Independent component analysis of starch deficient *pgm* mutants

Matthias Scholz, Yves Gibon, Mark Stitt and Joachim Selbig

Max Planck Institute of
Molecular Plant Physiology,
Germany

Abstract: Changes in enzymatic activities in response to carbon starvation were investigated in *Arabidopsis thaliana* in two distinct experiments. One compares the Columbia ecotype (Col-0) and its starch deficient *pgm* mutant (plastidial phosphoglucomutase), the other investigates the enzymatic activities of Col-0 under extended night conditions.

A classical technique for detecting and visualizing relevant information from the measured data is *principal component analysis (PCA)*. We show that *independent component analysis (ICA)* is more suitable for our questions and the results are more precise than those obtained with PCA. This higher informative power is only achieved when ICA is combined with suitable pre-processing and evaluation criteria. It is essential to first reduce the dimensionality of the data set, using PCA. The number of principal components determines the quality of ICA significantly, therefore we propose a criterion for estimating the optimal dimension automatically. The measure of kurtosis is used to sort the extracted components.

We found that ICA could detect on the one hand the time component of the extended night experiment, and on the other hand a discriminating component in the *pgm* mutant experiment. In both components the most important enzymes were the same, confirming the carbon starvation phenotype in the mutant.

Key Words: *Arabidopsis thaliana*, *pgm* mutant, feature extraction

Contact: `scholz@mpimp-golm.mpg.de`

1 Introduction

Techniques for visualizing data sets and for extracting important variables in a ‘blind’ unsupervised way are very helpful for biologists to interpret the given data. Biological background information such as group affiliations (class labels) are not used in *unsupervised algorithms*. Such techniques are an attempt to present the major or global information contained within the data set, independently from the aim of the experiment. An unnoticed supervising effect could appear when adjusting some algorithm parameters by hand. Therefore, we define different criteria for automatic analysis.

One well-established technique for dimensionality reduction and visualization is the classical *principal component analysis (PCA)*, where the extracted information is represented by a set of new variables, termed *components* or *features*. Various PCA-algorithms have

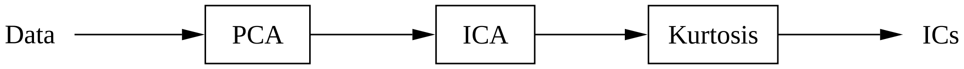

Figure 1: The proposed ICA procedure. First, the data set is reduced by PCA thereby maintaining all of the relevant variances. ICA is applied to this reduced data set and the extracted independent components are sorted by their kurtosis value.

been described in [DK96].

In the field of molecular biology, PCA has become a popular tool for visualizing data sets and for extracting relevant information [WCHB03, UWLK$^+$03]. However, PCA is only powerful if the biological question is related to the highest variance in the data set. Elsewhere, other techniques may be more helpful as shown in [GYHS03] and [JBGS03] for supervised techniques in combination with validation and pre-processing.

More general questions about the underlying data structure are better investigated by an unsupervised technique which would detect relevant components, independently from the background knowledge of the experiment. Such unsupervised concepts allow a better understanding of the molecular response in biological experiments. In addition to experimental characteristics, unexpected factors can also be detected.

Different techniques were developed to overcome the disadvantages of the original PCA. Several extensions of PCA are done in a nonlinear way, for example a nonlinear PCA [SV02] or locally linear embedding [RS00]. However, due to the limited number of samples in molecular data sets, linear alternatives might be more reliable. A very promising linear technique is *independent component analysis (ICA)*. In ICA an independence condition is optimized, which often gives more meaningful components than by optimizing only the variance, as is done by PCA. Because of this the components of ICA are termed *independent components (ICs)*, meaning that different ICs represent different non-overlapping information.

For applying ICA we assume that the observed data are conditioned by unknown fundamental factors, which are independent from each other. By searching for components as statistically independent as possible these required factors can be detected. These fundamental factors are often termed *sources* and the application field is called *blind source separation, BSS*.

The concept of independent component analysis was first proposed by Comon [Co94], with subsequent developments by Bell and Sejnowski [BS95]. One of the first motivations for ICA was sound signal separation. Currently ICA is becoming more important for biomedical applications. Here, applications on time series, such as EEG data [MWJ$^+$02] have to be distinguished from applications on rather static data like gene expression [Li02, MMSM02]. There exists a wide variety of methods for performing ICA. For time series ICA algorithms such as TDSEP [ZM98] have been developed, whereas algorithms such as FastICA [HO00] or [BW04] are more suitable for static data. Detailed descriptions about ICA are given in [HKO01] and in [CA02].

Although the dimensionality of the considered enzymatic data is not as high as that of many metabolite or gene expression data sets, we show that ICA gives optimal results only in conjunction with PCA as a pre-processing step.

ICA is able to extract as many ICs as the data set has dimensions (number of variables).

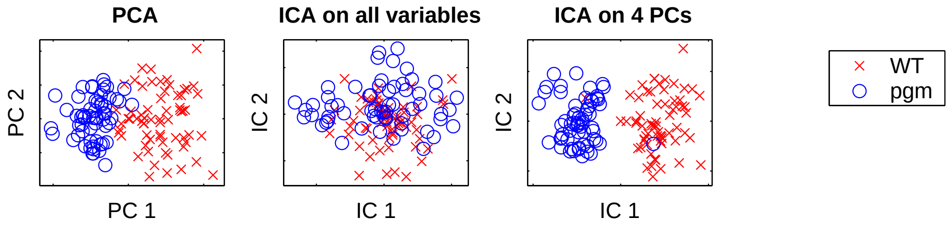

Figure 2: *Pgm* mutant experiment: ICA is compared to PCA. The *pgm* mutant and the wild type are marked differently. On the left, PCA gives already a good result. The first component (PC 1) discriminates the two groups. By applying ICA to the total data set, the result is worse than the result of PCA. However, by using PCA for pre-processing before applying ICA, a more strongly discriminating component can be extracted, as shown on the right.

For technical reasons the ICs have to be sorted. In [Li02], the ICs were sorted by a combination of a contrast function and a variance criterion. Here we capture first the relevant variances by PCA and then the ICs can be extracted and sorted without considering the variance. The kurtosis distribution measure is used for ranking the independent components. The proposed procedure is illustrated in Figure 1, see also [SGS$^+$04]
We have established a set of microplate-based assays for 17 enzyme activities from various metabolic pathways in rosette leaves of *Arabidopsis thaliana* (Gibon et al., in preparation). We applied this first platform to compare the evolutions of transcript levels and the corresponding enzyme activities within a diurnal cycle in the wild type and in the starch less mutant *pgm*. We then determined these activities in plants submitted to an extended night, to investigate the response to carbon starvation and to complement results obtained at both transcript and metabolite level [TBG$^+$04].

2 PCA – pre-processing

With PCA a reduced data set with maximal variance is searched for. By applying PCA for visualization it is assumed that the most interesting information will be directly related to the highest variance in the data. The best projection or visualization is then given by the first two principal components (PCs) of highest variance. However, this is often a false assumption, and the relevant information is not related to the highest amount of variance, but can nevertheless be related to a significantly high level. Thus PCA can be used to reduce the high dimensionality of the data while maintaining the relevant variances. Such a pre-processing step can preserve all of the relevant information and reduces only the noise given by small variances.

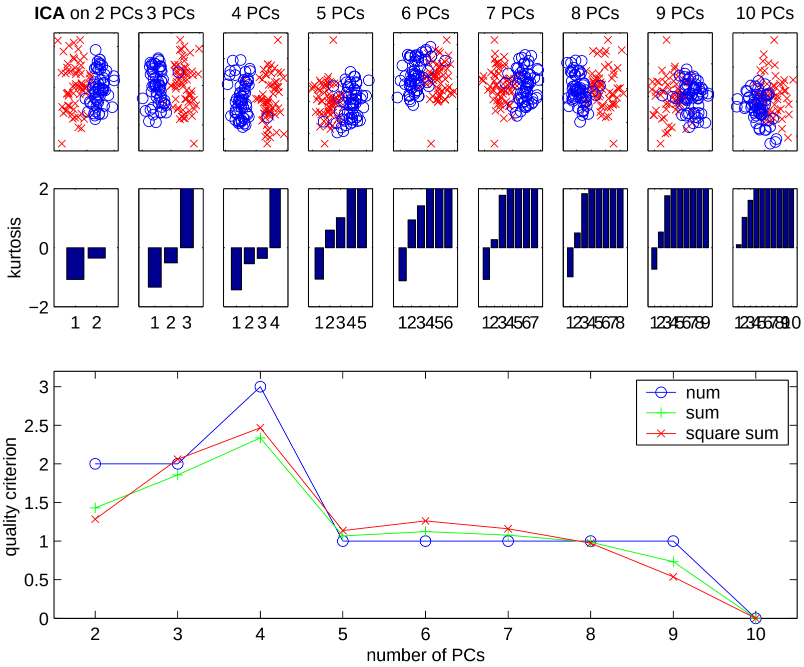

Figure 3: ICA is applied to different numbers of principal components (PCs). On the top row IC 1 is plotted against IC 2 for different numbers of PCs. The best separation between the two groups can be found at 4 PCs. In the second row the kurtosis values of each IC are plotted. At 4 PCs ICA detects the highest number of ICs with a negative kurtosis value (‘num’). Below the proposed criteria for detecting the optimal number of PCs are plotted. All criteria have their optimum at 4 PCs.

3 ICA – independent component analysis

As the data set has been reduced in PCA by optimizing the variance, independent component analysis (ICA) can now be applied to this reduced data set for optimizing an independence condition.

Similarly to PCA, ICA extracts also a set of components. In contrast to PCA these components are constructed in order to minimize the dependence and are therefore termed *independent components (ICs)*. Independence is a stronger condition than the non-correlation in PCA and gives often more meaningful components. The components of ICA do not have to fulfill an orthogonality condition.

To generate independent components, different criteria (contrast functions) can be optimized: higher-order dependencies, entropy or kurtosis. In this article, ICA was performed by the CuBICA4 algorithm [BW04], which provides good and reproducible results.

4 Significant components - kurtosis

ICA is able to extract as many components as the data set has dimensions. These components have no order. For practical reasons we had to define a criterion for sorting these components to our interest. One measurement which can match our interest very well, is kurtosis.
Kurtosis is a classical measure of non-Gaussianity, and is computationally and theoretically relatively simple. It indicates whether the data are peaked or flat, relative to a Gaussian (normal) distribution. A Gaussian distribution has a kurtosis of zero. Positive kurtosis indicates a 'peaked' distribution (super-Gaussian) and negative kurtosis indicates a 'flat' distribution (sub-Gaussian).

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

where $z = (z_1, z_2, ..., z_n)$ represents a variable or component with mean μ and standard deviation σ, n is the number of samples. The kurtosis is the fourth auto-cumulant after mean (first), variance (second), and skewness (third).

From purely Gaussian distributed data, no unique independent components can be extracted, therefore, ICA should only be applied to data sets where we can find components that have a non-Gaussian distribution.
Examples of super-Gaussian distributions (highly positive kurtosis) are speech signals, because these are predominantly close to zero. However, for molecular data sub-Gaussian distributions (negative kurtosis) are more interesting. Negative kurtosis can indicate a cluster structure or at least a uniformly distributed factor. The former can resolve between two experimental conditions (high and low enzymatic responses), whereas the latter can represent a continuously changed experimental factor such as the temperature or the light intensity. Thus the components with the most negative kurtosis can give us the most relevant information.

5 Optimal PCA-dimension

By using the PCA as a pre-processing step, the number of PCs, hence the optimal reduced dimensionality is usually unknown. Thus we had to find a way to estimate this dimension. Here, the estimation was aligned with the aim of our analysis, i.e. to find as many relevant components as possible. As a negative kurtosis indicates relevant components, the dimension, where we can extract the highest number of independent components with negative kurtosis is the optimal dimension.
As an alternative to counting simply the number of components with negative kurtosis, the square sum over these negative values can be used. This might be a more reliable criterion, as a kurtosis close to zero has little effect.

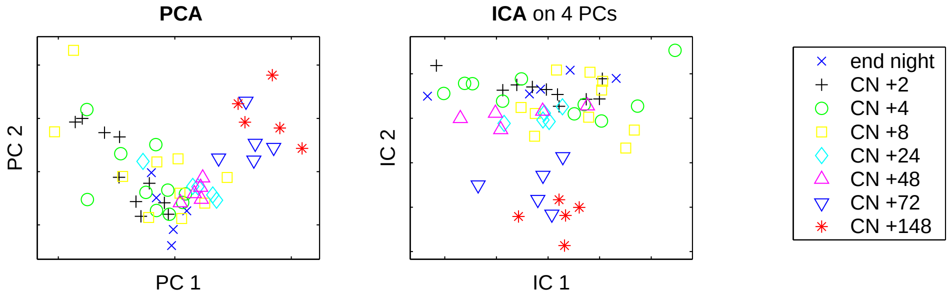

Figure 4: Continuous night (CN) experiment. PCA is compared to ICA. The time steps of the experiment are marked differently. Already in PCA a separation of the late time steps ($72h$ and $148h$) can be found, but the direction of this time factor is not directly related to PC 1 or PC 2, it is on the diagonal. ICA is able to arrange this time factor to a single component, IC 2.

6 Influence values

As the detected independent components often have a biological interpretation, it would be important to know, which variables (enzymes) have the highest influence on these components. The influences are given by the transformation matrices of PCA and ICA and are also termed *loadings* or *weights*.

PCA transforms a d-dimensional sample vector $x = (x_1, x_2, \ldots, x_d)^T$ into a usually lower dimensional vector $y = (y_1, y_2, \ldots, y_k)^T$, where d is the number of masses and k is the number of selected components. The PCA transformation is given by the eigenvector matrix V, $y = Vx$. Similarly, ICA transforms this vector y to the required vector $z = (z_1, z_2, \ldots, z_k)^T$, containing the independent values z_i for each IC i. For that a demixing matrix W is estimated by ICA, $z = Wy$. V gives the influences of each variable (mass) on each of the PCs, whereas W gives us the influence of each PC on each of the ICs. We can combine both matrices $U = W * V$ to a direct transformation $z = Ux$, where U gives vector-wise the required influences of each variable on each of the ICs.

7 Experiment

One experiment compares enzyme activities from the wild type Columbia (Col-0) and the *pgm* mutant, which cannot synthesize starch. The other experiment investigates the response to carbon starvation, obtained by extending the night for up to $148h$. The responses of 17 different enzymes were investigated. The number of samples is 125 in the mutant experiment and 55 in the continuous night experiment. The variables (enzymes) are normalized to unit variance.

Continuous night experiment **IC 2: time factor**		**Mutant experiment** **IC 1: Mutant ↔ WT**	
Enzyme	infl.	Enzyme	infl.
AcidInvertase ←	-0.26	AcidInvertase ←	-0.24
GLDHam ←	-0.22	GLDHam ←	-0.23
ShikD	0.19	Fd-GOGAT	0.16
Fumarase	0.15	NR Vmax	0.16
Glycerokinase	-0.11	NAD-GAPDH	0.12
PEPCase	0.11	NADP-GAPDH	0.12
NR Vmax	0.10	Glycerokinase	0.11
NADP-GAPDH	-0.09	ShikDH	0.10
AspAT	0.08	G6PDH	-0.10
Glycerokinase	0.07	Fructokinase	0.10

Table 1: Enzyme influence. The 10 enzymes of highest influence are given for IC 2 of the continuous night experiment and for IC 1 of the *pgm* mutant experiment. The first two most important enzymes are identical, and hence the continuous night time component is quite similar to the discriminating component of the *pgm* mutant experiment.

In the *pgm* mutant experiment, PCA already gives a good result. That means that the relevant experimental conditions are represented by a high amount of variance in the enzymatic data, but this is not necessarily the optimal projection of the data. The result can be improved by applying ICA in the proposed procedure. The first component of ICA (IC 1) has a higher discriminating power than the first component of PCA (PC 1), see Figure 2. The most relevant independent components are detected in an automatic manner. All proposed criteria point to the optimal number of PCs in PCA pre-processing, see Figure 3. Note that such a high discriminating component could not detected by applying ICA to all variables without PCA pre-processing.

In the continuous night experiment, the projection of the first two principal components shows that the early time steps are close to each other and only the last two time steps (+72h and +148h) are distinct from the others. However, the direction of time information is not directly assigned to one of the principal components, it is on the diagonal. In contrast to this, ICA is able to arrange this time component automatically to one of the independent components, IC 2, see Figure 4

For both interpreted components, the discriminating component IC 1 of the *pgm* mutant experiment and for the time component IC 2 of the continuous night experiment, the first two most important enzymes are identical as can be seen in Table 1.

It would be interesting to know at what time point a *pgm* mutant has an identical response to a wild type experiencing an extended night. For this diagnostic task, the *pgm* enzyme data are projected into the independent component space of the continuous night experiment by using the transformation matrix of continuous night. The *pgm* mutant shows the same enzymatic response as a wild type under 48h to 72h extended night condition, see Figure 5.

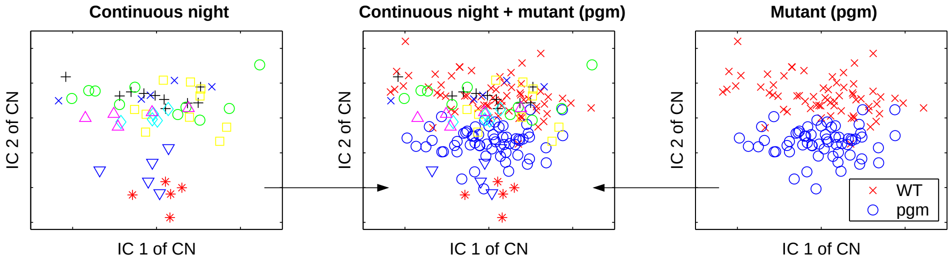

Figure 5: The mutant (pgm) experiment is joined with the continuous night (CN) experiment. On the left, the ICA result of the continuous night experiment from Figure 4 is shown. On the right the *pgm* samples are transformed into the component space of the continuous night experiment. The plot shows that the time component IC 2 of the continuous night experiment have also some discriminating effect in the *pgm* experiment. In the middle both data sets are superimposed. The *pgm* samples fall into the region within $48h$ and $72h$ of the continuous night experiment.

8 Conclusion

We have demonstrated that independent components of ICA can have greater discriminating power and can be more intepretable than the principal components of PCA. This higher informative power is only achieved when ICA is combined with suitable pre-processing and evaluation criteria.

The kurtosis measure is used for estimating the optimal number of principal components (PCs) in the PCA pre-processing step and is also used for sorting the detected independent components (ICs). Applied to the *pgm* mutant experiment, the first independent component (IC 1) discriminates between the *pgm* mutant and the wild type. In the continuous night experiment, the first component could not be interpreted and might be an artefact, but the second component (IC 2) could be interpreted as the time component of the experiment. We found that the two most strongly implicated enzymes are identical in both interpreted components. Thus, as expected, the starch deficient *pgm* mutant shows similar behaviour to a wild type plant grown in darkness. This could also be shown by combining both data sets.

Although in this study ICA is applied to enzymatic data, the proposed approach is not restricted to these kind of data and could be applied to high dimensional data from metabolomic, proteomic and transcriptomic investigations. The described approach is available for public use in *MetaGeneAlyse* [DKS03], a web-based analysis tool for molecular biology.

9 Acknowlegement

We would like to thank Gareth Catchpole for helpful comments on the manuscript.

References

[BS95] Bell, A. J. and Sejnowski, T. J.: An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*. 7:1129–1159. 1995.

[BW04] Blaschke, T. and Wiskott, L.: CuBICA: Independent component analysis by simultaneous third- and fourth-order cumulant diagonalization. *IEEE Transactions on Signal Processing*. 52(5). 2004.

[CA02] Cichocki, A. and Amari, S.: *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. Wiley. 2002.

[Co94] Comon, P.: Independent component analysis, a new concept? *Signal Processing*. 36(3):287–314. 1994.

[DK96] Diamantaras, K. and Kung, S.: *Principal Component Neural Networks*. Wiley. New York. 1996.

[DKS03] Daub, C. O., Kloska, S., and Selbig, J.: MetaGeneAlyse: analysis of integrated transcriptional and metabolite data. *Bioinformatics*. 19(17):2332–2333. 2003. http://metagenealyse.mpimp-golm.mpg.de/.

[GYHS03] Goodacre, R., York, E. V., Heald, J. K., and Scott, I. M.: Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry. *Phytochemistry*. 62(6):859–863. 2003.

[HKO01] Hyvärinen, A., Karhunen, J., and Oja, E.: *Independent Component Analysis*. J. Wiley. 2001.

[HO00] Hyvärinen, A. and Oja, E.: Independent component analysis: Algorithms and applications. *Neural Networks*. 4–5(13):411–430. 2000.

[JBGS03] Johnson, H. E., Broadhurst, D., Goodacre, R., and Smith, A. R.: Metabolic fingerprinting of salt-stressed tomatoes. *Phytochemistry*. 62(6):919–928. 2003.

[Li02] Liebermeister, W.: Linear modes of gene expression determined by independent component analysis. *Bioinformatics*. 18(1):51–60. 2002.

[MMSM02] Martoglio, A. M., Miskin, J. W., Smith, S. K., and MacKay, D. J. C.: A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. *Bioinformatics*. 18:1617–1624. 2002.

[MWJ$^+$02] Makeig, S., Westerfield, M., Jung, T.-P., Enghoff, S., Townsend, J., Courchesne, E., and Sejnowski, T. J.: Dynamic Brain Sources of Visual Evoked Responses. *Science*. 295(5555):690–694. 2002.

[RS00] Roweis, S. T. and Saul, L. K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*. 290(5500):2323–2326. 2000.

[SGS$^+$04] Scholz, M., Gatzek, S., Sterling, A., Fiehn, O., and Selbig, J.: Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics Advance Access published on April 15*. 2004. doi:10.1093/bioinformatics/bth270.

[SV02] Scholz, M. and Vigário, R.: Nonlinear PCA: a new hierarchical approach. In: Verleysen, M. (Ed.), *Proceedings ESANN*. pp. 439–444. 2002.

[TBG$^+$04] Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L. A., Rhee, S. Y., and Stitt, M.: MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal*. 37(6):914–939. 2004.

[UWLK$^+$03] Urbanczyk-Wochniak, E., Luedemann, A., Kopka, J., Selbig, J., Roessner-Tunali, U., Willmitzer, L., and Fernie, A. R.: Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO reports*. 4(10):989–993. 2003.

[WCHB03] Ward, J. L., C. Harris, J. L., and Beale, M. H.: Assessment of ^{1}H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsis thaliana*. *Phytochemistry*. 62(6):949–957. 2003.

[ZM98] Ziehe, A. and Müller, K.-R.: TDSEP - an efficient algorithm for blind separation using time structure. In: *Proc. ICANN'98, Int. Conf. on Artificial Neural Networks*. pp. 675–680. 1998.