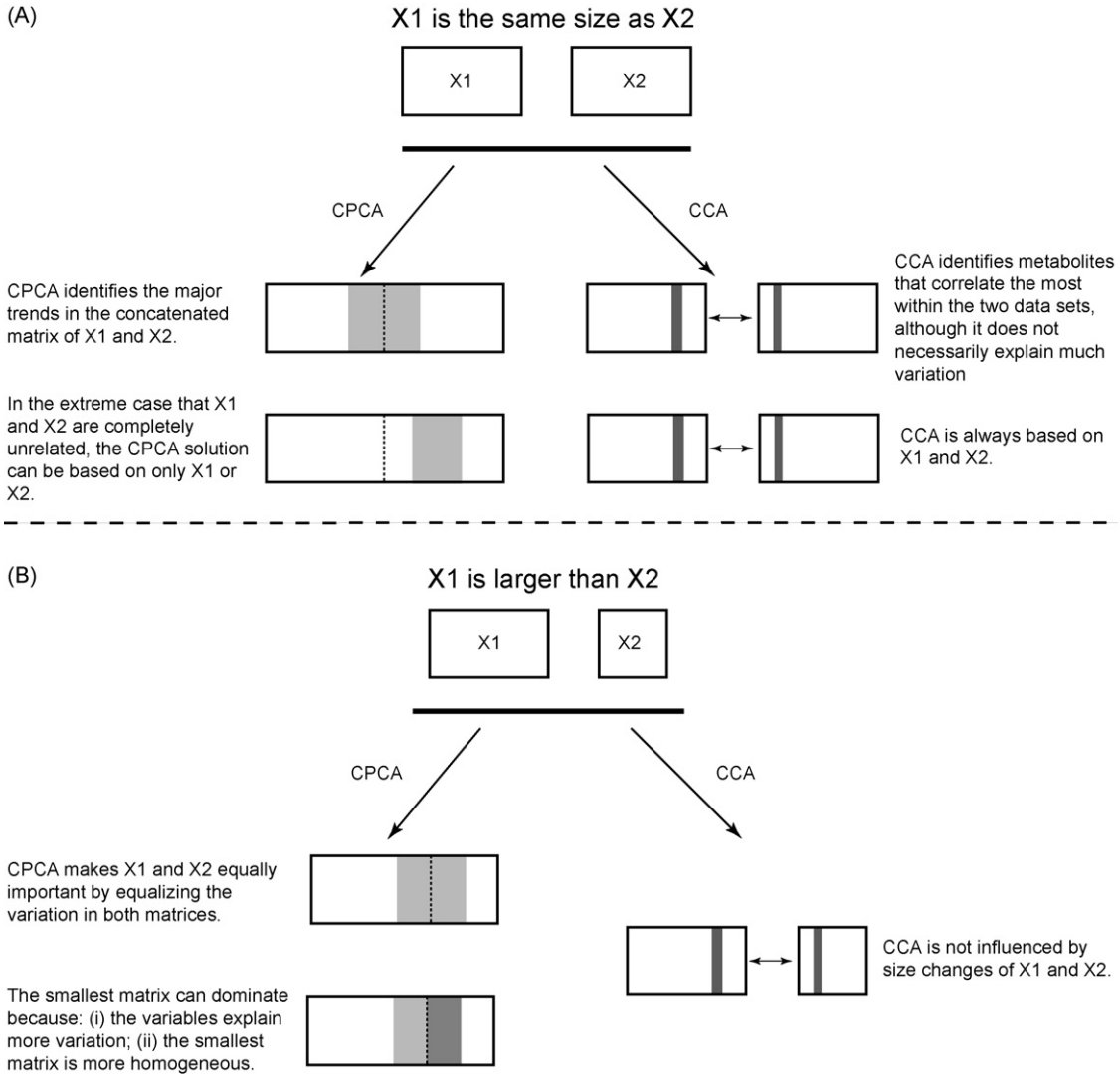


Fig. 1. Comparison of CPCA and CCA. (A) Properties of CPCA and CCA when $\mathbf{X}_1$ and $\mathbf{X}_2$ are equal sized. (B) Properties of CPCA and CCA when $\mathbf{X}_1$ is larger than $\mathbf{X}_2$. The gray areas indicate the linear combinations of the different metabolite concentrations captured by CPCA or CCA. The different shades of gray are a measure of the homogeneity of the variation of the captured metabolites; increasingly darker shades mean that the variation is more homogeneous.

PCs, the effects of single metabolites are limited to contributions to these components. Furthermore, the dimensionality is controlled by the number of components deemed relevant. Therefore the CCA as applied in this paper becomes the maximization of:

$$r = \mathrm{corr}(\boldsymbol{T}_1^* \boldsymbol{a}_1, \boldsymbol{T}_2^* \boldsymbol{b}_2) \quad (4)$$

Here, $\boldsymbol{T}_1^*$ and $\boldsymbol{T}_2^*$ are matrices consisting of the selected PCs from the following PCA decomposition:

$$\boldsymbol{X_k} = \boldsymbol{T_k^* P_k^{*T}}. \quad (5)$$

Here, k indicates the decomposition of the kth data block.

1.1.3. Validation

CPCA and CCA both provide information on the relative importance of every metabolite to the effects discovered by the analysis, namely the weights for each metabolite. These metabolite weights are the starting point for further exploration of the meaning of the results. It is therefore important that a certain degree of confidence of these metabolite weights can be obtained. For this, a validation scheme based on permutations is developed which is generic for CPCA and CCA. Note that this validation does not validate biological relevance.

The significance of every metabolite for the end solution was determined by permuting the values of one metabolite at a time across its sample direction. After the permutation all data analysis steps were performed identical to the unpermuted analysis. The permuted models will be very similar to the unpermuted models, as only one metabolite per model is permuted. The weight obtained for the permuted metabolite in the permuted model is compared with the weight for the unpermuted model. A larger weight in the permuted model indicates that the weight in the unpermuted model is not significant. The permutation is repeated 500 times to obtain a distribution of permuted weights per metabolite, hence in total $500\times (J_1 + J_2)$ permutations were performed. A metabolite weight is considered significant if its unpermuted weight is larger than the permuted weight in 90% or more of the permutations.

The CCA also returns an association measure for the correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$. This measure can also be validated by a permutation approach. In this case, the experiment order of one data matrix is permuted simultaneously for all its metabolites and the resulting association is compared with the association of the unpermuted data. Generally, an association is considered significant if it is in 90% of the permutations larger than the association obtained with permuted data.

2. Materials and methods

2.1. Data

The first data set is obtained from *Pseudomonas putida* S12 (maintained at TNO (Zeist, the Netherlands)) [13] metabolome samples. Cultures of *P. putida* S12 were grown as previously described [14]. In short, samples were obtained in triplicate from cultures grown on four different 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 [15] and LC–MS [16]. The GC–MS and LC–MS data sets were fused together by concatenating the measurement tables [17]. The final data set was manually cleaned up, removing spurious and double entries and consisted of 9 experiments and 161 metabolites. The second data set (*Escherichia coli* NST 74, a phenylalanine overproducing strain, and *E. coli* W3110, the wild type strain) were grown under different experimental conditions as described elsewhere [17]. Samples were analyzed by GC–MS [15] and LC–MS [16] and fused together

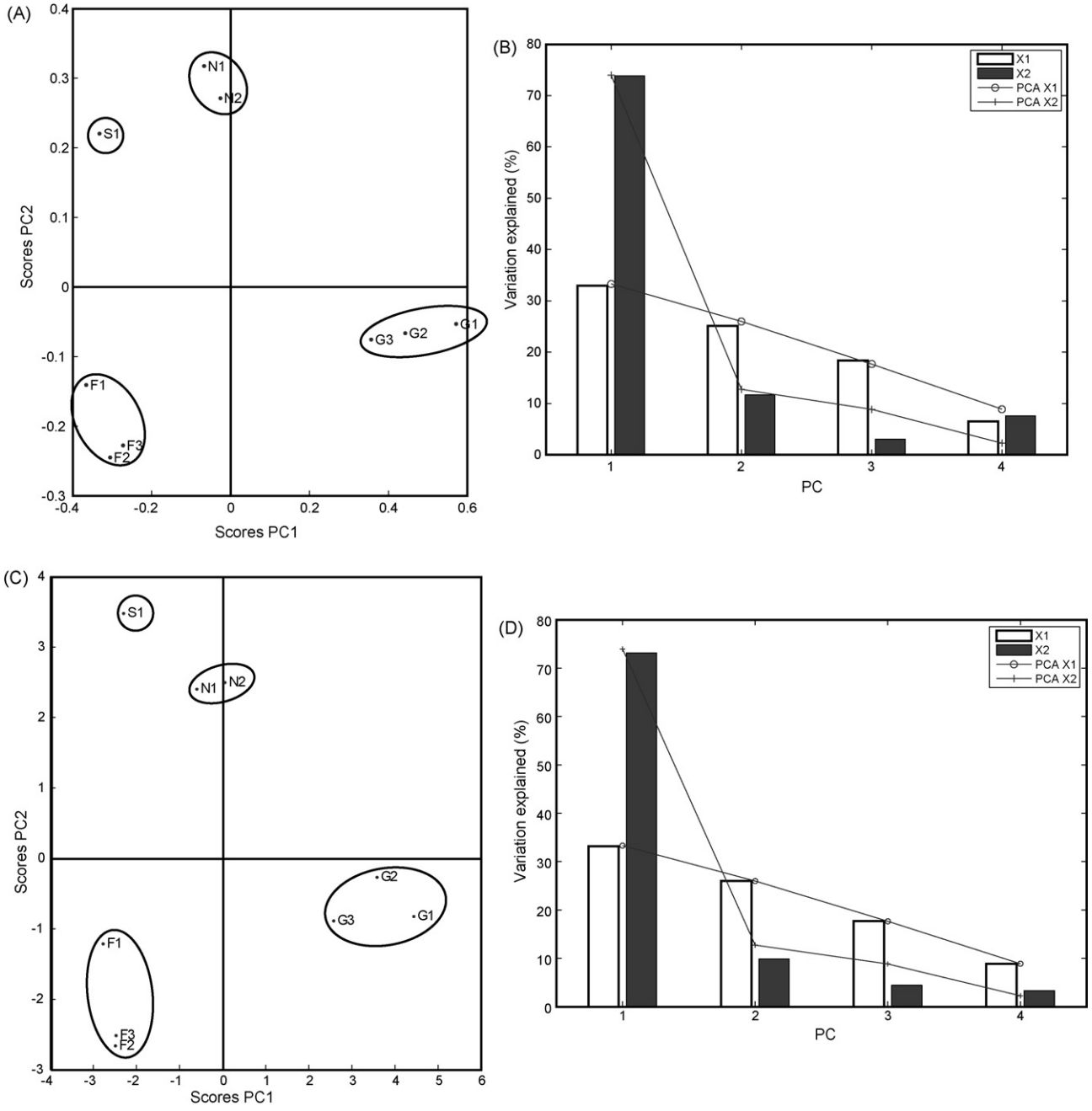


Fig. 2. CPCA and PCA results of the *P. putida* S12 data set. (A) The score plot of the CPCA. The *x*-axis indicates the scores on PC1, the *y*-axis the scores on PC2. The metabolome samples obtained from fermentations with the same carbon source are circled. N, S, G, and F refer respectively to gluconate, succinate, D-glucose, and D-fructose as sole carbon source in the fermentation. (B) The explained variation per data block (white bars for $\mathbf{X}_1$ and gray bars for $\mathbf{X}_2$) and the maximal explained variation for that data block (lines with 'O' for $\mathbf{X}_1$ and with '+' for $\mathbf{X}_2$) for CPCA. On the *x*-axis the calculated PCs are given and on the *y*-axis the amount of explained variation as a percentage of the total variation. (C) The score plot of the PCA. The *x*-axis indicates the scores on PC1, the *y*-axis the scores on PC2. The metabolome samples obtained from fermentations with the same carbon source are circled. N, S, G, and F refer respectively to gluconate, succinate, D-glucose, and D-fructose as sole carbon source in the fermentation. (D) The explained variation per data block (white bars for $\mathbf{X}_1$ and gray bars for $\mathbf{X}_2$) and the maximal explained variation for that data block (lines with 'O' for $\mathbf{X}_1$ and with '+' for $\mathbf{X}_2$) for PCA. On the *x*-axis the calculated PCs are given and on the *y*-axis the amount of explained variation as a percentage of the total variation.

[17]. The final data set was manually cleaned up, removing spurious and double entries and consisted of 28 experiments and 188 metabolites.

2.2. Data analysis

CPCA [8] and CCA [9] were implemented in Matlab 7.3.0 (The Mathworks, Inc.). In the data analysis, the data was range scaled [12]. The significance of the data analysis results was validated as described in Section 1.1.3.

3. Results

CPCA and CCA were illustrated by their application on two different metabolomics data sets. The first data set consisted of metabolomes obtained from *P. putida* S12 fermentations in which *P. putida* S12 was grown on four different carbon sources. The $\mathbf{X}_2$ matrix (9 experiments, 19 metabolites) contained the concentrations of the measured nucleotides and the $\mathbf{X}_1$ matrix (9 experiments, 142 metabolites) contained the metabolome minus the nucleotides. The nucleotides were chosen as they are closely related in the metabolic network and nucleotides are cellular energy carriers that could respond to the chosen experimental conditions. This data set proved to be a straightforward data set with large effects induced by the selected experimental conditions. The second data set consisted of *E. coli* metabolomes obtained from cells cultivated under 28 different experimental conditions. $\mathbf{X}_2$ (28 experiments, 13 metabolites) contained all the measured intermediates of the phenylalanine biosynthesis pathway and $\mathbf{X}_1$ (28 experiments, 175 metabolites) contained the remaining metabolome. The phenylalanine biosynthesis pathway was selected as the experimental conditions aimed to induce variation in phenylalanine production. This data set is a complex data set in which different effects play a role, like the environmental conditions and different growth phases in the batch process. None of concentrations of the metabolites were simultaneously in $\mathbf{X}_1$ and $\mathbf{X}_2$ to avoid trivial results.

3.1. CPCA

The CPCA analysis of the combined metabolome/nucleotide matrix from the *P. putida* S12 data set lead to a clear separation of the metabolomes resulting from growth on the four carbon sources on the first two PCs (Fig. 2A). The metabolites that contribute to the first component were for $\mathbf{X}_1$ metabolites related to the carbon catabolism pathways and to central metabolism, such as, glyceraldehyde-3-phosphate, dihydroxyacetone-phosphate, glucose-6-phosphate, and pyruvate (Appendix). For $\mathbf{X}_2$ most metabolites contributed significantly to the first component. It is noteworthy that the mono-phosphate (xMPs) and di-phosphate (xDPs) nucleotides had a positive contribution, while the tri-phosphate nucleotides (xTPs) had a negative contribution. This suggests a negative correlation between the behavior of the high energy nucleotides (xTPs) and the low energy nucleotides (xMPs and xDPs). Therefore, this could point towards a difference in cell energy availability for the samples separated by the first component: succinate and D-fructose on the left and D-glucose on the right.

The variation explained for each $\mathbf{X}$ sub matrix was compared with the maximal explained variation possible for that matrix (Fig. 2B). Both $\mathbf{X}$ sub matrices are very close to the maximal explained variation for the first two PCs. This indicated that these PCs indeed described a common direction in $\mathbf{X}_1$ and $\mathbf{X}_2$. After the second PC, the variation in $\mathbf{X}_1$ remained maximally explained, while the variation $\mathbf{X}_2$ was not maximally explained in PC 3. Here, the differences in variation between the two blocks became visible.

Table 1
Phenylalanine concentration in *E. coli* metabolomics samples. The phenylalanine concentrations are sorted in descending order. The experimental conditions under which the metabolome samples were obtained are presented elsewhere [15].

Sample	Concentration (nmol/mg dry weight)
6.3	4.17
10.3	3.38
2.2	3.19
10.4	2.96
6.2	2.91
4.5	2.81
10.2	2.71
4.3	2.65
1.4	2.62
1.3	2.57
7.4	2.35
5.4	2.25
7.3	1.86
5.3	1.61
6.1	1.43
4.2	1.33
1.2	1.08
4.1	0.70
3.3	0.63
7.2	0.37
1.1	0.28
5.1	0.26
10.1	0.23
5.2	0.19
9.4	0.093
9.3	0.051
9.1	0.015
9.2	0.010

A PCA was performed on the combined metabolome/nucleotide matrix from the *P. putida* S12 data set. The results of the PCA of the *P. putida* S12 data are similar to the CPCA. The score plot (Fig. 2C) shows a similar grouping of the metabolomes resulting from growth on the four carbon sources as for CPCA. Also the explained variation per $\mathbf{X}$ sub matrix (Fig. 2D) is highly similar, albeit in the PCA the explained variation of $\mathbf{X}_2$ is slightly less than the maximal explained variation. Here, the CPCA results remained closer to the maximal explained variation of $\mathbf{X}_2$ (Fig. 2B), although these differences are small. The ranking of the most important metabolites (Appendix) for the first component of the PCA is highly similar to the ranking of the CPCA. The loadings of CPCA and PCA start to deviate for the second component and further.

For the more complex *E. coli* data set, the score plots of the first PC of the CPCA analysis (Fig. 3A) showed an effect related to high and low phenylalanine productivity (Table 1) in the first PC. The most important metabolites relating to this effect were for $\mathbf{X}_1$ phenyllactate, 3,5-dihydroxypentanoate (tentatively identified), a number of unidentified metabolites, and the amino acids valine and isoleucine (Appendix). In some organisms phenyllactate is synthesized by (R)-4-hydroxyphenyllactate dehydrogenase [18] with phenylpyruvate as substrate as an alternative end point of the phenylalanine pathway [19], however, this enzyme activity has not been described previously for *E. coli* [19,20]. In phenylalanine biosynthesis, phenylpyruvate is converted in phenylalanine as the final step of the pathway. Limiting or eliminating the production of phenyllactate could thus potentially improve phenylalanine production. For $\mathbf{X}_2$, the most important metabolites were phenylalanine and metabolites that are regulatory important [20] in the phenylalanine biosynthesis route, such as, chorismate and erythrose-4-phosphate. The enzymes that convert chorismate and erythrose-4-phosphate are subject to end product inhibition by phenylalanine [20].

For the second PC there was not a clear explanation for the behavior of the experimental conditions. The most important

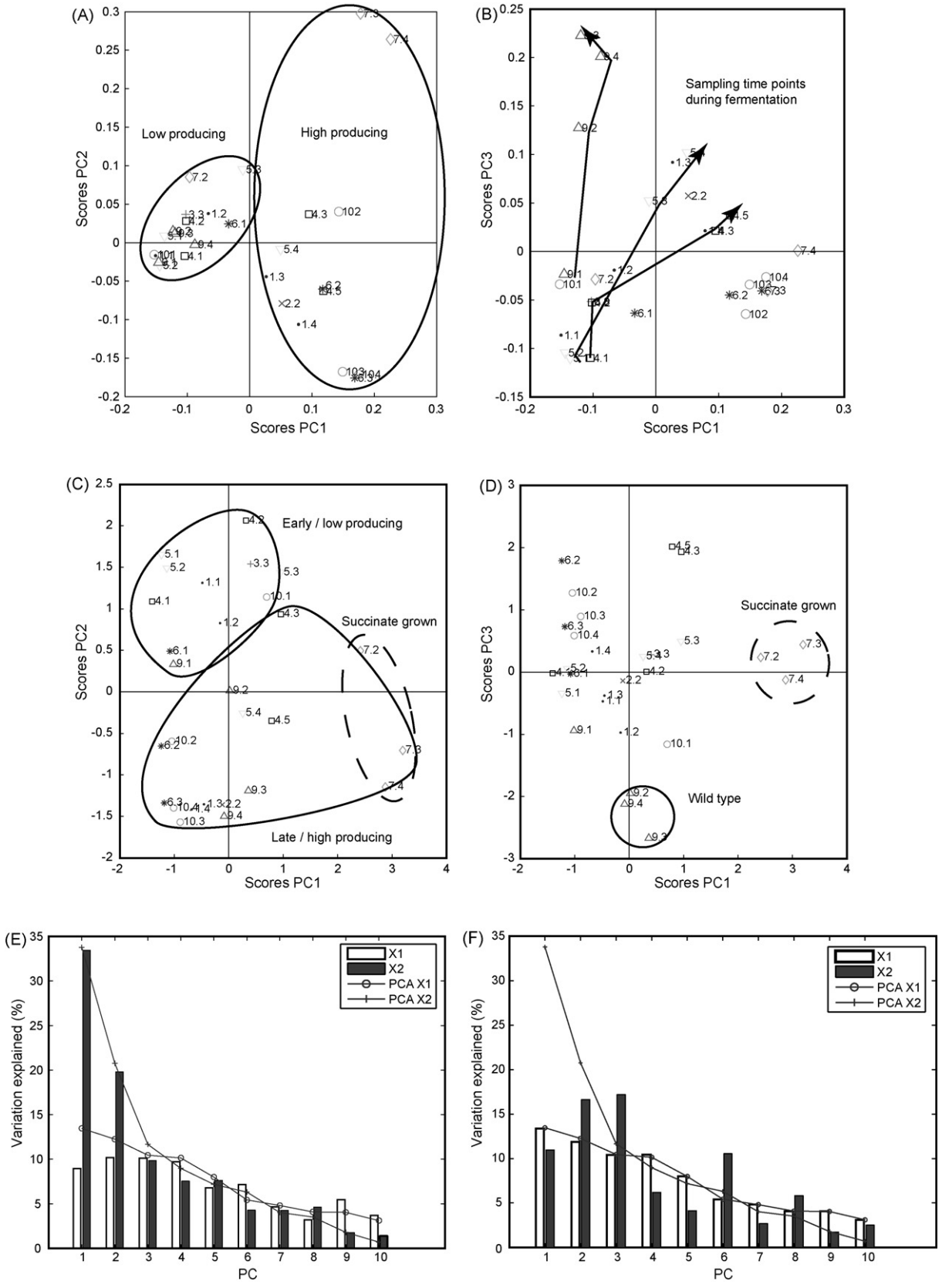


Fig. 3. CPCA results of the E. coli data set. (A) The score plot of the CPCA for PC 1 and 2. The *x*-axis indicates the scores on PC1, the *y*-axis the scores on PC2. The numbers in the figure refer to experimental conditions. The circles indicate the difference between experimental conditions that resulted in high and low phenylalanine production (Table 1). (B) The score plot of the CPCA for PC 1 and 3. The *x*-axis indicates the scores on PC1, the *y*-axis the scores on PC3. The numbers in the figure refer to experimental conditions. The arrows indicate the order of sampling from early to late time points during the batch fermentation. (C) The score plot of the PCA for PC 1 and 2. The *x*-axis indicates the scores on PC1, the *y*-axis the scores on PC2. The numbers in the figure refer to experimental conditions. The circles indicate the different subgroups based on early sampling/low phenylalanine production; late/high phenylalanine production; and the succinate grown cells (Table 1). (D) The score plot of the PCA for PC 1 and 3.

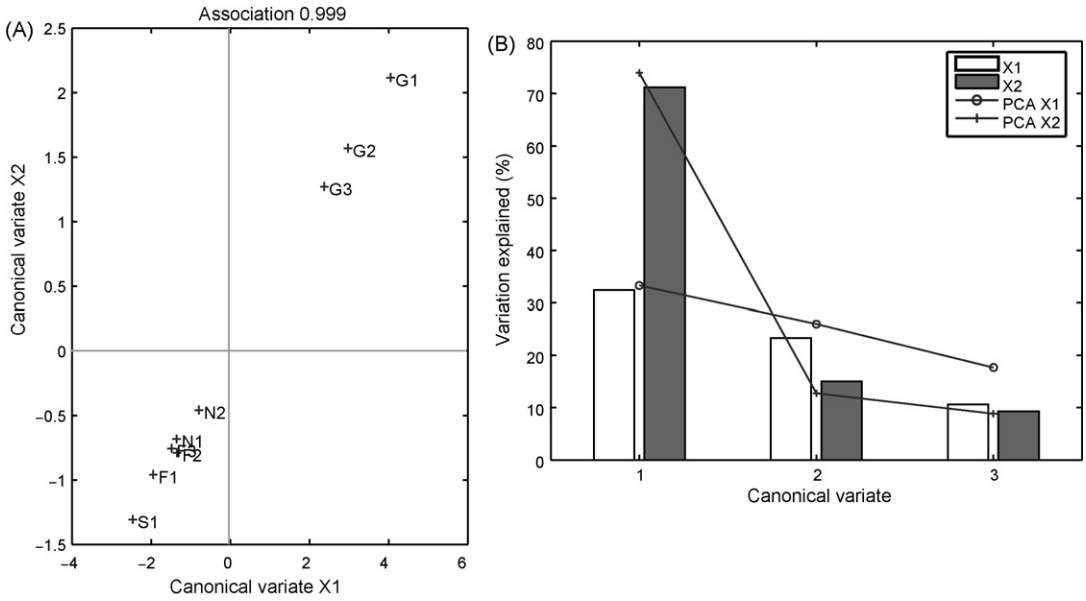


Fig. 4. CCA results for the *P. putida* S12 data set. (A) The nature of the correlation of canonical variate 1. The association measure represents the strength of the correlation. On the *x*-axis the canonical variate of $\mathbf{X}_1$ is given and on the *y*-axis the canonical variate of $\mathbf{X}_2$. (B) The explained variation per data block (white bars for $\mathbf{X}_1$ and gray bars for $\mathbf{X}_2$) and the maximal explained variation for that data block (lines with 'O' for $\mathbf{X}_1$ and with '+' for $\mathbf{X}_2$). On the *x*-axis the calculated canonical variates are given and on the *y*-axis the amount of explained variation as a percentage of the total variation.

metabolites for this PC, however, suggested that the PC was related to general amino acid metabolism. The most important metabolites were for $\mathbf{X}_1$ urea, aspartate, malate, and fumarate: these metabolites are part of the citric acid cycle and the urea cycle. Also the amino acids isoleucine and valine were important for this PC as well as for the first PC. For $\mathbf{X}_2$, important metabolites were glutamate and ketoglutarate, used in amino group transfer reactions; tyrosine and tryptophan, end products of the other branch of the aromatic amino acid biosynthesis pathway; and phenylpyruvate, the precursor to phenylalanine.

The third PC seemed to describe a time effect as is indicated with arrows for fermentations 4, 5, and 9 (Fig. 3B). The most important metabolites for $\mathbf{X}_1$ consisted of unidentified metabolites, UDP-N-AAGDAA, and guanine. For $\mathbf{X}_2$ only prephenate was significant. Although UDP-N-AAGDAA is part of the peptidoglycan biosynthesis pathway and thus related to cell wall synthesis, it is unfortunately difficult to speculate about biological processes behind this time effect due to the large number of unidentified metabolites in $\mathbf{X}_1$.

The comparison of the explained variation per X block with the maximal explained variation for that X block showed that the CPCA analysis seemed to depend most on $\mathbf{X}_2$. The explained variance of $\mathbf{X}_2$ in the solution closely followed the maximal explained variation (Fig. 3E), while this is not the case for $\mathbf{X}_1$. This can be caused by two factors; first, $\mathbf{X}_2$ contains much less metabolite concentrations than $\mathbf{X}_1$, and second, $\mathbf{X}_2$ is more homogeneous than $\mathbf{X}_1$ because the selected metabolites are part of the same pathway.

A comparison of the score plots of CPCA (Fig. 3A and B) with the score plots of the PCA (Fig. 3C and D) reveals that the PCA decomposition differs from the CPCA decomposition. The first two components of the PCA seem to describe a mixed effect of fermentation time and phenylalanine production. On the top left of the score plot of components 1 and 2 (Fig. 3C) the samples pertaining to early time points and low phenylalanine production are clustered, while the bottom of the figure contained the samples pertaining to late time points and/or high phenylalanine productivity. Furthermore, the samples of succinate grown cells form a cluster on the right of the figure. This cluster is indicated by the dashed lines.

The interpretation of components 1 and 3 of the PCA (Fig. 3D) is less clear than the interpretation of components 1 and 2. A clear cluster of experiments that could be identified consisted of the samples pertaining to the experiments with wild type *E. coli* strains. Also, the cluster of samples originating from succinate grown cells is still present.

When the variation explained per sub matrix by the PCA decomposition is compared with the maximal explained variation for the sub matrices (Fig. 3F), it is clear that $\mathbf{X}_1$, the largest data block, dominates the analysis. The variation explained in $\mathbf{X}_1$ is very close to the maximal explained variation for this block. On the other hand, the variation explained in $\mathbf{X}_2$ is far from the maximal explained variation. This result is in contrast to the CPCA results, which are dominated by $\mathbf{X}_2$.

Not surprisingly given the results already presented, the ranking of the most important metabolites for both data blocks for the PCA analysis of the *E. coli* data differ strongly from the CPCA (Appendix).

3.2. CCA

CCA searches for the largest correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$. For the *P. putida* S12 data set of both $\mathbf{X}_1$ and $\mathbf{X}_2$ the dimensions were reduced by PCA after range scaling. For $\mathbf{X}_1$ four and for $\mathbf{X}_2$ three PCs were used. The correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$ was very large – all the experiments are on the diagonal line – and the significant association is 0.999 (Fig. 4). This value for the association was significant after validation by permutation of the experimental conditions and repetition of the data analysis. The metabolites responsible for this large correlation

The *x*-axis indicates the scores on PC1, the *y*-axis the scores on PC3. The numbers in the figure refer to experimental conditions. The wild type and succinate grown samples are indicated by circles. (E) The explained variation per data block (white bars for $\mathbf{X}_1$ and gray bars for $\mathbf{X}_2$) and the maximal explained variation for that data block (lines with 'O' for $\mathbf{X}_1$ and with '+' for $\mathbf{X}_2$) for CPCA. On the *x*-axis the calculated PCs are given and on the *y*-axis the amount of explained variation as a percentage of the total variation. (F) The explained variation per data block (white bars for $\mathbf{X}_1$ and gray bars for $\mathbf{X}_2$) and the maximal explained variation for that data block (lines with 'O' for $\mathbf{X}_1$ and with '+' for $\mathbf{X}_2$) for PCA. On the *x*-axis the calculated PCs are given and on the *y*-axis the amount of explained variation as a percentage of the total variation.

were for $\mathbf{X}_1$ metabolites related to catabolic pathways, such as, glyceraldehyde-3-phosphate, dihydroxyacetone-phosphate, and glucose-6-phosphate (Appendix). This was similar to the CPCA results. The responsible metabolites for $\mathbf{X}_2$ were the xMPs and the xTPs. Unlike the CPCA results, the xDPs were less important. For both data sets, the variation modeled by the correlation between the two matrices was close to the maximal explained variation for those matrices. This indicated that the behavior of the metabolite concentrations in $\mathbf{X}_1$ and $\mathbf{X}_2$ correlates very well and that the correlation is a major effect in the behavior of these concentrations.

CCA on the *E. coli* data set identified also a strong correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$ with a significant association of 0.981 (Fig. 5A). The order of the experiments in the correlation plot for the first canonical variate seemed related to the difference between the wild type strain and the high producing strains. This effect was not as strong as for the CPCA analysis. For instance, condition 6.3, that lead to the highest phenylalanine production (Table 1) was close to zero in Fig. 5A, and thus not important for canonical variate 1. Unfortunately, the metabolites of $\mathbf{X}_1$ that contributed most to this correlation were unidentified metabolites (phenyllactate can be found at the 10th position), for $\mathbf{X}_2$, metabolites were phenylalanine 3-dehydroquinate, tryptophan, and erythrose-4-phosphate (Appendix). The second largest correlation between the two data sets was still large with a significant association of 0.966. Here the fermentations on succinate as a carbon source stood out (Fig. 5B). In $\mathbf{X}_1$, the metabolites urea, isoleucine, malate, fumarate, and aspartate were important; this is similar as the results for the second PC in the CPCA analysis. However, slightly different metabolites in $\mathbf{X}_2$ were important, shikimate, phenylalanine, phosphoenolpyruvate, ketoglutarate, glutamate, and phenylpyruvate. The explained variance for the correlation was not following the maximal explained variance for both *E. coli* data matrices (Fig. 5C) as closely as for the *P. putida* S12 data set. This means that for these two matrices the directions that correlate the best were not the most dominant directions in the separate matrices.

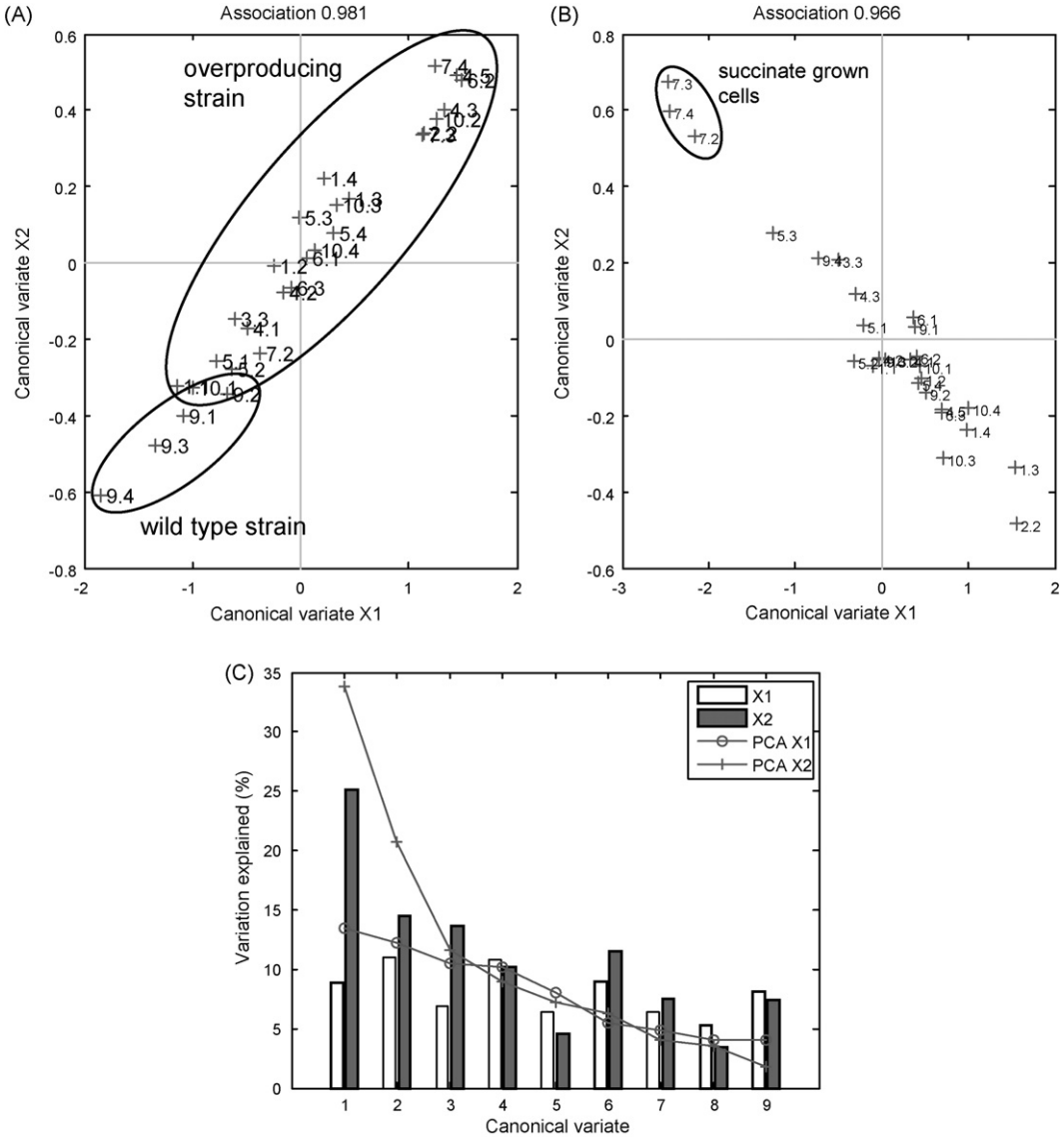


Fig. 5. CCA results for the *E. coli* data set. (A) The nature of the correlation of canonical variate 1. The association measure represents the strength of the correlation. On the *x*-axis the canonical variate of $\mathbf{X}_1$ is given and on the *y*-axis the canonical variate of $\mathbf{X}_2$. The ovals indicate grouping of the metabolomes resulting from fermentations with wild type and the overproducing strain. (B) The nature of the correlation of canonical variate 2. The association measure represents the strength of the correlation. On the *x*-axis the canonical variate of $\mathbf{X}_1$ is given and on the *y*-axis the canonical variate of $\mathbf{X}_2$. The metabolomes obtained from succinate grown cells are indicated with a circle. (C) The explained variation per data block (white bars for $\mathbf{X}_1$ and gray bars for $\mathbf{X}_2$) and the maximal explained variation for that data block (lines with O for $\mathbf{X}_1$ and with + for $\mathbf{X}_2$). On the *x*-axis the calculated canonical variates are given and on the *y*-axis the amount of explained variation as a percentage of the total variation.

4. Discussion

CPCA and CCA are valuable methods to emphasize specific areas of the metabolic network in analysis of metabolomics data. They make it possible to focus on groups of metabolites be it functionally or chemically related metabolites as for the *P. putida* S12 data, or a metabolic pathway as for the *E. coli* data set; both methods result in biological meaningful results. Compared to a normal PCA, these methods were able to focus the analysis on a biologically relevant subset of the metabolome. This focus lead, especially for the complicated *E. coli* data to an improved biological interpretation of the data.

CPCA and CCA address different biological and data analysis questions. CPCA searches for the direction that explains most of the variation in the weighted and concatenated matrices. When the variation within and between both data sets shows similar major trends, the variation described will closely resemble the maximal variation explained for both data sets, as was the case for the *P. putida* S12 data sets (Fig. 2). On the other hand, when variation in the two data sets is not similar, CPCA will still identify the largest variation in the concatenated data set and this direction can be dominated by one matrix; as proved the case for the *E. coli* data set (Fig. 3).

CCA is not consensus based; it retains the nature of the matrices and identifies the largest correlation between the two data sets. Due to the PCA step performed before the CCA analysis, CCA focused on large trends in variation in the matrices. The results of CCA for a data set with a simple structure and coherent behavior, like the *P. putida* S12 data set, was similar to the CPCA analysis (Fig. 4).

The difference between the two methods becomes clear from the analysis of the *E. coli* data set. As a consequence of the complex nature of the data set, there is no common dominant variation in both $\mathbf{X}_1$ and $\mathbf{X}_2$ and the CPCA became dominated by $\mathbf{X}_2$ that contained the measured intermediates of the phenylalanine pathway (Fig. 3C). In contrast, CCA identifies the largest correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$ even though this direction is not dominant in either $\mathbf{X}_1$ or $\mathbf{X}_2$ (Fig. 5C).

Based on the biological question to be answered CPCA is better suited for identifying large common effects between the metabolome and the specified metabolites. CCA searches those trends in the two data sets that correlate the strongest, without compromising towards major trends.

5. Conclusion

In this paper, we include knowledge of metabolic pathways and chemical relatedness to focus the data analysis. This opened up the possibility to study the behavior of these metabolites in more detail than with an unsupervised method. Besides applications in metabolomics, these methods can also be applied for the comparison of, for instance, metabolomics and transcriptomics, or proteomics data.

Acknowledgments

This research was funded by the Kluyver Centre for Genomics of Industrial Fermentation, which is supported by the Netherlands Genomics Initiative (NROG).

Appendix. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.aca.2009.08.029.

References

[1] J.E. Jackson, A User's Guide to Principal Components, John Wiley & Sons, Inc., 1991.
[2] I.T. Jolliffe, Principal Component Analysis, second ed., Springer-Verlag, New York, 2002.
[3] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley-Interscience, 1990.
[4] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 14863–14868.
[5] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1–17.
[6] A. Hoskuldsson, J. Chemom. 2 (1988) 211–228.
[7] R. Hoogerbrugge, S.J. Willig, P.G. Kistemaker, Anal. Chem. 55 (1983) 1710–1712.
[8] A.K. Smilde, J.A. Westerhuis, S. de Jong, J. Chemom. 17 (2003) 323–337.
[9] W.J. Krzanowski, Principles of Multivariate Analysis, a User's Perspective, Oxford University Press Inc., New York, 1988.
[10] K. Van Deun, A.K. Smilde, M.J. van der Werf, H.A.L. Kiers, I. Van Mechelen, BMC Bioinformatics 10 (2009) 246.
[11] J. Pagès, Food Qual. Pref. 16 (2005) 642–649.
[12] R.A. van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, BMC Genomics 7 (2006) 142.
[13] S. Hartmans, M.J. van der Werf, J.A.M. de Bont, Appl. Environ. Microbiol. 56 (1990) 1347–1351.
[14] M.J. van der Werf, K.M. Overkamp, B. Muilwijk, M. Koek, Bianca, R.H. Jellema, L. Coulier, T. Hankemeier, Mol. Biosyst. 4 (2008) 315–327.
[15] M. Koek, B. Muilwijk, M.J. van der Werf, T. Hankemeier, Anal. Chem. 78 (2006) 1272–1281.
[16] L. Coulier, R. Bas, S. Jespersen, E. Verheij, M.J. van der Werf, T. Hankemeier, Anal. Chem. 78 (2006) 6573–6582.
[17] A.K. Smilde, M.J. van der Werf, S. Bijlsma, B.J.C. van der Werff-van der Vat, R.H. Jellema, Anal. Chem. 77 (2005) 6729–6736.
[18] A. Chang, M. Scheer, A. Grote, I. Schomburg, D. Schomburg, Nucleic Acids Res. 37 (2009) D588–D592.
[19] M. Kanehisa, S. Goto, Nucleic Acids Res. 28 (2000) 27–30.
[20] I.M. Keseler, J.C. Vides, S.G. Castro, J.L. Ingraham, S. Paley, I.T. Paulsen, M.P. Gil, P.D. Karp, Nucleic Acids Res. 33 (2005) D334–D337.