Contents lists available at SciVerse ScienceDirect

Journal of Chromatography B

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

Independent component analysis in non-hypothesis driven metabolomics: Improvement of pattern discovery and simplification of biological data interpretation demonstrated with plasma samples of exercising humans[☆]

Xiang Li[a,e], Jakob Hansen[b], Xinjie Zhao[a], Xin Lu[a], Cora Weigert[c,d], Hans-Ulrich Häring[c,d], Bente K. Pedersen[b], Peter Plomgaard[b], Rainer Lehmann[c,d,*], Guowang Xu[a,**]

[a] CAS Key Laboratory of Separation Science for Analytical Chemistry, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, 16023 Dalian, China
[b] The Centre of Inflammation and Metabolism, Department of Infectious Diseases and Copenhagen Muscle Research Center, Rigshospitalet, Faculty of Health Sciences, University of Copenhagen, 2100 Copenhagen, Denmark
[c] Division of Clinical Chemistry and Pathobiochemistry (Central Laboratory), University Hospital Tuebingen, D-72076 Tuebingen, Germany
[d] Paul-Langerhans-Institute Tübingen (Inst. for Diabetes Research and Metabolic Diseases of the Helmholtz Centre Munich at the University of Tübingen), Eberhard Karls University Tübingen, D-72076 Tübingen, Germany
[e] Qinhuangdao Entry-Exit Inspection and Quarantine Bureau of P.R.C., 066004 Qinhuangdao, China

ARTICLE INFO

Article history:
Received 24 November 2011
Accepted 25 June 2012
Available online 1 July 2012

Keywords:
Independent component analysis
Metabolomics
Exercise
Metabolic profiling
GC–MS

ABSTRACT

In a non-hypothesis driven metabolomics approach plasma samples collected at six different time points (before, during and after an exercise bout) were analyzed by gas chromatography–time of flight mass spectrometry (GC–TOF MS). Since independent component analysis (ICA) does not need a priori information on the investigated process and moreover can separate statistically independent source signals with non-Gaussian distribution, we aimed to elucidate the analytical power of ICA for the metabolic pattern analysis and the identification of key metabolites in this exercise study. A novel approach based on descriptive statistics was established to optimize ICA model. In the GC–TOF MS data set the number of principal components after whitening and the number of independent components of ICA were optimized and systematically selected by descriptive statistics. The elucidated dominating independent components were involved in fuel metabolism, representing one of the most affected metabolic changes occurring in exercising humans. Conclusive time dependent physiological changes of the metabolic pattern under exercise conditions were detected. We conclude that after optimization ICA can successfully elucidate key metabolite pattern as well as characteristic metabolites in metabolic processes thereby simplifying the explanation of complex biological processes. Moreover, ICA is capable to study time series in complex experiments with multi-levels and multi-factors.

1. Introduction

A huge amount of complex data are generated especially by non-hypothesis driven metabolomics approaches, including information based on analytical characteristics like ion masses including metabolites, fragments but also noise, as well as biological effects (e.g. metabolic processes, environmental influences, etc.). Therefore, mining useful information in the collected data is a key step in metabolomics, and chemometrics plays an important role in this context. Currently, principal component analysis (PCA) and partial least-squares discriminant analysis (PLS-DA) are the commonly applied methods. These approaches extract principal components or latent variables from the data after dimension reduction. Unrelated factors or noisy components are excluded to focus solely on useful information [1]. PLS-DA is an extension of PCA which includes known class information aiming to maximize the separation between groups, and it is a typical supervised method. If the knowledge about the biological processes or the analytical information (e.g. noise) is ambiguous or lacking, supervised methods may fail to separate biological information from miscellaneous data [2], or they bear a high risk of over-fitting the data and misinterpretation of the observations may happen [2,3].

An alternative strategy is independent component analysis (ICA) [4,5], which is usually applied in the area of blind source separation (BSS). ICA has been widely and successfully utilized in signal processing [6,7], image feature extraction [8,9], medical imaging

[☆] This paper belongs to the Special Issue Chemometrics in Chromatography, Edited by Pedro Araujo and Bjørn Grung.

[*] Corresponding author. Tel.: +49 07071 29 83193; fax: +49 07071 29 5348.

[**] Corresponding author. Tel.: +86 411 84379559; fax: +86 411 84379559.
E-mail addresses: Rainer.Lehmann@med.uni-tuebingen.de (R. Lehmann), xugw@dicp.ac.cn (G. Xu).

1570-0232/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jchromb.2012.06.030

[10,11], genomics and protein profiling [12,13], process monitoring [14,15]. ICA can separate the source signals with non-Gaussianity and statistically independent data, thus it has also successfully been applied in the areas of proteomics [13,16,17], transcriptome [18] and metabolomics [19–24]. Although the numbers of principal components after whitening and the independent components are important for the explanations of the data and the biological process [18], few papers were focused on the optimization of ICA [18]. Furthermore, to the best of our knowledge the performance and potential benefits of optimized ICA in particular in non-hypothesis metabolomics approaches have not been investigated and described in detail.

In this study we optimized, evaluated and applied ICA in a non-hypothesis-driven metabolomics approach based on descriptive statistics. By combining characteristic metabolites discovered by ICA with network analysis the biological interrelationship of independent component in metabolomics networks is demonstrated. Optimized ICA was applied in a non-targeted metabolomics investigation of human plasma samples collected at six different time points before, during and after a single bout of exercise. Investigating this complex metabolic process with multi-factorial influences, the analytical performance of optimized ICA had been confirmed by the detection of important metabolic pattern and identification of key metabolites.

2. Theory

The typical problem solved by ICA is BSS. Subsequent to the receipt by the receiver (e.g. by mass spectrometer in metabolomics), the source signals are mixed into mixed signals. BSS can separate these source signals based on the mixed signals without the information of source signals and the mixing approaches. If the source signals are statistically independent from each other, then they can be separated by ICA.

The mathematical expression of ICA is:

$$x = As \tag{1}$$

Here $x=(x_1,x_2,...,x_k)^{\mathrm{T}}$ represents mixed signals, i.e. the detected data (signals, metabolites concentrations, etc.); $s=(s_1,s_2,...,s_l)^{\mathrm{T}}$ represents the source signals, i.e. independent components; A is the mixing matrix and represents the mixing approaches in signal mixing.

The FastICA [25] algorithm for ICA is based on maximization of non-Gaussianity and has the advantages of reliable, robust and fast convergence. In the following section we investigate the applicability of FastICA as the ICA method in our metabolomics study.

The non-Gaussianity can be estimated by different methods like kurtosis and negentropy. The kurtosis is defined as:

$$\mathrm{kurtosis}\,(z) = \frac{\sum_{i=1}^{n}(z_i-\mu)^4}{(N-1)\sigma^4} - 3 \tag{2}$$

Here z represents the variable with the mean value μ and the variance of σ. A positive kurtosis means the variable is super-Gaussian, and a negative kurtosis means the variable is sub-Gaussian. The kurtosis of Gaussian distribution is 0.

FastICA algorithm is based on negentropy, and the detailed description and mathematical proof of FastICA can be found in the literature [5,25]. As the information of independent components and the mixing matrix are unknown, the variances (i.e. amplitudes of the signals, including the signs) and orders of the independent components cannot be determined by ICA [26].

Introducing BSS in metabolomics, x in Eq. (1) can be considered as the metabolomics data recorded by the analytical instrument, and s (or A, depending on the input format of x, for details see Section 4) can be considered as the impact factors with some non-Gaussianity and statistical independency in the metabolic process. Then A (or s) can be considered as the weights of the metabolites contributing to the independent components, consequently the metabolites with large weights are the most important metabolites in the investigated metabolic context.

Of note, the number of detected metabolites in metabolomics data sets is generally much higher than the number of investigated samples. Furthermore the data often contain variables not related to the metabolic process (e.g. analytical noise). A useful preprocessing strategy in ICA is to whiten the observed variables, and the most commonly used whitening method is PCA. PCA whitening decreases the data into a few principal components, and these principal components are used as the variables in ICA. Thus the variables of ICA are largely decreased and the variables with smaller variance can be excluded (e.g. noise). The number of principal components after whitening determined the remaining variances (i.e. information) of the raw data and the exclusion of irrelevant components. Thus this strategy may support the elucidation of relevant data thereby simplify the interpretation of the remaining data set.

Scholz et al. [19] suggested the number of independent components based on kurtosis measure of the independent components, i.e. only the independent components with negative kurtosis were selected. But this method is based on the pre-requisite that the information of the data is solely included in the components with sub-Gaussian distribution. Furthermore, the number of principal components after whitening is not optimized by this approach. Therefore, it may be worthwhile to determine the number of principal components and independent components simultaneously, with no limitation to the non-Gaussian distribution of data, i.e. applicable to all metabolomics data.

We applied in our approach the kurtosis as measure of non-Gaussianity and studied the descriptive statistics for ICA with different numbers of principal components and independent components. Selecting these parameters and investigating the network of characteristic exercise metabolites in human plasma discovered by independent components, the relevance of independent components and the performance of ICA in metabolomics were investigated and discussed in our study.

3. Experimental

3.1. Samples

Eight volunteers were enrolled in the exercise study. The protocol of the study was approved by the local ethical committee (H-D-2007-0127) conformed to the Declaration of Helsinki before commencement, and all subjects gave the written informed consent. The investigation was conducted in accordance with the ethical principles of good clinical practice. The volunteers performed one-leg knee-extensor exercise for 120 min as described elsewhere [27]. 48 blood samples were collected at the following time points: before the exercise bout (=0 min), during the exercise bout at 60 min and 120 min, as well as in the recovery phase (the subjects had to recover by lying in a bed) at 150 min, 180 min and 300 min after the start of the experiment. The plasma was immediately stored at −80 °C until sample preparation.

3.2. Sample preparation

The plasma was thawed on ice. 100 μl of plasma was added to 400 μl of methanol, then 20 μl internal standards (16.6 μg/mL sorbic acid and 16.6 μg/mL D3-methyl lauric acid) in methanol solution were added into the mixture. The mixture was vortexed

Table 1
Statistic description of the randomly generated ICA models.

Dimension*	Number of independent components	Case**	500 models	1000 models	2000 models	5000 models
1	1	1	IC(24)	IC(24)	IC(24)	IC(24)
2	1	1	IC(41)	IC(41)	IC(41)	IC(41)
		2	IC(50)	IC(50)	IC(50)	IC(50)
2	2	1	IC(41,47)	IC(41,47)	IC(41,47)	IC(41,47)
		2	IC(38,50)	IC(38,50)	IC(38,50)	IC(38,50)
3	1	1	IC(72)	IC(72)	IC(72)	IC(72)
		2	IC(77)	IC(77)	IC(77)	IC(77)
3	2	1	IC(72,77)	IC(72,77)	IC(72,77)	IC(72,77)
3	3	1	IC(17,72,77)	IC(17,72,77)	IC(17,72,77)	IC(17,72,77)

* Dimension was the number of principal components after whitening by PCA.
** Cases 1 and 2 were named randomly.

and proteins were removed by centrifugating at 13,000 rpm for 20 min at 4 °C. Following that the supernatant was collected and lyophilized. The lyophilized samples were derivatized with 50 μl pyridine and 100 μl BSTFA (bis(trimethylsilyl)trifluoroacetamide) at 80 °C for 30 min.

3.3. Instrumental conditions

A LECO Pegasus 4D GC × GC–TOF MS instrument (LECO Corporation, St. Joseph, MI, USA, run in the GC–TOF MS mode) with an Agilent 6890 N GC was used. The carrier gas was Helium (99.9995%) with a constant flow of 1.2 mL/min. The GC column was a 30 m × 250 μm × 0.25 μm DB-5 column (J&W Scientific, Folsom, CA). 2 μl of sample was injected into the GC column by Agilent 7683B autosampler (Agilent, Palo Alto, CA, USA) at a split ratio of 1:5. The temperature of the GC oven was set as follows: start at 70 °C, held for 3 min, then ramped at 10 °C/min up to 320 °C and held for 5 min. The ion source was set at 230 °C, the temperature of GC inlet was 300 °C, and the temperature of the transfer line was 280 °C. Solvent delay time was 430 s to avoid solvent peaks. The detector voltage was 1600 V and the electron energy was −70 V. Mass spectra of m/z 33–600 were collected at 5 Hz (5 spectra/s).

3.4. Data processing

The raw data collected by Leco ChromaTOF V3.25 software (Leco Corporation, St. Joseph, MI, USA) were exported to CSV (comma separated value) files and peak alignment was performed by COW (correlation optimized warping) [28,29]. The peak tables of each sample were merged into the data matrix (samples × variables) by a home-made program written by Matlab (Mathworks, Natick). The peak areas of the metabolites were normalized to the internal standards with the closest retention time. The peak areas of internal standards were calculated by Leco ChromaTOF, based on the characteristic ions of m/z 169 for sorbic acid and m/z 105 for D3-methyl lauric acid. The final data matrix contained 173 metabolites per sample and 48 samples in total (8 subjects × 6 time points per individual).

FastICA was performed with the algorithm given by Hyvärinen [5] in the Matlab. A home-made Matlab program was designed to perform descriptive statistics of ICA. The characteristic metabolites discovered by ICA were identified by mass spectrum similarity searches of NIST MS search 2.0 (NIST/EPA/NIH Mass Spectral Library, NIST 05). The Kendall rank correlation coefficients were calculated by the SPSS software, and the metabolic network of the characteristic metabolites were constructed by the Cytocape software.

4. Results and discussion

4.1. The descriptive statistics for ICA

The results of ICA were simply described as follows:

(1) IC_n^m: the metabolomics data were reduced to n dimensions, i.e. PCA was performed to compress the data to n principal components, and then ICA was run to get m independent components (ICs).
(2) $\mathrm{IC}(k1, k2, \ldots)$: ICA model with the independent components where the kurtosis values are $k1$, $k2$, $k3$,... As the order of independent components could not be determined by the ICA algorithm, we ordered them by their kurtosis values. To define a specific ICA model, it was written as $\mathrm{IC}_n^m(k1, k2, \ldots)$.
(3) $S(k)$: the independent component with kurtosis of k.
(4) $A(k)$: the mixing matrix of the independent component with kurtosis of k.

ICA maximizes non-Gaussianity of the data, and the FastICA algorithm measures non-Gaussianity based on the negentropy. We used kurtosis here, a classical measure of non-Gaussianity, to denominate and describe the results of ICA, i.e. each IC and the mixing matrix of the IC.

A novel approach based on descriptive statistics was established to optimize ICA model. We had observed that IC_n^m $(n > 1)$ might have different convergence results when calculating ICA model at a time. These convergence results had different kurtosis values, mixing matrices and ICs (details see below). This situation might occur when the maximization process of the ICA has different extremes. To search all of the possibilities of the maximization results, we established different number of randomly generated ICA models, having the same number of dimensions and independent components. By descriptive statistics of these models with their ICs, mixing matrices and kurtosis values, we can deduce how to determine the number of principal components after whitening and the number of independent components by optimization.

In previous investigations [19,21,30] the data 'x' were modelled to '$A \times s$' by two different approaches: if x was variables × observations, then the independent component 'x' represents the weight matrix of the metabolites, and the mixing matrix 'A' represents the impact factors in the metabolic process [21]; if x was observations × variables, then the mixing matrix 'A' represents the weight matrix of the metabolites, and the independent component 'x' represents the impact factors in the metabolic process [19,30]. These two different input approaches may both be applied in metabolomics in theory, however we applied an alternative strategy.

It is noteworthy that in our research when the input was variables × observations, the algorithm will not be convergent in some models, therefore we did not apply this input approach for our further descriptive statistics. Applying the input observations × variables we achieved descriptive statistics results of the models given in Table 1. If the kurtosis value was rounded to integer, the changing of kurtosis value was less than 2% (even less than 1% in most of cases), and $S(k)$ and $A(k)$ were very similar for the cases where kurtosis values were the same after rounding. Therefore all of the kurtosis values were rounded to integers, and the independent components with the same kurtosis value were considered as identical independent component. This process was denominated as degeneracy in the presented study.

To denote the metabolic changes during the exercise and the recovery phase, the mixing matrix values at each time point of every individual were averaged. Thus the time trend plot of the mixing matrix represented the changes of different impact factors at each time point. The averaging process can also decrease the influences given by individual differences of the volunteers and the system noises.

The data were reduced to n dimensions for IC_n^m after whitening, however it was not possible to estimate the number of different ICA modelling results (different cases) when the number of independent components was fixed as m. The descriptive statistics of randomly generated ICA models can give every possible results and corresponding probabilities, see Table 1. It shows that the numbers of models had no influence on the cases. 500, 1000, 2000 and 5000 models will give the same numbers of cases and the same independent components, regardless the number of models, as shown in Table 1. For example, after degeneracy, 500, 1000, 2000 and 5000 models will all generate two different case of IC_2^1, i.e. IC(41) and IC(50). The only minor differences were the frequencies. The frequency of IC(41) in 500 models was 0.432, the frequency of IC(50) in 500 models was 0.568, while the frequency of IC(41) in 5000 models was 0.462, the frequency of IC(41) in 5000 models was 0.538. No case in Table 1 had low frequency, showing that they were not generated by chance. The changes in the number of the generated models will not lead to the changes in the cases of IC_n^m. Thus we chose the results of 1000 models for further discussion.

If m equals 1, all of different convergences results should be achieved, i.e. all of the extremes during maximization may appear. If the number of different ICA modelling results (cases) for IC_n^1 was w, the case number for $\mathrm{IC}_n^m (m = 2, \ldots n)$ should be IC_w^m (if $w < m$, it should be 1), i.e. the independent components in IC_n^m should be the permutation and combination of the independent components of IC_n^1. The results of Table 1 were well in line with the deduction described above, e.g. the case number of IC_3^1 was 2, and the cases number of IC_3^2 and IC_3^3 was 1. Furthermore, both $S(k)$ and $A(k)$ of IC_3^2 and IC_3^3 were very similar to those of IC_3^1. But the case number of IC_2^2 was 2, the same as IC_1^1. When investigating two cases of IC_2^2, they were substantially the same, as shown in Fig. 1. Fig. 1A shows that $\mathrm{IC}_2^2(38, 50)$ and $\mathrm{IC}_2^2(41, 47)$ had very similar mixing matrices, i.e. $A(47)$ and $A(50)$ were similar, and $A(38)$ and $A(41)$ were similar. Head-to-tail depictions given in Fig. 1B and C, clearly demonstrate that the ICs of $\mathrm{IC}_2^2(38, 50)$ and $\mathrm{IC}_2^2(41, 47)$ were also comparable. The kurtosis values of the similar ICs in $\mathrm{IC}_2^2(38, 50)$ and $\mathrm{IC}_2^2(41, 47)$ differed by only 8%, i.e. their non-Gaussianities were also similar. Consequently $\mathrm{IC}_2^2(38, 50)$ and $\mathrm{IC}_2^2(41, 47)$ were substantially the same case. Thus degeneracy of them was reasonable.

The number of independent components in IC_3^3 was higher than for IC_3^1, i.e. the case number of IC_3^1 was only 2. Fig. 2 shows the mixing matrices and ICs in IC_3^3. $A(77)$ and $A(72)$ were similar (Fig. 2A), and the linearity between them at each time point was also very good ($R^2 = 0.9676$). But as shown in the head-to-tail description in Fig. 2B, $S(77)$ and $S(72)$ were different. Fig. 2C shows $S(72)$ was similar to $S(17)$, although there are obvious differences in the absolute values ($S(72)$ was larger than $S(17)$). Moreover, the kurtosis values of the ICs of IC_3^3 showed a pronounced difference, e.g. the kurtosis values of $S(17)$ and $S(77)$ were 120% different. In summary, the cases of IC_3^3 were obviously different, and they could not be taken into degeneracy. $A(77)$ was similar to $A(72)$, but $S(77)$ was not similar to $S(72)$. The situation above suggested the independent components in IC_3^3 were all confused. A possible explanation of this situation could be that the maximization process of the ICA could not find three extremes simultaneously, resulting in a mixing between two extremes (i.e. two cases in IC_3^1). Therefore this situation could be a clew for selecting optimal number of principal components and independent components.

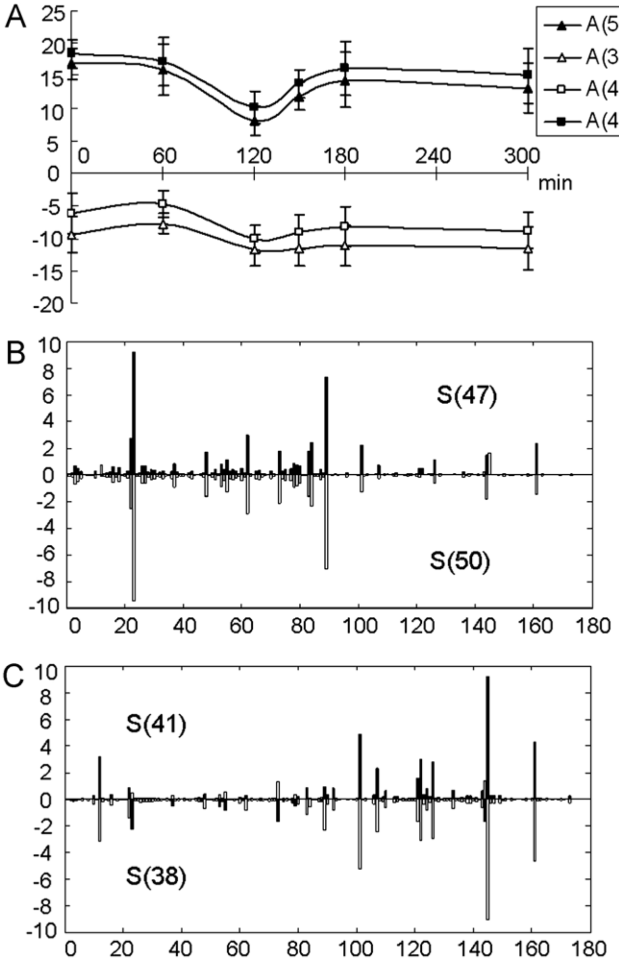

Fig. 1. (A) Mixing matrices ($A(38)$ and $A(50)$) of $\mathrm{IC}_2^2(38, 50)$, and mixing matrices ($A(41)$ and $A(47)$) of $\mathrm{IC}_2^2(41, 47)$, (B) and (C): head-to-tail depictions of independent components of $\mathrm{IC}_2^2(38, 50)$ and $\mathrm{IC}_2^2(41, 47)$, (B) $S(47)$ and $S(50)$; (C) $S(41)$ and $S(38)$. The x-axis is the serial numbers of the detected metabolites.

It can be concluded from the results above that the data can reach the maximized non-Gaussianity in IC_2^2. If one more independent component is calculated the algorithm will lead to a mixing result. Thus optimal number of independent components should be 2. Furthermore, the case number of IC_3^1 and IC_2^1 were the same indicating that one more principal component would contain no more information. Thus the number of principal components after whitening should be 2.

Based on descriptive statistics, the number of principal components after whitening and the number of independent components can be determined as follows:

(1) When the number of independent components is within an appropriate range: if the case number of independent

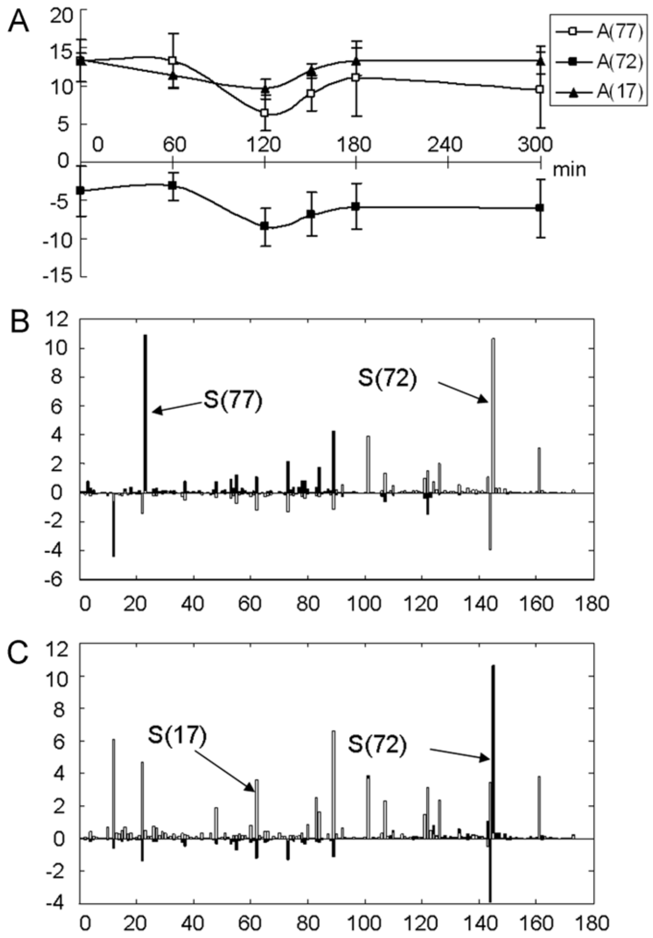

Fig. 2. (A) Mixing matrices (*A*(17), *A*(72) and *A*(77)) of $\mathrm{IC}_3^3(17, 72, 77)$, (B) and (C): head-to-tail depictions of independent components of $\mathrm{IC}_3^3(17, 72, 77)$, (B) *S*(72) and *S*(77), (C) *S*(17) and *S*(72). The *x*-axis is the serial numbers of the detected metabolites.

components in IC_n^1 is *w*, then the case number of independent components in $\mathrm{IC}_n^m(m = 2, \ldots n)$ will be C_w^m (when $w < m$, it should be 1).

(2) When the *A* matrices and *S* matrices are all similar for some independent components, and their kurtosis values are also similar, they can be taken into degeneracy.

(3) When the *A* matrices (or *S* matrices) are similar for some independent components, but the *S* matrices (or *A* matrices) are obviously different and their kurtosis values are not similar, they cannot be taken into degeneracy.

(4) Subsequent to degeneracy: if the number of different independent components in all $\mathrm{IC}_n^m(2, \ldots n)$ cases are higher than in IC_n^1, it suggests that the convergence of ICA reach a mixing result, thus the number of independent components should be less than *m*.

(5) If the case number of IC_{n+1}^1 is the same as IC_n^1, it suggests that the number of principal components after whitening should be *n*.

4.2. The metabolic locus analysis based on ICA

The impact factors at each time point were analyzed based on the optimal number of principal components and independent components for the data. IC_2^2 led to two results which can be taken into degeneracy, i.e. IC(38,50) and IC(41,47). Further investigations were exemplarily performed with IC(38,50).

Two different time loci in the trend plot of the mixing matrix in Fig. 1A became obvious. *A*(50) decreased at 120 min (of note, since in ICA the signs of the ICs are not defined, this change may also be an increase). This was followed by a switch back to a similar level comparable to the situation before the exercise bout. *A*(38) was slightly increasing from 0 to 60 min. Then the levels were decreasing until the end of the exercise bout at 120 min (Fig. 1A) and no obvious change occurred thereafter. The metabolites, which had the highest values in *S*(38) and *S*(50) matrices (Fig. 1B and C), can be considered as the characteristic metabolites for corresponding independent components.

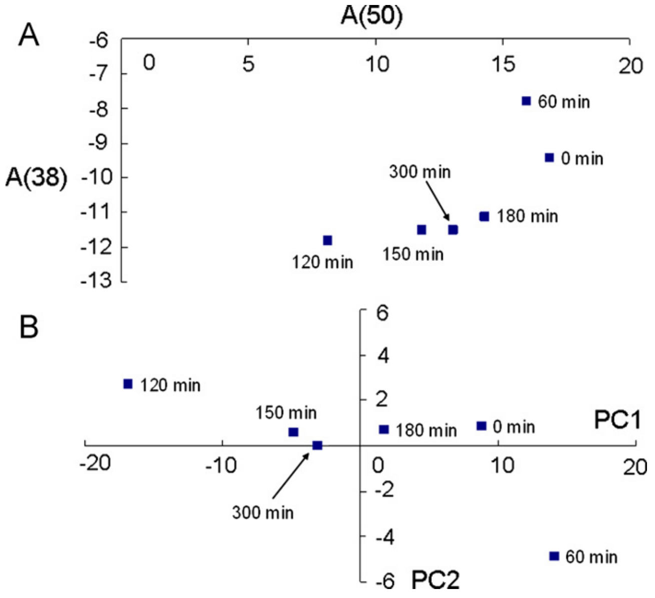

Fig. 3. Metabolic locus plots of time dependent metabolic alterations by combining (A) the mixing matrices of both independent components in ICA and (B) the scores of both principal components in PCA.

By combining the mixing matrices of both independent components, a comprehensive time locus plot was elucidated (Fig. 3A) representing a holistic view on the time dependent metabolic alterations. From pre-exercise (0 min) to 60 min, the metabolic locus was along *A*(38) (*y*-axis), after 60 min, the metabolic locus was along both *A*(38) (*y*-axis) and *A*(50) (*x*-axis). The final metabolic locus within the recovery phase (300 min) was towards the pre-exercising metabolic state (0 min), which fits well to the expected metabolic changes during recovery following an exercise bout, and these changes were primarily along *A*(50). Furthermore, the plot shown in Fig. 3A revealed that the alterations at 60–120 min and 120–180 min were most pronounced. This is well in line with the changes in the physiological processes in a 120 min exercise experiment performed near to exhaustion followed by the recovery phase where many metabolites drops from the end of the exercise bout at 120 min to the 180 min time point in the recovery phase. It confirms the power of the applied ICA analysis. These changes of the metabolic pattern were all associated with both *A*(38) and *A*(50), which shows that these metabolic changes are interrelated.

As an alternative approach PCA was applied to model the data and two principal components were selected (Fig. 3B). Subsequent to the averaging of the scores at the observations of the investigated time points, a metabolic locus can also be generated based on PCA. However, the extents of alterations in metabolism from 0 min to 60 min and 60 min to 120 min were similar to those seen in Fig. 3A from the ICA. PCA can only model the data to elucidate the directions of the largest variances (principal components), but these principal components have no evident physiological relevance. In contrast, independent components from ICA are detected based on their biological importance in the investigated context as demonstrated in the following section.

Table 2
Characteristic metabolites of independent components of IC(38,50) detected by GC–TOF MS. Only metabolites having absolute values >1 in the independent component matrices were selected.

Characteristic metabolites of *S*(50)		Characteristic metabolites of *S*(38)	
No.	Identification*	No.	Identification*
22	Unknown	12**	3-Hydroxybutyric acid
23**	Urea	22	Unknown
48**	Unknown	73	Unknown
55**	Threonic acid	83	Gluconic acid
62**	Unknown	89	Glucose***
73	Unknown	101	Glucose***
83	Gluconic acid	107**	Palmitic acid
84**	Deoxy myo-inositol	121**	Linoleic acid
89	Glucose***	122**	Oleic acid
101	Glucose***	126**	Stearic acid
144	Unknown	144	Unknown
145	Unknown	145	Unknown
161	Cholesterol	161	Cholesterol

* The identified compounds resulted from mass library searches. Most of them were confirmed by the analysis of standard compounds.
** Metabolites belong solely to one independent component.
*** Showed different retention times (possibly α- and β-D-glucose)

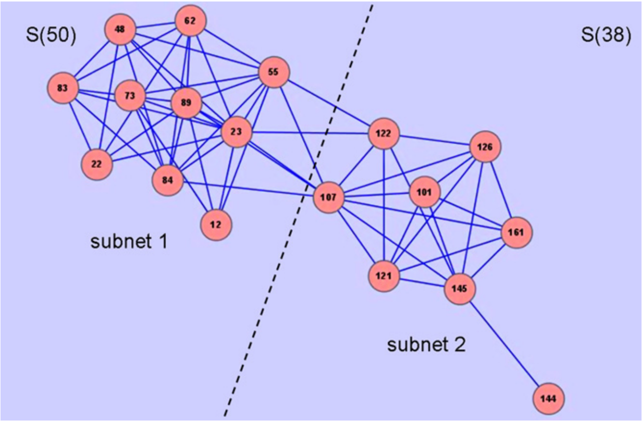

Fig. 4. Metabolic network based on characteristic metabolites of two independent components. All of the significantly correlated metabolites (Kendall rank correlation analysis, $p < 0.01$) were linked by beelines in the network.

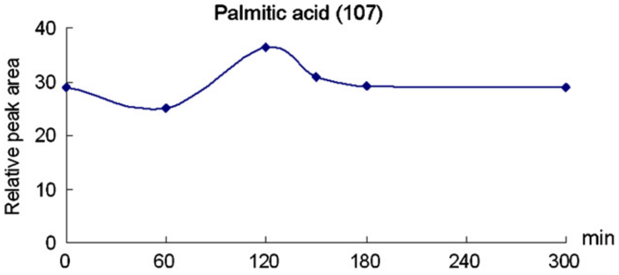

Fig. 5. Time trend plot of a characteristic metabolite of *S*(38), exemplarily shown for palmitic acid.

4.3. The metabolic network and the independent components

The relevance of independent components in ICA for non-hypothesis driven metabolomics was evaluated by elucidation of characteristic metabolites and metabolic network. In this study the selection criteria for these characteristic metabolites was an absolute value >1.0 in the independent component matrices. Table 2 shows exemplarily the identified metabolites and several unknowns of *S*(38) and *S*(50). The metabolites had been identified by mass library search of the GC–TOF MS data and the majority was confirmed further by the analysis of standard compounds. Both ICs contain 13 characteristic metabolites with an absolute value of >1. Of note, most of these metabolites are directly related to metabolic pathways of fuel supply (glucose and lipid metabolism), which is well in line with the enhanced energy demand of the body under exercise conditions and subsequent physiological reactions of the body. Switches between glucose and lipid oxidation are well known metabolic changes during exercise and in the recovery phase [31,32]. Of note, a very interesting finding is that several free fatty acids were detected in *S*(38) but not in *S*(50) (Table 2).

Since eight of these compounds were common characteristic metabolites of both two independent components we established a metabolic network by calculating Kendall rank correlation coefficients between these characteristic metabolites. The significantly correlated metabolites ($p < 0.01$) were linked by beelines, shown in Fig. 4. Interestingly two subnets became obvious (separated by a dash-dotted line; Fig. 4). The interrelationships within the subnets are high and the connections between the subnets are low. A very interesting finding is that all fatty acids can be found in subnet 2, also including the ketone body 3-hydroxybutyrate reflecting a catabolic metabolism with increased β-oxidation. This finding may indicate that the metabolites detected in subnet 2, which all can be found in *S*(38), are associated with changes in lipid metabolism. Switches between glucose and lipid oxidation during a 120 min exercise bout followed by 3 h recovery are very likely [31,32]. This hypothesis is confirmed by the time course investigation of free fatty acids, exemplarily shown for palmitic acid in Fig. 5 one of the dominating free fatty acids in blood. This characteristic metabolite of *S*(38) showed significant changes in the studied exercise time interval. The time kinetic of palmitic acid was dominated by a marked increase from 60 to 120 min of exercise and subsequently a rapid drop. Lactate peaked at 60 min and showed a dramatic fall thereafter (data not shown), indicating that anaerobic glycolysis played a major role in the first phase of exercise. We conclude based on our findings that between 60 and 120 min the energy was, at least in part, supplied mainly by the oxidation of fatty acids. The increase in plasma free fatty acids may reflect in this context increased lipolysis in adipose tissue to supply the exercising muscle with free fatty acid as fuel source for the generation of ATP via β-oxidation [33]. Based on these physiological consistent findings we conclude that ICA is suitable to detect physiological interrelationships by extraction from complex non-targeted metabolomics data sets. In addition our analytical strategy simplified the explanation of data resulting from non-hypothesis driven metabolomics approaches.

5. Conclusions

Independent component analysis (ICA) was optimized and subsequently applied to non-hypothesis driven GC–TOF MS metabolomics data of an exercise study. Descriptive statistics showed its ability to optimize ICA by selecting the numbers of principal components after whitening and the independent components. ICA data treatment elucidated conclusive time dependent physiological changes of the metabolic pattern in human plasma before, during and after a single bout of exercise. The dominating detected compounds mainly represent metabolites from the fuel metabolism, i.e. from the most affected metabolic pathways in exercising and recovering humans. ICA can successfully elucidate key metabolite pattern as well as metabolites in metabolic processes and to simplify the explanation of these biological findings.

Acknowledgements

This study was supported by grants from the foundation from the National Basic Research Program of China (No. 2012CB517506) from State Ministry of Science & Technology of China, the Key Foundation (No. 20835006) and the creative research group project (No. 21021004) from National Natural Science Foundation of China, and the Sino-German Center for Research Promotion (DFG and NSFC, GZ 364), the Kompetenznetz Diabetes mellitus (Competence Network for Diabetes mellitus) funded by the Federal Ministry of Education and Research to R.L. and H.U.H. (FKZ 01GI0803-04) and the German Federal Ministry of Education and Research to the German Center for Diabetes Research (DZD eV). The Centre of Inflammation and Metabolism (CIM) is supported by a grant from the Danish National Research Foundation (# 02-512-55). This study was further supported by the Danish Council for Independent Research–Medical Sciences, the Commission of the European Communities (Grant Agreement no. 223576 – MYOAGE), and the Novo Nordisk Foundation. CIM is part of the UNIK Project: Food, Fitness & Pharma for Health and Disease, supported by the Danish Ministry of Science, Technology, and Innovation. CIM is a member of DD2 – the Danish Center for Strategic Research in Type 2 Diabetes (the Danish Council for Strategic Research, grant nos. 09-067009 and 09-075724). The Copenhagen Muscle Research Centre is supported by a grant from the Capital Region of Denmark.

References

[1] L. Eriksson, H. Antti, J. Gottfries, E. Holmes, E. Johansson, F. Lindgren, I. Long, T. Lundstedt, J. Trygg, S. Wold, Anal. Bioanal. Chem. 380 (2004) 419.

[2] M. Defernez, E.K. Kemsley, Trends Anal. Chem. 16 (1997) 216.

[3] X. Li, X. Lu, J. Tian, P. Gao, H.W. Kong, G.W. Xu, Anal. Chem. 81 (2009) 4468.

[4] P. Comon, Signal Process. 36 (1994) 287.

[5] A. Hyvärinen, IEEE Trans. Neural Networks 10 (1999) 626.

[6] N. Murata, S. Ikeda, A. Ziehe, Neurocomputing 41 (2001) 1.

[7] S.I. Amari, A. Hyvarinen, S.Y. Lee, T.W. Lee, V.D. Sanchez, Neurocomputing 49 (2002) 1.

[8] F.J. Gonzalez-Serrano, H.Y. Molina-Bulla, J.J. Murillo-Fuentes, IEEE International Conference on Acoustics, Speech, and Signal Processing, Vols. I–Vi, Proceedings, 2001, 1997.

[9] V.T. Nguyen, J.C. Patra, Advances in Multimedia Information Processing—Pcm 2004, Pt 3, Proceedings, vol. 3333, 2004, p. 364.

[10] D. Glotsos, P. Spyridonos, P. Ravazoula, D. Cavouras, G. Nikiforidis, Neural Inform. Process. 3316 (2004) 1058.

[11] W.M. Zeng, A.Q. Qiu, B. Chodkowski, J.J. Pekar, Neuroimage 46 (2009) 1041.

[12] W. Liebermeister, Bioinformatics 18 (2002) 51.

[13] D. Mantini, F. Petrucci, P. Del Boccio, D. Pieragostino, M. Di Nicola, A. Lugaresi, G. Federici, P. Sacchetta, C. Di Ilio, A. Urbani, Bioinformatics 24 (2008) 63.

[14] Z. Ge, Z. Song, Ind. Eng. Chem. Res. 46 (2007) 2054.

[15] Z. Gea, Z. Song, J. Chemometr. 23 (2009) 636.

[16] Y. Chen, W. Hoehenwarter, W. Weckwerth, Plant J. 63 (2010) 1.

[17] A.J. Overgaard, H.G. Hansen, M. Lajer, L. Pedersen, L. Tarnow, P. Rossing, J.N. McGuire, F. Pociot, Proteome Sci. 8 (2010) 4.

[18] K. Yonekura-Sakakibara, A. Fukushima, R. Nakabayashi, K. Hanada, F. Matsuda, S. Sugawara, E. Inoue, T. Kuromori, T. Ito, K. Shinozaki, B. Wangwattana, M. Yamazaki, K. Saito, Plant J. 69 (2012) 154.

[19] M. Scholz, S. Gatzek, A. Sterling, O. Fiehn, J. Selbig, Bioinformatics 20 (2004) 2447.

[20] S. Trenkamp, P. Eckes, M. Busch, A.R. Fernie, Metabolomics 5 (2009) 277.

[21] F.P.J. Martin, S. Rezzi, I. Montoliu, D. Philippe, L. Tornier, A. Messlik, G. Holzlwimmer, P. Baur, L. Quintanilla-Fend, G. Loh, M. Blaut, S. Blum, S. Kochhar, D. Haller, J. Proteome Res. 8 (2009) 2376.

[22] J.W. Allwood, A. Erban, S. de Koning, W.B. Dunn, A. Luedemann, A. Lommen, L. Kay, R. Loescher, J. Kopka, R. Goodacre, Metabolomics 5 (2009) 479.

[23] B. Ebert, D. Zoeller, A. Erban, I. Fehrle, J. Hartmann, A. Niehl, J. Kopka, J. Fisahn, J. Exp. Bot. 61 (2010) 1321.

[24] H. Führs, A. Specht, A. Erban, J. Kopka, W.J. Horst, J. Exp. Bot. 63 (2012) 329.

[25] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley & Sons, Inc., New York, 2001.

[26] A. Hyvärinen, E. Oja, Neural Networks 13 (2000) 411.

[27] J. Hansen, C. Brandt, A.R. Nielsen, P. Hojman, M. Whitham, M.A. Febbraio, B.K. Pedersen, P. Plomgaard, Endocrinology (2010).

[28] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, J. Chromatogr. A 805 (1998) 17.

[29] Giorgio Tomasi, F.v.d. Berg, C. Andersson, J. Chemometr. 18 (2004) 231.

[30] K. Morgenthal, S. Wienkoop, M. Scholz, J. Selbig, W. Weckwerth, Metabolomics 1 (2005) 109.

[31] J.O. Holloszy, W.M. Kohrt, P.A. Hansen, Front. Biosci. 3 (1998) D1011.

[32] A.E. Jeukendrup, Regulation of fat metabolism in skeletal muscle, in: I. Klimes, E. Sebokova, B.V. Howard, E. Ravussin (Eds.), Lipids and Insulin Resistance: The Role of Fatty Acid Metabolism and Fuel Partitioning, New York Acad Sciences, New York, 2002, pp. 217–235.

[33] J.A. Kanaley, C.D. Mottram, P.D. Scanlon, M.D. Jensen, J. Appl. Physiol. 79 (1995) 439.