Multi-way analysis of flux distributions across multiple conditions

Maikel P. H. Verouden[a], Richard A. Notebaart[b], Johan A. Westerhuis[a]*, Mariët J. van der Werf[c], Bas Teusink[d] and Age K. Smilde[a]

With the availability of genome sequences of many organisms and information about gene-protein-reaction (GPR) associations with respect to these organisms genome-scale metabolic networks can be reconstructed. In cellular systems biology these networks are used to model the behavior of metabolism in context of cell growth in terms of fluxes (reaction rates) through reactions in the network. Because the flux through each reaction can generally vary within a range, many flux distributions of the entire network are possible. However, since reactions are connected by common metabolites, reactions that are functionally coherent, are expected to highly correlate in terms of their flux value over different flux distributions.

In this paper the genome-scale network of a lactic acid bacterium, named *Lactococcus lactis* MG1363, is used to generate flux distributions for multiple *in silico* environmental conditions, mimicking laboratory growth conditions. The flux distributions per condition are used to calculate a correlation matrix for each condition. Subsequently the correlations between the reactions are analyzed in a multivariate approach across the *in silico* environmental conditions in order to identify correlations that are invariant (i.e. independent of the environment) and correlations that are variant across conditions (i.e. dependent of the environment). The applied multivariate methods are Parallel Factor Analysis (PARAFAC) and Principal Component Analysis (PCA). The discussion of the results of both methods leads to the question whether latent variable models are suitable analyzing this type of data. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: PARAFAC; PCA; correlation; flux distributions; genome-scale network

1. INTRODUCTION

The field of molecular cell biology experienced rapid progress from the moment that methods became available to sequence the genomes of organisms. The genome represents all the heritage material of organisms. Parts of the genome are referred to as genes and code for proteins which carry out all kinds of functions within the cell. A very important functionality of the cell is metabolism. This process allows for the uptake of nutrients from the environment and converts them into energy and building blocks of which the cell is built up from. The latter is essential for reproduction (growth). Conversion of chemical components (like nutrients) into intermediates or endproducts is a metabolic reaction and is carried out by proteins, called enzymes, which are encoded on the genome. The whole network of all possible metabolic reactions in a cell is called the metabolic network. Since for many organisms, including bacteria and even human, the genome sequence is available, it is possible to reconstruct genome-scale metabolic networks [1,2]. Many enzymes that are encoded on the genome can be deduced if the genome is annotated, i.e. genes and their functions are identified. Subsequently, reactions can be assigned to enzymes (and thus genes) by exploring specific enzyme-reaction databases. The collection of gene-protein-reaction (GPR) associations form a network of interacting reactions through the use of common low weight chemical compounds, also known as metabolites. For example, a certain reaction produces a certain metabolite

and another reaction converts it up to the final production of biomass components (i.e. necessary components of which the cell is built up from). A reconstructed genome-scale metabolic network contains hundreds of GPRs and can be obtained using automated procedures [3–5].

It is of interest in cellular systems biology to model the behavior of metabolism in the context of cell growth at genome-scale [6–9]. This behavior can be expressed in terms of fluxes

* Correspondence to: J. A. Westerhuis, Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands.
  E-mail: j.a.westerhuis@uva.nl

a M. P. H. Verouden, J. A. Westerhuis, A. K. Smilde
   Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands

b R. A. Notebaart
   Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen, P.O.Box 9010, 6500 GL Nijmegen, The Netherlands

c M. J. van der Werf
   TNO Quality of Life, P.O.Box 360, 3700 AJ Zeist, The Netherlands

d B. Teusink
   Systems Bioinformatics, Centre for Integrative Bioinformatics, VU University Amsterdam, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands

through reactions, which is in fact a rate of metabolite conversion, as response to nutrient uptake. Fluxes can be determined by laboratory experiments or by *in silico* simulations. Due to current experimental limitations in determining all fluxes at genome-scale, the flux through each reaction is said to vary within a range. The latter means that many flux distributions of the entire network are possible. Since reactions are connected on the basis of common metabolites it is expected to find correlations between reactions in terms of their flux value over different flux distributions. It is thus possible to infer functionally coherent reactions on the basis of the entire network behavior [10,11].

The question now arises why we are interested in correlated reactions? It leads to the definition of reaction modules which can be useful for in-depth investigation of the reactions without considering the entire network, e.g. by adding a kinetic model to the reactions in such a module to provide more mechanistic detail. Moreover, it could lead to insights into the regulation of enzymes that catalyze reactions [12]. Recently a number of methods have been developed to infer correlations between reactions within specific environments [10,11,13,14]. Usually, the flux of each reaction in the network across many possible flux distributions is examined to calculate the correlation between reactions. However, for the analysis of metabolic systems it may not only be interesting to study reaction correlations in individual environments, but also across different environmental conditions. This could lead to the definition of correlating reactions that appear to be dependent or independent of the environment. In other words, which reaction correlations are robust against changing environment and which are not? Reactions that correlate within and between conditions can then be considered as strong functionally associated. Since for each environmental condition the correlations between the reactions in the metabolic network from many possible flux distributions can be calculated, the question arises how to identify reaction correlations that are invariant, i.e. do not change, and which reaction correlations are variant across environmental conditions.

In this paper, flux distributions of several different *in silico* environmental conditions mimicking laboratory growth conditions are used to calculate correlations between reactions in a genome-scale metabolic network of a bacterium. Subsequently the correlations between the reactions across these *in silico* environmental conditions are analyzed in a multivariate approach with the goal to identify correlations that are invariant and reaction correlations that are variant across conditions. The applied multivariate data analysis methods are Parallel Factor Analysis (PARAFAC) and Principal Component Analysis (PCA).

Section 2 introduces the genome-scale metabolic network and explains how the flux distributions in the different environments have been obtained. Section 3 reveals how the flux distributions are preprocessed, how the correlations are calculated and elaborates on PARAFAC and PCA in relation to the analysis of correlations across multi-environmental conditions. In Section 4, the results of analyzing changes in correlations between reactions across the environmental conditions are given and commented on. Finally in Section 5 the applicability of PARAFAC and PCA for analyzing these changes will be discussed and some important findings are concluded.

2. MATERIALS

2.1. Reconstruction metabolic network

In this study we focus on the calculation of reaction correlations within and between environmental conditions using an *in silico* metabolic network of a lactic acid bacterium called *Lactococcus lactis* MG1363. The metabolic network has been reconstructed on the basis of the sequenced genome [15] using a semi-automatic approach [3] and manual adjustments. Genome-scale metabolic networks contain information about which genes and gene products (i.e. protein) catalyzes which metabolic reactions. Using this semi-automatic approach [3] we determined equivalent genes between *L. lactis* and organisms such as *Lactococcus plantarum* and *Escherichia coli*, for which a manually curated genome-scale metabolic network has been published. Thereby, we inferred metabolic reactions for *L. lactis* which together form a network of interacting chemical compounds leading to the production of biomass components (i.e. components of which the bacterium is built up from). The network consists of 616 metabolites (*M*) involved in 598 metabolic reactions, of which 87% has been associated to proteins/genes. The latter means that 13% of the metabolic reactions has been added to allow the bacterium to grow in the *in silico* simulation. The model includes 513 genes and 445 proteins (including complexes). Moreover, 91 exchange reactions and 1 biomass reaction have been included, leading to a total of 690 reactions (*N*). Exchange reactions define the link between the metabolic network and the environment, thereby allowing for uptake and secretion of nutrients from and to the environment. The biomass reaction serves as a sink for metabolites that are precursors for growth (i.e. metabolites that are used to generate biomass).

2.2. Constraint-based modeling to infer flux states of metabolism

We have applied constraint-based modeling to explore flux distributions of metabolism of *L. lactis* MG1363 by *in silico* simulation in different environments [8]. The different environments are theoretical scenarios that mimic possible real life laboratory environmental conditions. For constraint-based modeling it is essential to structure the *in silico* metabolic network into the so-called the stoichiometric matrix **S** [*M* × *N*]. Each row of the matrix represents a metabolite, each column a reaction and each element the stoichiometry coefficient of the metabolite in that reaction. The stoichiometry coefficient is negative when the metabolite is consumed in a particular reaction and positive when produced. After structuring the metabolic network in the stoichiometric matrix **S**, dynamic mass balances around metabolites are defined in terms of fluxes (metabolite conversion rates) through each reaction and the stoichiometry of those reactions around the metabolites [8,16,17] in the form of

$$\frac{d\mathbf{c}}{dt} = \mathbf{S}\mathbf{v} \qquad (1)$$

where **v** [*N* × 1] denotes a vector of fluxes through all reactions in the network and **c** [*M* × 1] is a vector representing all metabolites in the network. At steady state there is no accumulation or depletion of metabolites in a metabolic network, therefore, the rate of production of each metabolite in the network must equal its rate of consumption, i.e. the change in the amount of any metabolite

Table I. Order, design and number of blocked reactions of the environmental conditions						
Environmental condition	Medium		Aeration			# Blocked
	Rich	*Minimal*	*Anaerob*	*Aerob*	*Aerob resp.*	
1	1	0	1	0	0	10
2	1	0	0	1	0	3
3	1	0	0	0	1	0
4	0	1	1	0	0	36
5	0	1	0	1	0	29
6	0	1	0	0	1	26

within the network over time becomes zero for all metabolites in all reactions ($\frac{dc_m}{dt} = 0$ where c_m is the mth metabolite in the vector **c**). In mathematical terms this is written as,

$$\mathbf{Sv} = \mathbf{0} \qquad (2)$$

Equation (2) limits the solution space of the allowable flux distributions to the nullspace of matrix **S** and eliminates the time derivatives in Equation (1). The steady-state assumption is relevant for intracellular reactions since these reactions are typically much faster than the rate of change in the resultant phenotype such as cell growth (biomass production) [6]. Besides the stoichiometric constraint there is another important constraint, called capacity constraint, to restrict the flux through reactions. This constraint defines the range of flux values that can be taken by each reaction in the *in silico* network and is written as

$$v_{i,\min} \leq v_i \leq v_{i,\max} \qquad (3)$$

Setting capacity constraints is especially important for the exchange reactions, because this is the way to vary the influx of nutrients from the environment in between certain minimal and maximal rates. It allows for *in silico* simulation of metabolism in the context of the environment, i.e. simulating theoretical scenarios reflecting possible real life laboratory environmental conditions.

Imposing these two constraints results in a space of allowable flux distributions of the network [18]. In practice the number of possible flux distributions is too large for direct interpretation and therefore approaches have been developed to select flux distributions. One of these methods is based on random sampling of points (i.e. flux distributions) from the solution space of allowable flux distributions [10]. In order to calculate correlations between reactions in the metabolic network from the space of allowable flux distributions we apply random sampling to sample 2000 flux distributions using the COBRA toolbox [13] with default settings. We have performed the approach for several environments to study not only correlations within single environments, but also across environments. In total six different environmental conditions (theoretical scenarios reflecting real life laboratory conditions) are examined, including:

- Anaerobic growth (without the presence of oxygen), scenario in which the flux capacity constraint for the exchange reaction of oxygen has been set to zero.
- Aerobic growth (in the presence of oxygen), scenario in which the flux capacity constraint for the exchange reaction of oxygen has been set in between a minimum and a maximum value (unequal to zero).

- Aerobic respiratory growth (in the presence of oxygen under addition of haem, which aids the transport of oxygen into *L. lactis*), scenario in which the flux capacity constraint for the exchange reaction of oxygen and the flux capacity constraints for haem-dependent reactions have been set in between a minimum and a maximum value (unequal to zero).

In rich medium, containing glucose and all amino acids and minimal medium, containing glucose and seven amino acids that are minimally required for growth of *L. lactis*. In the rich and minimal medium the flux capacity constraints for all exchange reactions of amino acids present in the medium have been set between boundaries (minimal and maximum flux). For the minimal medium the flux capacity constraints for all exchange reactions of amino acids that are not present in the medium have been set to zero. Table I displays the order and design of the environmental conditions as used throughout this paper.

The sampled flux distributions of fluxes through the reactions in the *L. lactis* metabolic network for each environmental condition have been stored in a matrix $\mathbf{X}_k^*$ [$I \times J^*$], with k representing the environmental condition number as given in column 1 of Table I, I the number of sampled flux distributions (2000) and J^* the number of reactions in the reconstructed metabolic network (690).

3. METHODS

3.1. Blocked reactions and futile cycles

Before correlation matrices can be calculated from the flux distributions it is necessary to remove reactions that have zero flux in all environmental conditions for every flux distribution, and reactions that form futile cycles. Reactions with zero flux for every flux distribution in all conditions are blocked [11], i.e. inactive, for all environmental conditions. Futile cycles [19] are created when two metabolic reactions run simultaneously in opposite directions or a series of metabolic reactions form a circular path. These cycles have no overall effect other than wasting energy, but may have a role in metabolic regulation [20]. Here, however, these cycles are an artifact of the constrained based modeling, because only mass balance and flux capacity constraints are active, and need to be removed. Assignment of additional constraints, e.g. thermodynamic or energetic [8,21–23] can help to prevent futile cycles from appearing in a metabolic network.

The flux distribution matrices for the different environmental conditions from which the overall blocked reactions and reactions involved in futile cycles have been removed, denoted by $\mathbf{X}_k$ [$I \times J$], may still contain reactions with zero fluxes for all flux

distributions. These reactions, however, are blocked only within specific environmental conditions and are not removed from the flux distribution matrices ($\mathbf{X}_k$) in order to keep the reaction dimension (J) of all environmental conditions the same to allow for comparison between and over conditions.

3.2. Correlation matrices

The flux distribution matrices ($\mathbf{X}_k$) are used to calculate a Pearson correlation matrix, between the fluxes carried by pairs of reactions for all sampled flux distributions, for each environmental condition, as denoted by $\mathbf{\Phi}_k$ [$J \times J$]. Pearson correlation is used, because the sampled fluxes of the reactions follow normal distributions. As mentioned in Section 3.1 the flux distribution matrix of a specific condition may still contain reactions that are blocked for that specific condition and contain zero flux for all sampled flux distributions of that reaction. Obviously no correlation can be calculated between a reaction containing zero flux for all flux distributions and any other reaction. In order to create a valid correlation matrix a correlation of zero has been imputed for the correlations between a reaction containing zero flux for all flux distributions and a reaction containing fluxes. A correlation of one has been imputed for correlations between reactions that both contain zero flux for all flux distributions.

The Pearson correlation coefficient for a pair of reactions (s and t) is denoted by φ_{st} and by definition falls within in the range $-1 \leq \varphi_{st} \leq 1$. Two special cases within this range exist [14]:

$\varphi_{st} = \pm 1$: The fluxes carried by the pair of reactions, s and t, are in a fixed ratio for all sampled flux distributions. The reactions, therefore, behave the same and belong to same reaction subset [24].

$\varphi_{st} = 0$: The fluxes carried by the pair of reactions, s and t, are not related to each other for all the sampled flux distributions. The reactions belong to different reaction subsets.

Correlations between 0 and 1 or -1 can be found for fluxes around branching points. A branching point in a metabolic network is a metabolite that has one flux through an incoming reaction and two fluxes through two outgoing reactions. When the inbound flux of a metabolite can have any value within its flux capacity constraint and the outbound fluxes are not in a fixed ratio, the correlation between the incoming reaction and an outgoing reaction will take a value in the range given above for φ_{st}.

The sign of the correlation φ_{st} between a pair of reactions, s and t depends on how the stoichiometry of both reactions is initially specified in the stoichiometry matrix $\mathbf{S}$. If for example a reaction system containing metabolites A, B and C with reactions s and t among those metabolites is specified as

$$s: \quad A \rightarrow B$$
$$t: \quad B \rightarrow C$$
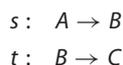
with the arrow indicating which metabolite is initially specified as reactant and product. The flux carried by reaction s will be exactly the same as the flux carried by reaction t. The reactions behave the same, which means they belong to the same subset and the correlation will, therefore, be $\varphi_{st} = 1$. However, if the system was initially specified as

$$s: \quad A \rightarrow B$$
$$t: \quad C \rightarrow B$$
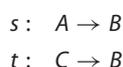
The fluxes carried by reaction s and reaction t would have the same values but always have opposite signs. The reactions would still be in the same subset, but their correlation would be $\varphi_{st} = -1$. Because of this sign indeterminacy of the correlation, absolute correlations will be used throughout this paper. However, the positive semidefinite property of the correlation matrices is removed by using absolute correlations. The resulting absolute correlation matrices are still symmetric and there still exists dependency between the elements in the absolute correlation matrices, meaning that changing an absolute correlation between a pair of reactions also changes the correlations with reactions that highly correlate with this pair.

To visualize the correlation matrices of the environmental conditions one specific order of the reactions will be applied. To determine the order of the reactions hierarchical clustering with average linkage has been used for environmental condition 1 of Table I (anaerobic growth in rich medium) with one minus the absolute correlation serving as distance measure. We will use this reaction order throughout the paper.

3.3. Mean centering

In this paper we are interested in whether correlations between reactions stay the same or change across different environmental conditions. We will, therefore, apply mean centering of the absolute correlations across environmental conditions [25] prior to multivariate data analysis. The correlations between reactions that are invariant across all environmental conditions, i.e. their values remain the same, will have only zeros after mean centering. These invariant correlations contain no variation and are, therefore, not described by the applied multivariate methods. The variant correlations, i.e. correlations that change across environmental conditions, contain variation after mean centering on which the applied multivariate methods will focus.

In Section 4.3 we will show and comment on the invariant correlations. The variant correlations will be analyzed with PARAFAC and PCA.

3.4. PARAFAC

By stacking absolute correlation matrices of multi-environmental conditions, as denoted by $|\mathbf{\Phi}_k|$ [$J \times J$], on top of each other a datacube, represented by $\underline{\mathbf{\Phi}}$ [$K \times J \times J$], is obtained that has a three-way structure. Multi-way analysis after mean centering, therefore, is the obvious approach for identifying correlations between reactions that change across environmental conditions.

The most appropriate multi-way method for analyzing correlation matrices of multiple environmental conditions is an INDSCAL (individual difference scaling) model [26,27]. Although, our datacube contains absolute correlation matrices that are no longer positive semidefinite but still symmetric with remaining dependency between the correlations within each environmental condition, we assume that the INDSCAL model still applies. The INDSCAL model, consisting of loading matrices $\mathbf{A}$ [$K \times R$] and $\mathbf{B}$ [$J \times R$] both containing the same number of components R, can be represented by

$$|\mathbf{\Phi}_k| - \overline{\mathbf{\Phi}} = \mathbf{B}\mathbf{D}_k\mathbf{B}^T + \mathbf{E}_k \tag{4}$$

where $|\mathbf{\Phi}_k|$ is the kth horizontal slice of $\underline{\mathbf{\Phi}}$, $\overline{\mathbf{\Phi}}$ [$J \times J$] contains the mean of $\underline{\mathbf{\Phi}}$ over all environmental conditions ($k = 1, \ldots, K$), $\mathbf{D}_k$ is a diagonal matrix with the kth row of loading matrix

A on its diagonal (elements $a_{k1}, \ldots, a_{kR}$) and $\mathbf{E}_k$ [$J \times J$] denotes the residual term of the kth horizontal slice of $\underline{\Phi}$. It has been shown that a symmetric case of a canonical decomposition (CANDECOMP) model [27] can be used to estimate the parameters of the INDSCAL model [28, p. 388–389]. In practice a PARAFAC model [29,30], which is mathematically equivalent to a CANDECOMP model, can also be used to estimate the parameters of the INDSCAL model. The resulting PARAFAC model, as shown in Equation (5) [31, p. 59–64], consists of loading matrices **A**, **B** and **C** [$J \times R$] with $\mathbf{B} \approx \mathbf{C}$ [32].

$$|\mathbf{\Phi}_k| - \overline{\mathbf{\Phi}} = \mathbf{B}\mathbf{D}_k\mathbf{C}^T + \mathbf{E}_k \qquad (5)$$
$$\text{with } \mathbf{B} \approx \mathbf{C}$$

3.5. PCA

Datamatrix **V** [$K \times (J \times J)$] is created by putting the vectorized absolute correlation matrices of each environmental condition, $|\mathbf{\Phi}_k|$, into its rows. The rows of **V** represent the environmental conditions as given in Table I.

After mean centering the PCA model [33,34] summarizes the mean centered datamatrix **V** in a bilinear model containing a set of scores **T** [$K \times R$] and loadings **P** [$R \times (J \times J)$], with the number of components $R \ll \min(K, J \times J)$, while the residuals in **E** contain the non-systematic variation that cannot be modeled.

$$\mathbf{V} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T\mathbf{V} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \qquad (6)$$

The PCA model does not consider the dependency between the absolute correlations when building the model, but this dependency is, however, still present in datamatrix **V**.

4. RESULTS

4.1. Blocked reactions and futile cycles

Within the six specific environmental conditions given in Table I in total 210 reactions are blocked for all conditions and 42 reactions involved in futile cycles have been identified and removed from the data. The remaining number of reactions, therefore, equals $J = 438$.

As stated in Section 3.1 each individual environmental condition can still contain blocked reactions, but these are specific for that environmental condition. The last column of Table I shows the number of blocked reactions per specific condition. The number of blocked reactions for environmental condition 4–6 is higher than for environmental condition 1–3. The explanation for this observation is that in environmental condition 4–6 a minimal growth medium has been used that contains only seven amino acids, whereas in environmental condition 1–3 a rich growth medium has been used containing all amino acids. Transport and exchange reactions for amino acids not present in the minimal medium, as given in Table II, become blocked. Both within environmental conditions 1–3 and 4–6 there is a decrease in the number of blocked reactions that can be explained by the change in aeration condition.

Table II. Reactions that get blocked going from rich medium to minimal medium environmental conditions under the same aeration condition

Abbreviation	Description
EX_ala-L(e)	L-Alanine exchange
EX_arg-L(e)	L-Arginine exchange
EX_asn-L(e)	L-Asparagine exchange
EX_asp-L(e)	L-Aspartate exchange
EX_gln-L(e)	L-Glutamine exchange
EX_gly(e)	Glycine exchange
EX_lys-L(e)	L-Lysine exchange
EX_orn-L(e)	Ornithine exchange
EX_phe-L(e)	L-Phenylalanine exchange
EX_pro-L(e)	L-Proline exchange
EX_ser-L(e)	L-Serine exchange
EX_thr-L(e)	L-Threonine exchange
EX_trp-L(e)	L-Tryptophan exchange
EX_tyr-L(e)	L-Tyrosine exchange
ARGORNt3	Arginine/ornithine antiporter
ARGabc	L-arginine transport via ABC system
ARGt2	L-arginine transport in via proton symport
ASNt2	L-asparagine transport in via proton symport
ASPt2	L-aspartate transport in via proton symport
GLNabc	L-glutamine transport via ABC system
LYSt6	L-lysine transport in/out via proton symport
PHEt6	L-phenylalanine transport in/out via proton symport
PROabc	L-proline transport via ABC system
SERt6	L-serine transport in/out via proton symport
TRPt6	L-tryptophan transport in/out via proton symport
TYRt6	L-tyrosine transport in/out via proton symport

Table III. Reactions that are blocked under anaerobic growth condition		
No.	Abbreviation	Description
1	EX_o2(e)	O_2 exchange
2	ALOX	oxidative decarboxylation of acetolacate (chemical)
3	GSHPO	glutathione peroxidase
4	GTHRD	glutathione-disulfide reductase
5	NOX2	NADH oxidase (H_2O forming)
6	NPR	NADH peroxidase
7	O2t	O_2 transport in via diffusion
8	CYTB_B2	menaquinol oxidase (7:1 protons)
9	G3PD4	glycerol-3-phosphate dehydrogenase (menaquinone 7)
10	NADH4	NADH dehydrogenase (menaquinone 7 and no proton)

Table III shows which reactions are blocked under anaerobic growth conditions. The first seven reactions in this table get unblocked when changing from anaerobic to aerobic growth conditions (change from environmental condition 1 to 2 and 4 to 5), this is explained by the fact that these reactions all involve oxygen usage. The last three reactions get unblocked when changing from aerobic to aerobic respiratory conditions (change from environmental condition 2 to 3 and 5 to 6) and can be explained, because they are involved in the respiratory cycle.

4.2. Correlation matrices

Pearson correlation matrices have been calculated for all environmental conditions (Table I). Because of sign indeterminacy of the correlations between reaction pairs in the metabolic network (described in Section 3.2) absolute correlations are used. Taking the absolute value of a correlation matrix; however, removes the positive semidefinite property, but the symmetry and dependency between the correlations within the matrix remains unchanged.

Figure 1 shows the absolute correlation matrix for, respectively, environmental condition 2 (aerobic growth in rich medium) and 5 (aerobic growth in minimal medium) of Table I. The reactions in both correlation matrices have been ordered as discussed in Section 3.2. Reactions behaving the same can be seen in Figure 1 as black blocks on the diagonal with correlation one. There is a very large group of reactions behaving the same and there are several small sets of reactions that perfectly correlate. When comparing Figure 1(a) and (b) changes in correlations can be observed. Some clusters change in size, meaning that reactions become uncorrelated to the other reactions, and sometimes the correlation between clusters is different.

To enhance the visualization of the blocks of reactions on the diagonal and the correlation between blocks, logical correlation matrices have been created by setting all correlations within each environmental condition smaller than 1 to a value of 0 and the correlations of the blocked reactions within each condition to a value of 3. Setting the blocked reactions to a value of 3 enables the visualization of reactions that become blocked and unblocked when comparing two environmental conditions. Figure 2(a) shows the difference between the logical correlation matrices of Figure 1(a) and (b), and clearly visualizes the changes in correlations between aerobic growth in a rich medium and a minimal medium. The black dots (value of −3) in Figure 2(a) are

correlations between reactions that get blocked when *L. lactis* grows in a minimal medium (Table II) and involve transport and exchange reactions of amino acids that are not present in the minimal growth medium. The gray bars (value of −1) in Figure 2(a) contain reaction clusters (Table IV) whose correlation changes with an existing cluster of reactions (they add to the existing cluster, as visible in Figure 1(b)). The reactions in these clusters are related to the amino acid biosynthesis of amino acids that cannot be taken up from the medium, because they are simply not present. Figure 2(b) visualizes the changes in correlations between anaerobic and aerobic respiratory growth conditions in rich medium (only the part containing changes is shown, the rest of the figure contained only zeros). The reactions in the lower right corner of this figure (values of +3) are the ones that become unblocked going from anaerobic to aerobic respiratory growth condition (Table III). In this cluster of reactions there are some reactions (values of +2) that become unblocked under growth in aerobic respiratory condition and cluster together (numbers 1/7 and 3/4 from Table III form two separate clusters). In the upper left part of Figure 2(b) there are also three reactions visible in light gray, i.e. EX_diact(e): diacetyl exchange; DIACTt: diacetyl diffusion and ACTD: acetoin dehydrogenase, that form a cluster in rich medium under anaerobic growth. However, in rich medium under aerobic respiratory growth condition one of reactions (ACTD: acetoin dehydrogenase) drops out of this cluster.

Comparisons between two environmental conditions reveal the differences between those conditions, but do not reveal the underlying concepts of which correlations change (i.e. are variant) across multi-environmental conditions due to a change in a specific environmental factor (e.g. aeration condition). When the effects of environmental factors can be identified in a model of the performed experiments, this can possibly help the choice of the environmental factor settings when planning new experiments.

4.3. Invariant correlations

The invariant correlations can be identified by examining the standard deviation of the correlations across all environmental conditions, where a standard deviation close to 0 signifies correlations that are invariant. Figure 3(a) displays the standard deviations of the correlations across all conditions and clearly shows which clusters of reactions (seen as blocks with standard deviation close to 0) are invariant across conditions. For example the large block almost in the middle of Figure 3(a)
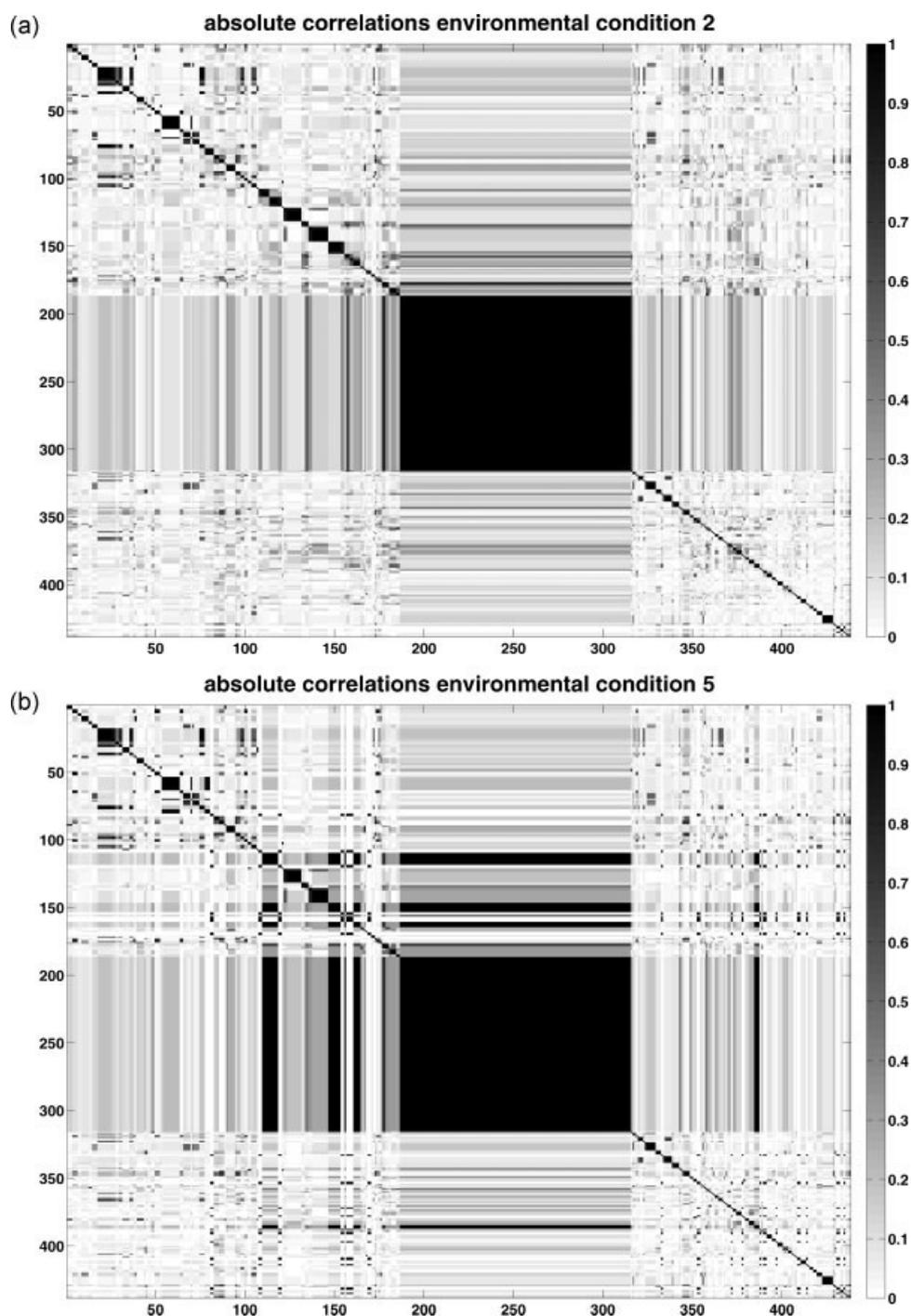
411

Figure 1. Sorted absolute correlations between reactions for: (a) condition 2 of Table I (rich medium aerob); (b) condition 5 of Table I (minimal medium aerob).

contains invariant correlations between reactions (129 out of 438 reactions), that are directly linked to the production of biomass (lipids, proteins, vitamins, polysaccharides and cell wall components). Table V shows some reactions from this biomass production related cluster with indication to which class they belong. The variant correlations between reactions that are linked to amino acid biosynthesis, as described in the previous section (Figure 2(a) and Table IV) add to the large block of invariant correlations and are also directly linked to biomass production

in the minimal medium. Figure 3(a) also shows which correlations between clusters or parts of clusters on the diagonal are invariant across conditions, e.g. correlations between and within reactions 55–63 and 124–131 are invariant (Table VI).

Invariance of correlations, merely, states that the correlations do not change across environmental conditions and does not imply that these correlations are high or low. By studying the mean of the correlations over environmental conditions an impression of the actual value of correlations can be obtained.
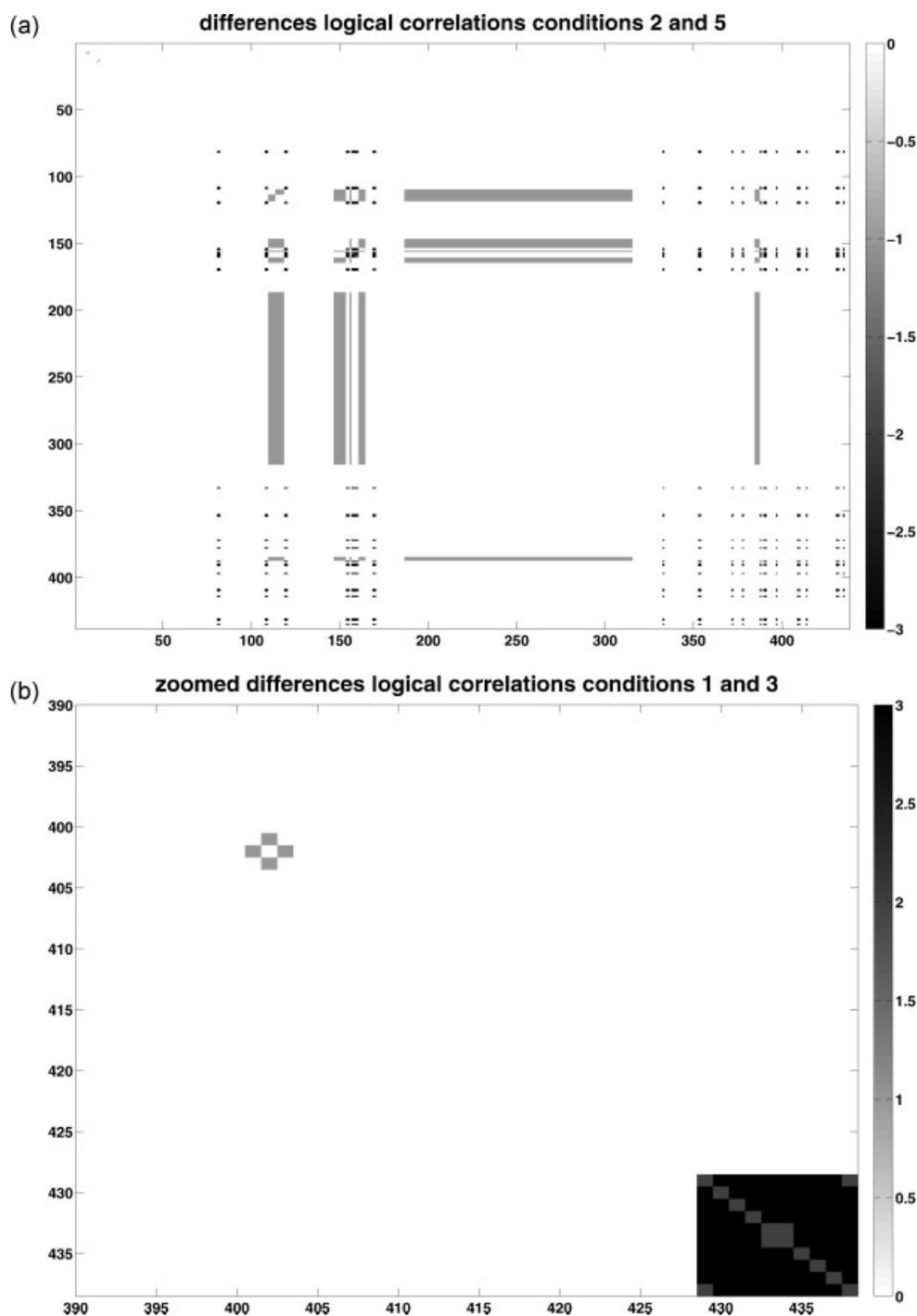
Figure 2. Difference between logical correlations correlations matrices for growth: (a) in rich medium and minimal medium under aerobic conditions; (b) under anaerobic and aerobic respiratory conditions in rich medium.

Figure 3(b) shows the mean correlation over all conditions and average correlations close to one show reaction pairs that behave very similar and belong to the same subset. Although the correlations between and within the reactions in Table VI are invariant and the reactions within each group belong to the same subset, there is no relation between the two subsets because the mean correlation of the reactions between the subsets is close to 0. The most interesting reactions are the ones that have invariant correlations across conditions and high mean correlation over conditions, as for example the reactions related to biomass production. Correlations with high mean and low standard deviation across conditions between subsets of reactions, of which the individual subsets also have high mean correlations with low standard deviation across conditions, are also very interesting because these subsets display the same behavior and might be closely related in the sense of functionality.

Table IV. Reaction cluster that add to an existing cluster of reactions when changing from aerobic growth in rich medium to minimal medium

Bar no.	Abbreviation	Description
1	G5SADs	L-glutamate 5-semialdehyde dehydratase (spontaneous)
	G5SD	glutamate-5-semialdehyde dehydrogenase
	GLU5K	glutamate 5-kinase
	P5CR	pyrroline-5-carboxylate reductase
	ANPRT	anthranilate phosphoribosyltransferase
	ANS1	anthranilate synthase
	IGPS	indole-3-glycerol-phosphate synthase
	PRAI	phosphoribosylanthranilate isomerase
	TRPS1	tryptophan synthase (indoleglycerol phosphate)
2	ADPDS	acetyl-diaminopimelate deacetylase
	ADPTA	acetyl-diaminopimelate transaminase
	APAT	apolipoprotein *N*-acyl transferase
	DAPDC	diaminopimelate decarboxylase
	DAPE	diaminopimelate epimerase
	DHDPRy	dihydrodipicolinate reductase (NADPH)
	DHDPS	dihydrodipicolinate synthase
3	PPND	prephenate dehydrogenase
4	ACGK	acetylglutamate kinase
	ACOTA	acetylornithine transaminase
	AGPR	*N*-acetyl-g-glutamyl-phosphate reductase
	ORNTAC	ornithine transacetylase
5	ARGSL	argininosuccinate lyase
	ARGSS	argininosuccinate synthase
	OCBT	ornithine carbamoyltransferase

4.4. PARAFAC

In order to find the underlying concepts of which correlations between reactions change due to changes in environmental conditions multivariate data analysis on the multi-environmental absolute correlation matrices has been applied. We start with a PARAFAC model, because the data (stacked absolute correlation matrices) in itself has a multi-way structure and the PARAFAC model keeps the dependency between the absolute correlations intact. In the PARAFAC model we have chosen for two components. Addition of a third component did not result in a better interpretable model. Mean centering across conditions is applied before data analysis to show differences between the correlations over all environmental conditions.

A two-component unconstrained PARAFAC model explains 61% of the variation in datacube $\underline{\Phi}$. Figure 4(a) shows that the scores for the environmental mode (A) are strongly related to each other and one component would be enough to describe the separation between the environmental conditions, whereas the expected number of components is at least two because of the two design factors (medium and aeration) in the environmental conditions. The strong relation between the scores for the environmental mode is also confirmed by the Tucker congruence value between component 1 and 2 of the environmental mode A ($T_{12}^A = -1.0000$), which indicates a degeneracy of that mode. The Tucker congruence values between component 1 and 2 for mode B and C are, $T_{12}^B = T_{12}^C = 0.9929$. Multiplication of the

Tucker congruence values of all three modes results in Tucker's congruence coefficient for component 1 and 2 ($T_{12} = -0.9858$). For a degenerated model, i.e. PARAFAC is unable to correctly fit the trilinear model, Tucker's congruence coefficient will be close to -1 [31, p. 107–108].

To resolve the relationship between the two components in the PARAFAC model we have built a two component PARAFAC model with an orthogonality constraint on environmental mode A. This model explains 45% of the variation in datacube $\underline{\Phi}$. The scores on the environmental mode shown in Figure 4(b) now display a direction clearly visualizing the separation between rich and minimal growth medium and a direction that describes some of the variation caused by changes in aeration. However, both effects are mixed over the two components. The Tucker congruence values of the second and third mode for component 1 and 2 ($T_{12}^B = T_{12}^C = 0.9196$) show that, although the degeneracy of the environmental mode ($T_{12}^A = 0$) has been resolved, a degeneracy problem within the components of mode B and C still exists. The outer product of the second (B) and third (C) mode for component one is shown in Figure 5(a) and component two is displayed in Figure 5(b). Although the values in both figures are not exactly the same, the pattern displayed is the very similar for both components, as could be expected from the degeneracy of the second and third mode. Comparison of the patterns in Figure 5 with the pattern in Figure 2(a) shows that both components mainly describe the variation in the correlations caused by a change in growth medium. The variation in correlations caused by a change in aeration condition, which results in changes of
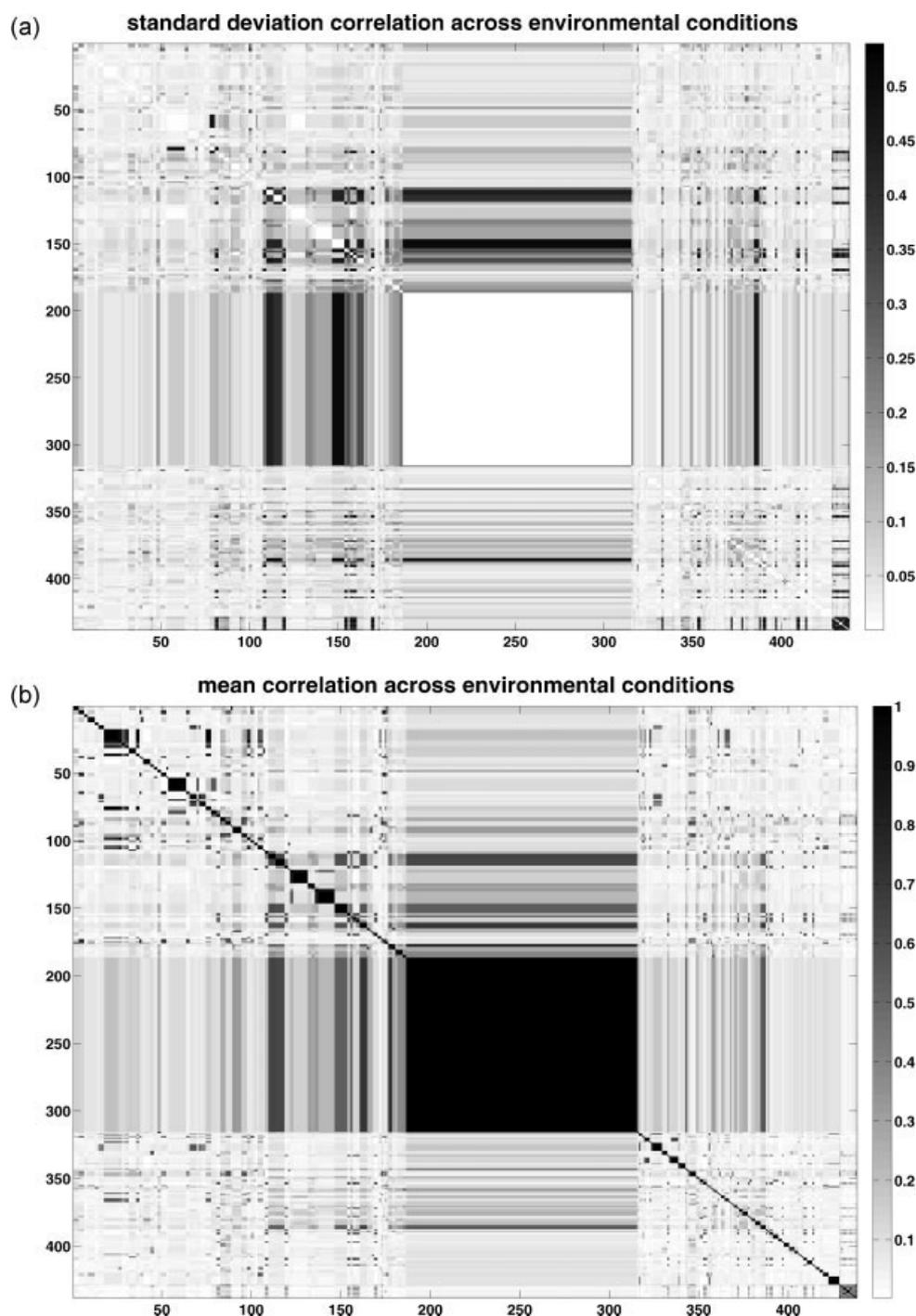
Figure 3. Standard deviation and mean of the correlations across all environmental conditions: (a) standard deviation; (b) mean.

blocked reactions (visible in the lower right part of Figures 2(b) and given in Table III), is not at all explained by the PARAFAC model. The outer product between the loadings of mode *B* and *C* for both components (Figure 5) also show that the large block of reactions related to biomass product get a value, whereas those correlations are invariant across conditions (Section 4.3). The reason is that it is impossible for multi-way component models to simultaneously explain zeros and non-zero values when there are

blocks with zero value present in the data. This raises the question whether PARAFAC is a good choice to analyze this data.

4.5. PCA

The second applied multivariate data analysis method is PCA after mean centering of the vectorized absolute correlations matrices of all environmental conditions is used. Mean centering

Table V. Examples of reactions in the block related to biomass production

Class	Abbreviation	Description
Coenzyme A synthesis	DPCOAK	dephospho-CoA kinase
	PNTK	pantothenate kinase
	PPCDC	phosphopantothenoylcysteine decarboxylase
DNA synthesis	DNAS_LLA	DNA synthesis, LLA specific
Fatty acid biosynthesis	ACACT1r	acetyl-CoA C-acetyltransferase
	FABM	Fatty acid enoyl isomerase (FabM reaction)
	MACPMT	Malonyl-CoA:[acyl-carrier-protein] S-malonyltransferase
	kaasIII	beta-ketoacyl-ACP synthase III
Lipid biosynthesis	DMATT	dimethylallyltranstransferase
	DPMVD	diphosphomevalonate decarboxylase
	HMGCOAS	Hydroxymethylglutaryl CoA synthase
	MEVK	mevalonate kinase
	UDCPDPS	Undecaprenyl diphosphate synthase
Peptidoglycan/teichoic acid biosynthesis	GALTAL	Galactose lipoteichoic acid ligase
	GLUR	glutamate racemase
	PGAMT	phosphoglucosamine mutase
	UDCPDP	undecaprenyl-diphosphatase
Phospholipid biosynthesis	GLYK	glycerol kinase
	CLPNS_LLA	Cardiolipin Synthase (lactis specific)
	LPGS_LLA	lysylphosphatidyl-glycerol synthetase
Polysaccharide metabolism	G1PTMT	glucose-1-phosphate thymidylyltransferase
	PGMT	phosphoglucomutase
	UDPG4E	UDPglucose 4-epimerase
Protein synthesis	ALATRS	Alanyl-tRNA synthetase
	LEUTRS	Leucyl-tRNA synthetase
	TRPTRS	Tryptophanyl-tRNA synthetase
Pyrimidine biosynthesis	DTMPK	dTMP kinase
RNA synthesis	RNAS_LLA	RNA synthesis, lactis specific
Vitamins and cofactor metabolism	DHFR	dihydrofolate reductase
	TMDS	thymidylate synthase

Table VI. Clusters of reactions that have invariant correlations between each other

| Cluster	Reaction number	Abbreviation	Description
1	55	CHORS	chorismate synthase
	56	DAHPS	3-deoxy-D-arabino-heptulosonate 7-phosphate synthetase
	57	DHQD	3-dehydroquinate dehydratase
	58	DHQS	3-dehydroquinate synthase
	59	PSCVT	3-phosphoshikimate 1-carboxyvinyltransferase
	60	RPE	ribulose 5-phosphate 3-epimerase
	61	SHK3D	shikimate dehdrogenase
	62	SHKK	shikimate kinase
	63	TKT2	transketolase
2	124	ADSL2	adenylosuccinate lyase
	125	AIRC	phosphoribosylaminoimidazole carboxylase
	126	GARFT	phosphoribosylglycinamide formyltransferase
	127	GLUPRT	glutamine phosphoribosyldiphosphate amidotransferase
	128	PRAGS	phosphoribosylglycinamide synthetase
	129	PRAIS	phosphoribosylaminoimidazole synthetase
	130	PRASCS	phosphoribosylaminoimidazolesuccinocarboxamide synthase
	131	PRFGS	phosphoribosylformylglycinamidine synthase
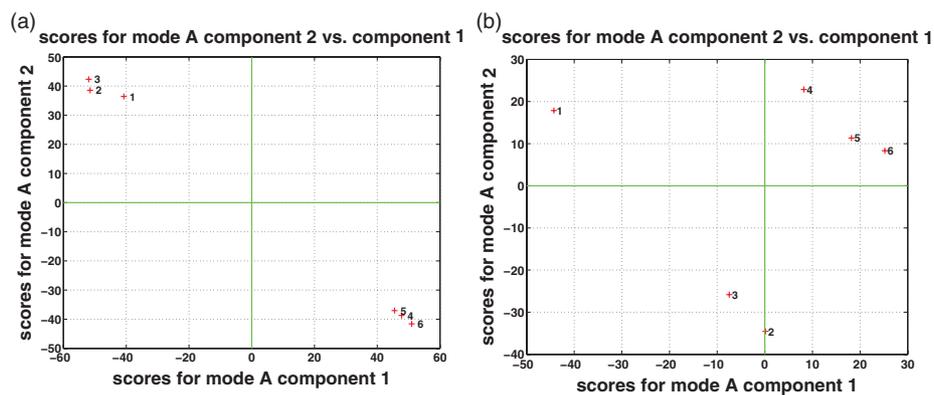
Figure 4. Scores on component 2 versus component 1 for the environmental mode (*A*) of, (a) an unconstrained PARAFAC model; (b) a PARAFAC model with orthogonality constraints on the environmental mode.
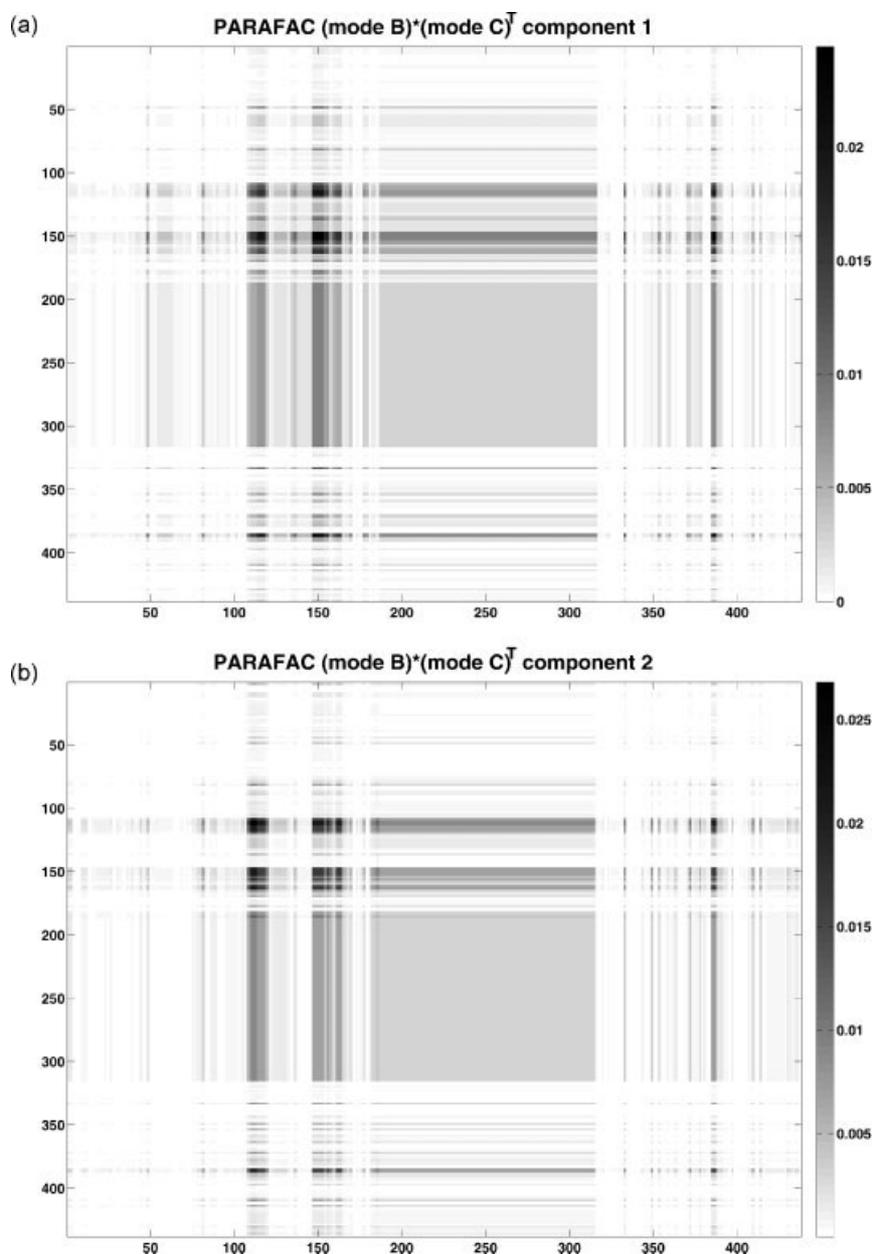

Figure 5. Outer product between mode *B* and mode *C* for PARAFAC mode with orthogonality constraints on the environmental mode *A*: (a) component 1; (b) component 2.

is again applied to show differences between correlations across environmental conditions.

A two-component PCA-model of the mean centered datamatrix **V** captures 82% of the variation. Two components have been chosen here and a scree graph of the eigenvalues versus the number of components [34] clearly showed an elbow at two components. This number of components matches the underlying design of the environmental conditions, namely growth medium and aeration condition. One component for the aeration condition might not be sufficient because it has three levels, but addition of an extra component did not reveal extra information with respect to the aeration condition.

A plot of the scores on principal component 2 (PC2) versus principal component 1 (PC1), as shown in Figure 6, reveals that PC1 (explaining 71% of the variation in the data) nicely separates the rich medium environmental conditions from the minimal medium conditions. PC2 does not show a clear separation
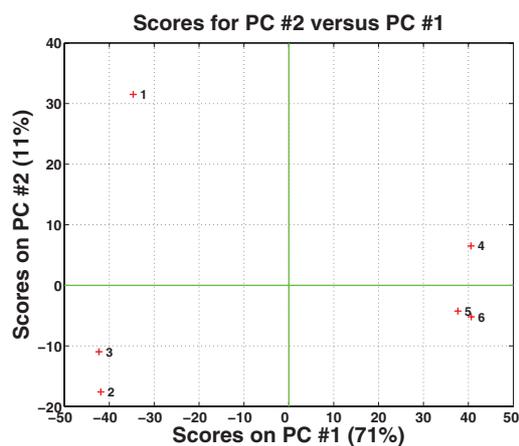

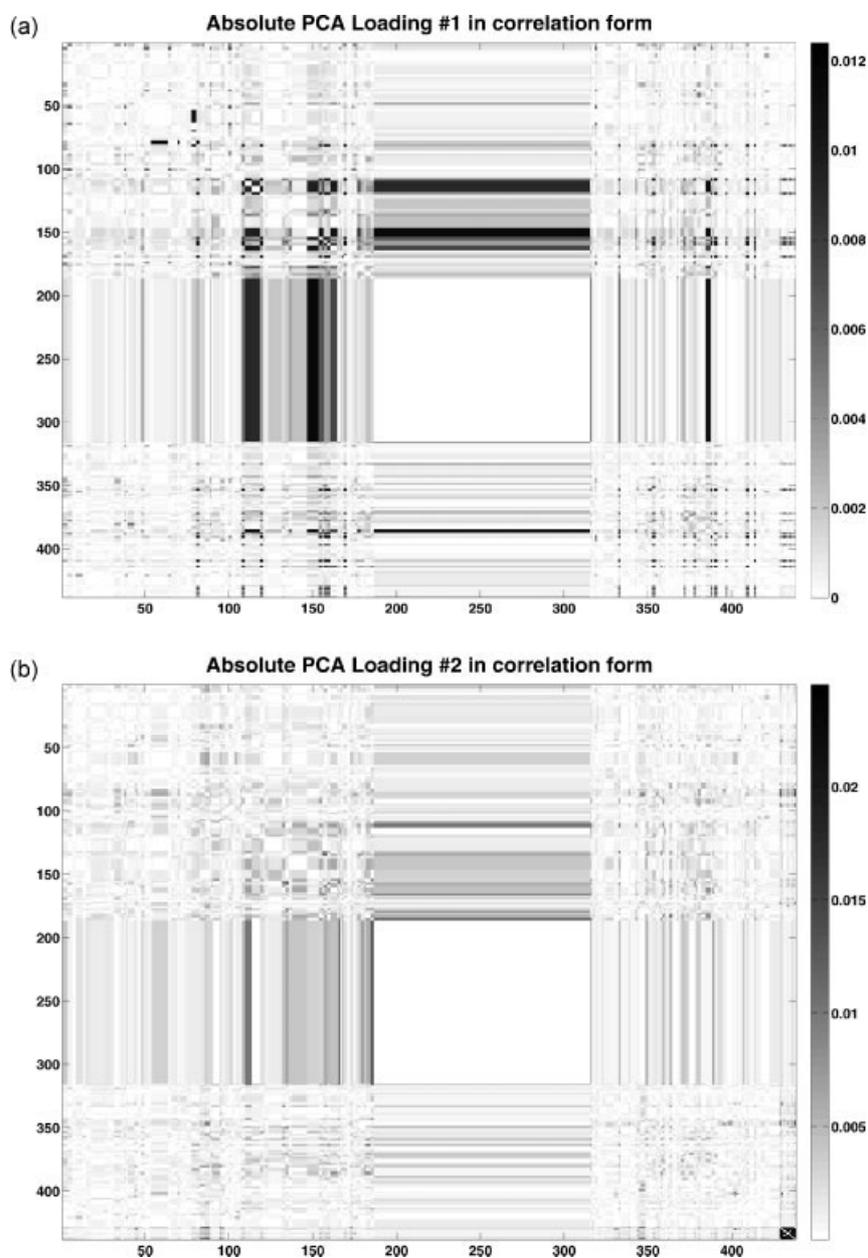**Figure 6.** Scores on PC2 versus PC1 for a two-component PCA model.

Figure 7. PCA loadings after folding back into correlation matrix structure: (a) loading for PC1; (b) loading for PC2.

between environmental conditions with respect to aeration conditions, but does show a part of the variation in the data linked to change in aeration. The fact that a change in growth medium accounts for the most variation in the correlations between reactions over all environmental conditions is not very surprising, as we have already seen in the comparison of two conditions (Figures 1 and 2), that a change in growth medium displays many changes in the correlations between only two conditions. Figure 7 shows the loadings of PC1 and PC2 of the PCA model after folding them back into a correlation matrix structure for ease of interpretability. Because the first PCA loading (Figure 7(a)) is linked to the change in medium, the correlations between reactions that vary the most by the change in growth medium can be seen as correlations with a high absolute loading weight. Figure 7(a) displays the same pattern of dots and bars (Figure 2(a)) as described in Section 4.1 and the reactions with highly variable correlations are indeed the same reactions as given in Tables II and IV. The second PCA loading (Figure 7(b)) display highly variable correlations in the low right corner. The reactions linked to these variable correlations are those that change by a change in aeration condition as described in Section 4.1 (Figure 2(b) and Table III).

5. DISCUSSION AND CONCLUSION

Correlation matrices have been calculated for each condition using flux distributions through reactions in a metabolic network of *L. lactis* for several environmental conditions. Because of sign indeterminacy of the correlations between reactions pairs in a metabolic network absolute correlations are used, but this removes the positive semidefinite property of a correlation matrix. The resulting absolute correlation matrices are still symmetric and the dependency between correlations remains.

When stacking the absolute correlation matrices of all conditions after mean centering correlations between reactions can be identified that are invariant across environmental conditions. For identifying correlations that change (i.e. are variant) across environmental conditions multivariate approaches are applied (PARAFAC and PCA).

The PARAFAC model has been used because the data of stacked absolute correlation matrices in itself has a three-way structure and PARAFAC keeps the dependency among the correlations intact. The two component PARAFAC model with orthogonality constraints on the environmental mode does not explain the variation in the data very well (only 45% of the variation is explained). The model suffers from degeneracy in the second (*B*) and third (*C*) mode and the outer product of these modes for both components only display the variant correlations of reactions effected by a change in growth medium. The variant correlations between reactions due to a change in aeration condition can not be identified at all with the PARAFAC model. Furthermore does PARAFAC have difficulty in modeling the specific structure of the data, which contains many zeros after mean centering due to correlations that are invariant across environmental conditions. Techniques like weighted PARAFAC [35], maximum likelihood PARAFAC [36] or MILES [37] might help in dealing with this specific structure in the data.

Instead of dealing with the specific structure in the data with techniques as suggested above, we have chosen to vectorize the absolute correlation matrices of the environmental conditions and applying PCA. Although PCA does not consider the dependency between the correlations in each environmental condition when building the model, this dependency is of course still present in the data after vectorizing it. By vectorizing the data PCA gets much more parameters for modeling the data compared to a PARAFAC model and combined with the fact that PCA ignores the dependency between the correlations, it is not surprising that in the PCA model much more of the variation present in the data is explained (82%) than in the PARAFAC model. However, by allowing so much more freedom in modeling the data by removing the dependency between the correlations a real risk of overfitting lurks. The loadings of the PCA model do; however, nicely show the variable correlations across conditions by the change in both growth medium and aeration condition.

The crucial question that remains, is whether latent variable models are suitable for analyzing this type of data. Perhaps clustering type models are better equipped for describing this data type in which blocks exist of invariant and variant correlations. In our further research we intend to focus on simplivariate models [38] and three-mode partitioning [39] for identifying invariant and variant blocks of correlations.

REFERENCES

1. Francke C, Siezen RJ, Teusink B. Reconstructing the metabolic network of a bacterium from its genome. *Trends Microbiol.* 2005; **13**(11): 550–558.
2. Reed JL, Famili I, Thiele I, Palsson BØ. Towards multidimensional genome annotation. *Nat. Rev. Genet.* 2006; **7**(2): 130–141.
3. Notebaart RA, van Enckevort FHJ, Francke C, Siezen RJ, Teusink B. Accelerating the re30 construction of genome-scale metabolic networks. *BMC Bioinf.* 2006; **7**: 296.
4. Satish Kumar V, Dasika MS, Maranas CD. Optimization based automated curation of metabolic reconstructions. *BMC Bioinf.* 2007; **8**: 212.
5. DeJongh M, Formsma K, Boillot P, Gould J, Rycenga M, Best A. Toward the automated generation of genome-scale metabolic networks in the seed. *BMC Bioinf.* 2007; **8**: 139.
6. Varma A, Palsson BØ. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coliW3110. *Appl. Environ. Microbiol.* 1994; **60**(10): 3724–3731.
7. Ibarra RU, Edwards JS, Palsson BØ. Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* 2002; **420**(6912): 186–189.
8. Price ND, Reed JL, Palsson BØ. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* 2004; **2**(11): 886–897.
9. Teusink B, Wiersma A, Molenaar D, Francke C, de Vos WM, Siezen RJ, Smid EJ. Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model. *J. Biol. Chem.* 2006; **281**(52): 40041–40048.
10. Price ND, Schellenberger J, Palsson BØ. Uniform Sampling of Steady-State Flux Spaces: means to design experiments and to interpret enzymopathies. *Biophys. J.* 2004; **87**(4): 2172–2186.
11. Burgard AP, Nikolaev EV, Schilling CH, Maranas CD. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res.* 2004; **14**(2): 301–312.
12. Notebaart RA, Teusink B, Siezen RJ, Papp B. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLoS Comput. Biol.* 2008; **4**(1): e26.

13. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, HerrgardMJ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. *Nat. Protoc.* 2007; **2**(3): 727–738.

14. Poolman MG, Sebu C, Pidcock MK, Fell DA. Modular decomposition of metabolic systems via null-space analysis. *J. Theor. Biol.* 2007; **249**(4): 691–705.

15. Wegmann U, O'Connell-Motherway M, Zomer A, Buist G, Shearman C, Canchaya C, Ventura M, Goesmann A, Gasson MJ, Kuipers OP, van Sinderen D, Kok J. Complete genome sequence of the prototype lactic acid bacterium Lactococcus lactis subsp. cremoris MG1363. *J. Bacteriol.* 2007; **189**(8): 3256–3270.

16. Lee JM, Gianchandani EP, Papin JA. Flux balance analysis in the era of metabolomics. *Brief. Bioinform.* 2006; **7**(2): 140–150.

17. Edwards JS, Covert M, Palsson BØ. Metabolic modelling of microbes: the flux-balance approach. *Environ. Microbiol.* 2002; **4**(3): 133–140.

18. Wagner C, Urbanczik R. The geometry of the flux cone of a metabolic network. *Biophys J.* 2005; **89**(6): 3837–3845.

19. Schwender J, Ohlrogge J, Shachar-Hill Y. Understanding flux in plant metabolic networks. *Curr. Opin. Plant Biol.* 2004; **7**(3): 309–317.

20. Samoilov M, Plyasunov S, Arkin AP. Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. *Proc. Natl. Acad. Sci. U.S.A.* 2005; **102**(7): 2310–2315.

21. Beard DA, Liang Sd, Qian H. Energy balance for analysis of complex metabolic networks. *Biophys. J.* 2002; **83**(1): 79–86.

22. Kümmel A, Panke S, Heinemann M. Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinf.* 2006; **7**: 512.

23. Henry CS, Broadbelt LJ, Hatzimanikatis V. Thermodynamics-based metabolic flux analysis. *Biophys. J.* 2007; **92**(5): 1792–1805.

24. Pfeiffer T, Sánchez-Valdenebro I, Nuño J, Montero F, Schuster S. METATOOL: for studying metabolic networks. *Bioinformatics* 1999; **15**(3): 251–257.

25. Bro R, Smilde AK. Centering and scaling in component analysis. *J. Chemom.* 2003; **17**(1): 16–33.

26. Horan C. Multidimensional scaling: combining observations when individuals have different perceptual structures. *Psychometrika* 1969; **34**(2): 139–165.

27. Carroll JD, Chang JJ. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika* 1970; **35**(3): 283–319.

28. Carroll JD, Pruzansky S. *The CANDECOMP-CANDELINC Family of Models and Methods for Multidimensional Data Analysis*, chap. 10. In Law *et al*. [40], 372–402.

29. Harshman R. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics* 1970; **16**: 1–84.

30. Harshman RA, Lundy ME. *The PARAFAC Model for Three-Way Factor Analysis and Multidimensional Scaling*, chap. 5. In Law *et al*. [40], 122–215.

31. Smilde A, Bro R, Geladi P. *Multi-way Analysis with Applications in the Chemical Sciences*. John Wiley & Sons, Ltd: Chichester, 2004.

32. ten Berge JM, Kiers HA. Some clarifications of the CANDECOMP algorithm applied to INDSCAL. *Psychometrika* 1991; **56**(2): 317–326.

33. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemom. Intell. Lab. Syst.* 1987; **2**(1–3): 37–52.

34. Jolliffe IT. *Principal Component Analysis, Springer Series in Statistics*, 2nd edn, Vol. XXIX. Springer-Verlag: New York, 2002.

35. Andersson GG, Dable BK, Booksh KS. Weighted parallel factor analysis for calibration of HPLC-UV/Vis spectrometers in the presence of Beer's law deviations. *Chemom. Intell. Lab. Syst.* 1999; **49**(2): 195–213.

36. Vega-Montoto L, Wentzell PD. Maximum likelihood parallel factor analysis (MLPARAFAC). *J. Chemom.* 2003; **17**(4): 237–253.

37. Bro R, Sidiropoulos ND, Smilde AK. Maximum likelihood fitting using ordinary least squares algorithms. *J. Chemom.* 2002; **16**(8–10): 387–400.

38. Hageman JA, Hendriks MMWB, Westerhuis JA, van der Werf MJ, Berger R, Smilde AK. Simplivariate models: ideas and first examples. *PLoS One* 2008; **3**(9): e3259.

39. Schepers J, van Mechelen I, Ceulemans E. Three-mode partitioning. *Comput. Stat. Data Anal.* 2006; **51**(3): 1623–1642.

40. Law HG, Snyder CW, Jr., Hattie JA, McDonald RP (eds.) *Research Methods for Multimode Data Analysis*. Praeger Publishers: New York, 1984.

420