Real-life metabolomics data analysis: how to deal with complex data?
Rubingh, C.M.

Link to publication

Citation for published version (APA):
Rubingh, C. M. (2010). Real-life metabolomics data analysis: how to deal with complex data?

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
The real life of CARINA MANON RUBINGH began at March 26, 1976, when she was born in Soest, the Netherlands. From that moment she wanted to understand why things are happening and how things are happening; she wanted to investigate life. After the common years of primary and secondary school, she continued her search at the Wageningen University (formerly, Wageningen Agricultural University), The Netherlands, from whom she received her Master degree in ‘Nutrition and Health’ in 1999. She started her professional career at TNO in Zeist, the Netherlands, in that same year.

Due to the diversity of research that is performed at TNO, some questions regarding life were answered. However, some topics got her attention to a great extent, which made her own contribution to the understanding of life, even if it is only for a bit. This thesis is a result of it.
REAL-LIFE METABOLOMICS DATA ANALYSIS:
HOW TO DEAL WITH COMPLEX DATA?

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 9 november 2010, te 12.00 uur

door Carina Manon Rubingh
geboren te Soest
Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The publication of this thesis was made possible by TNO.
### TABLE OF CONTENTS

1. **INTRODUCTION**  

2. **MATRIX CORRELATIONS FOR HIGH-DIMENSIONAL DATA: THE MODIFIED RV-COEFFICIENT**  

3. **ANALYZING LONGITUDINAL MICROBIAL METABOLOMICS DATA**  

4. **DISCOVERY OF SUBTLE EFFECTS IN A HUMAN INTERVENTION TRIAL THROUGH MULTILEVEL MODELING**  

5. **METABOLIC PROFILING OF THE RESPONSE TO AN ORAL GLUCOSE TOLERANCE TEST DETECTS SUBTLE METABOLIC CHANGES**  

6. **ASSESSING THE PERFORMANCE OF STATISTICAL VALIDATION TOOLS FOR MEGAVARIATE METABOLOMICS DATA**  

7. **THE COSTS OF COMPLEX MODEL OPTIMIZATION**  

8. **OUTRODUCTION**  

   **SUMMARY**  

   **SAMENVATTING**  

   **NAWOORD**
**analysis** (ə-nal'ə-sis) n. pl. –ses (-sēz) 1 The resolution of a whole into its parts or elements; opposed to synthesis. 2 A statement of the results of this; logical synopsis. 3 A method of determining or describing the nature of a thing by resolving it into its parts; to study literature by the analysis of texts. 4 Math. a The process of resolving a problem into its first elements. b the investigation of the relations of variable or indeterminate quantities by means of symbols, including some branches of algebra and differential and integral calculus.

**complex** (kem-pleks', kom'plex) adj. 1 Consisting of various parts or elements; composite. 2 Complicated; involved; intricate.

**data** (dā'tə, dat'ə, dä'tə) n. pl. of DATUM Facts or figures from which conclusions may be drawn; often constructed as a singular.

**explore** (ik-splôr’, -splôr’) v. -plored, -ploring v.t. 1 To search through or travel in or over, as new lands, for discovery. 2 To look into carefully; scrutinize.

**life** (līf) n. pl. lives (livz) 1 That state in which animals and plants exits which distinguishes them from inorganic substances and from dead organisms: characterized by metabolism and growth, reproduction, and internally initiated adaptations to the environment.

**metabolism** (mә-tәb’ә-liz’em) n. Biol. The aggregate of all physical and chemical processes constantly taking place in living organisms, including those which use energy to build up assimilated materials (anabolism) and those which release energy by breaking them down (catabolism). Also me·tab’o·ly (-ә-lē).

**real** (rē’әl, rēl) adj. 1 Having existence or actuality as a thing or state; not imaginary; a real event. 2 Being in concordance with appearance or claim; genuine; not artificial or counterfeit. 3 Representing the true or actual, as opposed to the apparent or ostensible; the real reason. 4 Unaffected; unpretentious; a real person. 5 Philos. Having actual existence, and not merely possible, apparent, or imaginary.

**structured, structuring** 1 To form into an organized structure; build. 2 To conceive as a structural whole.

**validate** (val’ә-dât) v.t. -dat-ed, -dat-ing 1 To make valid; ratify and confirm. 2 To declare legally valid; legalize.

Introduction
The homo sapiens is a curious species. He wants to know what is happening around him, and even more, he wants to understand why things are happening and how things are happening. Some of his questions may be spiritual, religious, or psychological: why do I exist? Is there life after death? How is it possible that you give me a call at the moment I’m thinking about you? Other questions may be more ‘down to earth’ and more directly related to himself: Why do I get bold? Now I’m gaining weight, will I get sick? Why am I so tired? Spiritual or down to earth, most questions are related to ‘life’: the homo sapiens wants to understand life.

There is no scientific discipline that can explain all of these questions and the combination of all of them may explain only a little bit about life. But it is a start in feeding the curiosity of the human species. There is a branch that meets the need for answers to questions like: Why do some people get diabetes and others don’t? Why is that certain bacteria producing more than that other bacteria? Why are obese at higher risk for developing cardiovascular diseases? How can we explain the health effects of a Mediterranean diet? These questions are all related to response of biological systems on environmental influences due to, for instance, toxicological exposure, nutrition or medical treatment, which can be investigated by metabolomics studies. These studies are set up to determine which metabolites are responsible for the effects of interest.

Metabolites in biological samples like plasma or urine are measured using various analytical techniques, which can generate a large amount of data containing information about a large number of correlated variables. This correlation can be twofold: 1) analytical correlations which are generated depending on the data pre-processing tool that is used (i.e. in case of peak picking or the derivatization method that is used, which result in the fact than one metabolite might be represented in the data set by more than one peak/variable), and 2) biological correlations, i.e. metabolites that correlate mutually, for instance a substrate and product of an enzymatic reaction operating close to equilibrium, or metabolites that give a similar response to a perturbation. The presence of these correlation structures makes it a challenge to get the relevant information out of the data.

Although basic statistical tools, like Analysis of Variance (ANOVA), may give some insight in the effects of interest, they are in fact inappropriate for the analysis of such amounts of data, because they do not take the correlation structure between variables into account. Effects of a metabolite might remain hidden if it is not contributing by itself but only in combination with another metabolite. Therefore, multivariate data analysis (MVA) is often used to analyze metabolomics data. MVA can be used to summarize the data by reducing dimensions of data, for regularization purposes, for variable selection, etcetera.

At the early days of metabolomics studies, MVA techniques like PCA and PLS were sufficient for statistical analysis. But nowadays, the metabolomics studies become more and more complicated and the urge for better and more sophisticated data analysis techniques is
increasing. The present thesis is a proof of this. Several challenges were met during metabolomics studies that were mainly performed at TNO Quality of Life (Zeist, The Netherlands). In the next chapters, solutions are given to meet these challenges.

**Outline of this thesis**

At first sight, the chapters of this thesis might give the impression that there is limited relationship between them:

In Chapter 2, a modified version of a matrix correlation measure is given and illustrated using metabolite data that were measured in *Escherichia coli*. Data of a microbial longitudinal study is used to demonstrate the power of a well-defined experimental design in combination with metabolomics and multi-way data analysis in Chapter 3. In Chapter 4 subtle treatment effects in slightly overweight men are discovered through multilevel modeling, which are discussed in a biological context in Chapter 5. In Chapter 6 the performance of statistical model validation tools such as cross-validation, jack-knifing model parameters and permutation tests are assessed, using a metabolomics data set based on measurements in plasma of lean and obese subjects. The costs for complex model optimization are calculated in Chapter 7.

Data of five different studies are used, obtained from different measurement platforms and analysed in different ways. However, they have some essential characteristics in common: it is all about metabolomics data, it is all about data handling, and, most important, it is all based on the *Real life*. Therefore, the methods and solutions that are described in this thesis are directly applicable in all other (biological) studies in which similar issues need to be addressed, as also is stated in the outroduction that is given in Chapter 8.

**Bridging**

If a closer look is taken at the subjects in this thesis, one will notice that the chapters have even more in common, besides the fact that it is about real life metabolomics. It is about the way complex metabolomics data needs to be addressed: *explore, model and validate*.

If new data is obtained, it must be explored before the actual data analysis is performed. A simple Principal Component Analysis can already give a lot of information about the data set: are there any expected or perhaps unexpected patterns seen, are there remarkable objects or are there variables that dominate? Another way to explore data is to calculate the correlation between data sets. There are many measures that can be used to express this correlation. A strong correlation between two data sets means that a process that is found to be important in one of the data sets may be confirmed by the other data set. A low correlation, on the other hand, may indicate that different processes are involved. Common used correlation measures cannot unthinkingly be used for high-dimensional data sets, as will be illustrated in this thesis.
Once data are explored, the modeling part can be started. Unnecessary to say that the type of modeling should be in line with the question that is to be answered. Modeling for the sense of modeling makes no sense. In case of complex metabolomics data it is very important to carefully select the modeling method. Relatively simple techniques like Principal component analysis or Partial Least Squares Regression are often not sufficient to reveal the subtle effects that are hidden in the data. More sophisticated methods which take into account the complex design of the data are in that case more successful, as will be illustrated in this thesis.

If the right method is chosen, it will hardly be a problem to find a statistical model when fitting high-dimensional metabolomics data. This sounds promising but it is a problem on its own: a good fit will not guarantee stable and reliable findings. A thorough validation is crucial in order to trust the results. Validation tools like cross-validation, jack-knifing or a permutation test will give insight in the reliability of the statistical model. However, these methods will also not bring universal happiness either, as will be illustrated in this thesis.
MATRIX CORRELATIONS FOR HIGH-DIMENSIONAL DATA: 

THE MODIFIED RV-COEFFICIENT

Age K Smilde¹, Henk AL Kiers², Sabina Bijlsma³, Carina M Rubingh³, Marjan J van Erk³

¹Biosystems Data Analysis, SILS, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam

²Heymans Institute, University of Groningen, Groningen

³TNO Quality of Life, Utrechtseweg 48, 3704 HE Zeist, The Netherlands

Bioinformatics 2009, 25. 401–405
Abstract

Motivation: Modern functional genomics generates high-dimensional datasets. It is often convenient to have a single simple number characterizing the relationship between pairs of such high-dimensional datasets in a comprehensive way. Matrix correlations are such numbers and are appealing since they can be interpreted in the same way as Pearson's correlations familiar to biologists. The high-dimensionality of functional genomics data is, however, problematic for existing matrix correlations. The motivation of this article is 2-fold: (i) we introduce the idea of matrix correlations to the bioinformatics community and (ii) we give an improvement of the most promising matrix correlation coefficient (the RV-coefficient) circumventing the problems of high-dimensional data.

Results: The modified RV-coefficient can be used in high-dimensional data analysis studies as an easy measure of common information of two datasets. This is shown by theoretical arguments, simulations and applications to two real-life examples from functional genomics, i.e. a transcriptomics and metabolomics example.

Availability: The Matlab m-files of the methods presented can be downloaded from http://www.bdagroup.nl.
Introduction

Functional genomics research generates high-dimensional data, e.g. transcriptomics, proteomics or metabolomics data. The central characteristic of these types of data is the low sample-to-variable ratio. Transcriptomics (or gene-expression) data typically has thousands of variables and the number of samples is in the order of tens to hundred. Similar characteristics hold for proteomics and metabolomics data. Often multiple datasets are available (i.e. measured) on the same samples of the biological system. This calls for data fusion methods: methods that are able to extract the mutual information from all datasets simultaneously (Alter et al., 2003).

A first useful step in such a data fusion strategy is to probe the similarity between pairs of datasets in a simple and comprehensive way (Smilde et al., 2005b). Matrix correlations can be used for this purpose. These correlations take values between zero and one, defining a scale of similarity between two matrices. This scale can be interpreted in much the same way as the absolute value of the Pearson correlation coefficient known to biologists. Hence, its use in functional genomics data fusion can be straightforward.

Matrix correlations have already a long history in multivariate analysis (Robert and Escoufier, 1976; Yanai, 1974). A comprehensive overview is given in Ramsay et al. (1984). For this article, we focus our attention on the RV-coefficient as a typical example of a matrix correlation already in use in metabolomics (Smilde et al., 2005b). While using the RV-coefficient in a transcriptomics study, we ran into problems: the RV-coefficient gave high values in almost all cases. This pointed to trivial results. We explain this trivial result (i.e. the break-down of the RV-coefficient for high-dimensional data) and give a solution to circumvent this unwanted behavior.

Methods

Matrix correlations

The idea of a matrix correlation is to provide a measure of the similarity of matrices. We start our explanation with matrices $X(I,J)$ and $Y(I,J)$ sharing the row-mode. The latter means that different types of measurements, e.g. transcriptomics and metabolomics, are performed on the same physical samples (the requirement that both matrices have an equal number of columns will be relaxed later). The mapping $r: \mathbb{R}^{IJ} \times \mathbb{R}^{IJ} \to [0,1]$ is called a correlation function if for all non-zero scalars $a$ and $b$ for $X$ and $Y$ not both zero holds that

$$
C_1: r(aX,Y) = r(X,bY) = r(X,Y)
$$

$$
C_2: r(X,Y) = r(Y,X)
$$

$$
C_3: r(X,Y) = 1 \text{ if } X = bY
$$

$$
C_4: r(X,Y) = 0 \text{ if } fX^TY = 0
$$

where iff is the abbreviation of if and only if (Ramsay et al., 1984). Matrices can be similar in a variety of ways; this means that rule $C_3$ can be changed, e.g. $X$ and $Y$ can have a correlation of one if they only differ by an orthogonal rotation ($X = YQ$ with $QQ^T = I$). In that case, the arrangement of the $I$ points (rows) of $X$ and those of $Y$ is essentially equal.
apart from the rotation. An example of a matrix correlation satisfying C1 to C4 is the absolute value of

\[ r_{\text{in}}(X, Y) = \frac{\text{tr}(X^T Y)}{\sqrt{\text{tr}(X^T X) \text{tr}(Y^T Y)}} \]  

which is based on the inner product of two matrices.

A commonly used matrix correlation which allows for a different number of columns in \( X(I \times J_1) \) and \( Y(I \times J_2) \) is the RV-coefficient (Robert and Escoufier, 1976):

\[ \text{RV}(X, Y) = \frac{\text{tr}(XX^T YY^T)}{\sqrt{\text{tr}(XX^T)^2 \text{tr}(YY^T)^2}} \]  

which is an orientation independent measure, i.e. rotations of the two matrices do not affect the RV-coefficient (it satisfies C1, C2 and C4, as well as a relaxed version of C3; Appendix). This is usually a desirable property since in many functional genomics applications similarities of the configuration of the samples generated by the two matrices is of interest and not their specific orientation. Stated otherwise, the relationships between the samples are of interest not their absolute positions in space.

The RV-coefficient can also be written using the singular value decomposition (SVD) of both \( X \) and \( Y \):

\[
X = U_1 D_1 V_1^T = T_1 V_1^T \\
Y = U_2 D_2 V_2^T = T_2 V_2^T
\]

where \( U_1 \) and \( U_2 \) are \( I \times I \) orthogonal matrices; \( D_1 \) and \( D_2 \) are \( I \times I \) diagonal matrices with the singular values of \( X \) and \( Y \), respectively, on their diagonals; \( V_1 (J_1 \times I) \) and \( V_2 (J_2 \times I) \) are column-orthogonal matrices (\( V_1^T V_1 = V_1^T V_2 = I \)). Then it holds that (Ramsay et al., 1984)

\[ \text{RV}(X, Y) = r(U_1 D_1^2 U_1^T, U_2 D_2^2 U_2^T) \]  

which can easily be verified by substitution and links the two matrix correlation coefficients \( r \) and RV. Equation (5) shows that directions in \( X \) and \( Y \) with more importance (i.e. with high singular values) are given more importance in calculating the RV-coefficients. This property is valuable in high-dimensional data because the interest is usually in communality between important dimensions of the matrices.

Alternative expressions for the RV are

\[ \text{RV}(X, Y) = \frac{\text{ssq}(Y^T X)}{\sqrt{\text{ssq}(XX^T) \times \text{ssq}(YY^T)}} \]  

or

\[ \text{RV}(X, Y) = \frac{\text{Vec}(XX^T) \text{Vec}(YY^T)}{\sqrt{\text{Vec}(XX^T) \times \text{Vec}(YY^T)}} \]

where \( \text{ssq} \) means sum-of-squares (the sum of squares of all elements of the corresponding matrix) and \( \text{Vec}(X) \) is the symbol for the vectorized version of \( X \) (see Appendix). The similarity with the Pearson correlation between two vectors \( x \) and \( y \) becomes clear when writing the latter as
\[ r_p(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \]  \tag{8}

where \( x_i, y_i \) and \( \bar{x}, \bar{y} \) are the typical elements and means, respectively, of the vectors \( x, y \). Rewriting (8) gives

\[ r_p(x, y) = \frac{\bar{x}'y}{\sqrt{\bar{x}'x} \sqrt{\bar{y}'y}} \]  \tag{9}

where \( \bar{x} \) is the column-centered version of \( x \) and likewise for \( \bar{y} \). Since \( \text{Vec}(XX') \) and \( \text{Vec}(YY') \) in (7) play the same roles as \( \bar{x} \) and \( \bar{y} \) it is clear that the RV-coefficient can be interpreted as a correlation coefficient. The interpretation of RV as an association measure becomes even more evident when the uncentered correlation is used, or Tucker’s congruence coefficient (Lorenzo-Seva and Ten Berge, 2006)

\[ r_p(x, y) = \frac{x'y}{\sqrt{x'x} \sqrt{y'y}} \]  \tag{10}

which shows that the RV-coefficient bears also similarities with Tucker’s congruence coefficient.

The RV-coefficient is only independent of a rotation or an overall scaling of the matrices \( (RV(X, Y) = RV(\alpha XQ_1, \beta YQ_2) \) for non-zero \( \alpha \) and \( \beta \) and orthogonal \( Q_1 \) and \( Q_2 \). All other preprocessing operations are influencing the RV coefficient, e.g. centering has a profound effect similar to the difference between centered and uncentered correlations. Hence, the user has to make a choice regarding the preprocessing and, thus, the metric in which to compare the matrices. Recommendations to this end are available in the literature (Bro and Smilde, 2003; van den Berg et al., 2006).

![Figure 1](image_url)

**Figure 1.** Simulation of RV-coefficients with different numbers of samples. Plotted is the mean, minimum and maximum for 100 repeats each of the RV as a function of the number of samples.
Problems with the RV-coefficient

While investigating two gene-expression dataset of sizes 5×130 and 5×113 from a functional genomics experiment (Kleeman et al., 2007) high RV-coefficients were found (values between 0.5 and 0.99). These could neither directly be understood from the underlying biology nor from independent calculations more extensively investigating the similarities between the two datasets. Hence, a small set of initial simulations was performed where random matrices of the same sizes were generated and RV-coefficients calculated. Despite the fact that the matrices were drawn from random numbers (N(0,1)) the RV-coefficient was always high. Increasing the sample size of the random matrices to 100 samples showed that the RV-coefficient depends on the sample size. In an extensive set of simulations with random numbers, the samples sizes were systematically increased while the other sizes (130 and 113) remained the same. The result is shown in Figure 1 and shows the problematic behavior. The behavior of the RV-coefficient was also investigated for moderate and strongly unequal sizes of the matrices. Figure 2 shows that the problematic behavior is already visible at much lower numbers of variables (simulations performed similarly as the ones of Figure 1). The reason for the unwanted behavior is as follows. According to (7), the RV-coefficient can be written as

\[ RV(X, Y) = a'b \]  

with

\[ a = \frac{\text{vec}(XX')}{\text{ssq}(XX')^{1/2}} \] 

\[ b = \frac{\text{vec}(YY')}{\text{ssq}(YY')^{1/2}}. \]

Now, suppose \( X \) and \( Y \) are fully random matrices, with elements drawn from standard normal distributions. When \( J_1 \) is large, \( XX' \) can be expected to have diagonal elements close to \( J_1 \) and off-diagonal elements close to 0.

Figure 2. Simulation of RV-coefficient with different numbers of samples and variables. Plotted are the means of the RV-coefficient from 100 repeats.
Specifically, we can write
\[
[XX']_{ii} = x'_i x_i = J_i + \varepsilon^X_{ii},
\]
\[
[XX']_{ij} = x'_i x_j = \varepsilon^X_{ij}, i \neq j,
\]
where \(x'_i\) denotes the \(i\)-th row of \(X\); the values \(\varepsilon^X_{ii} (i = 1, \ldots, I)\) can be considered as random draws from a distribution with zero mean and (the same) standard deviations \(\sigma_x\). Likewise, the values \(\varepsilon^X_{ij} (i, j = 1, \ldots, I, i \neq j)\) can be considered as random draws from a distribution with zero mean and (the same) standard deviations \(\tau_x\). The reasons for these distributional properties are as follows. Because \(x'_i\) has random elements from a standard normal distribution, \(E(x'_i x_i)\), i.e. the expected value of a sum of \(J_i\) squares of such values, equals \(J_i\); the variation across realizations of \(x'_i x_i\) is the same for all \(i\), because the elements of the vectors \(x'_i\) are drawn from the same distributions. Likewise, because \(x'_i\) and \(x'_j\) have independent random elements, \(E(x'_i x_j) = 0\), and the variation across realizations of \(x'_i x_j\) is the same for all \(i, j\) because the elements of all vectors \(x'_i\) and \(x'_j\) are drawn from the same distributions. When \(J_i\) is large, \(\sigma_x\) and \(\tau_x\) can be expected to be small compared with \(J_i\). Indeed, when the elements of \(X\) are drawn from standard normal distributions, it holds that \(\sigma_x = \sqrt{2J_i}\) and \(\tau_x = \sqrt{\frac{J_i}{J}}\), as follows from general results on stochastic theory for sums and products (Mood et al., 1974).

Using the above distributional results, we can give approximate descriptions of the normalized vectors \(a\) and \(b\). The numerator of \(a\) is the vector with \(I\) values of \((J_i + \varepsilon^X_{ii})\) and \(I(I - 1)\) values \(\varepsilon^X_{ij}\). Furthermore, we can approximate the denominator in \(a\) as
\[
\text{ssq}(XX')^{1/2} \approx (J_i^2 + I\sigma^2_x + i(i - 1)\tau^2_x)^{1/2},
\]
which can be explained as follows. The squares of the \(I\) diagonal elements of \(XX'\) sum to \(\sum_i (J_i + \varepsilon^X_{ii})^2 = \sum_i J_i^2 + 2 \sum_i J_i \varepsilon^X_{ii} + \sum_i (\varepsilon^X_{ii})^2\). Now using that \(\sum_i J_i^2 = J_i^2\), \(\sum_i \varepsilon^X_{ii} \approx 0\), and that \(\sum_i (\varepsilon^X_{ii})^2 \approx I\sigma^2_x\), we get \(J_i^2 + I\sigma^2_x + i(i - 1)\tau^2_x\) as approximation of the sum of the squared diagonal values. Furthermore, the sum of squared off-diagonal values is \(\sum_{i \neq j} (\varepsilon^X_{ij}) \approx I(I - 1)\tau^2_x\), which completes the explanation. Analogously, the numerator of \(b\) is the vector with \(I\) values of \((J_2 + \varepsilon^Y_{ii})\) and \(I(I - 1)\) values \(\varepsilon^Y_{ij}\), and the denominator in \(b\) can be approximated as
\[
\text{ssq}(YY')^{1/2} \approx (J_2^2 + I\sigma^2_y + I(I - 1)\tau^2_y)^{1/2}.
\]
Now, the RV-coefficient can be approximated as
\[
RV(X,Y) = a'b =
\]
\[
\approx \frac{(J_1 J_2)}{(J_2^2 + I\sigma^2_y + I(I - 1)\tau^2_y)^{1/2}}
\]
\[
= \frac{(J_1 J_2)}{(J_2^2 + I\sigma^2_y + I(I - 1)\tau^2_y)^{1/2}},
\]
taking into account that in the numerator of \(a'b\) all terms including the random values \(\varepsilon^X_{ii}\) and \(\varepsilon^Y_{ii}\) can be expected to roughly cancel. We can further simplify this expression by using the theoretical values for \(\sigma_x, \tau_x, \sigma_y,\) and \(\tau_y\) to obtain
\[
RV(X,Y) \approx \frac{(J_1 J_2)}{(J_2^2 + 2J_1 + (I - 1)J_1)^{1/2}}
\]
From (16) and (17) it can be seen that the value of RV for random data matrices depends on $I$: for small $I$, the RV is close to 1, whereas, as $I$ increases the denominator increases and the value approaches zero. The accuracy of these approximations depends on $I$ but as has been verified in simulations these approximations are typically quite good. Specifically, approximations by the above approach and mean values for RV over 100 random trials yielded the results as given in Table 1 showing that the average values of RV coefficients are very well approximated by the computation of RV according to (17). Apart from rounding errors, there is a close correspondence between the approximations [i.e. (17)] and the means of the simulated RV-values. These results show again that the RV-value is artificially high for small $I$. Interestingly, the RV-value for the limiting case, i.e. $I = 1$, leads to an RV-value of 1, as is easily verified as follows. In the case of $I = 1$, the matrices $XX'$ and $YY'$ reduce to single numbers, and normalizing these trivially leads to setting these numbers equal to 1, so that the RV-value (i.e. their product) also equals 1.

<table>
<thead>
<tr>
<th>$I$</th>
<th>Derived RV-values</th>
<th>Mean simulated RV-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.852</td>
<td>0.855</td>
</tr>
<tr>
<td>40</td>
<td>0.747</td>
<td>0.748</td>
</tr>
<tr>
<td>60</td>
<td>0.665</td>
<td>0.666</td>
</tr>
<tr>
<td>80</td>
<td>0.599</td>
<td>0.599</td>
</tr>
<tr>
<td>100</td>
<td>0.545</td>
<td>0.545</td>
</tr>
<tr>
<td>200</td>
<td>0.376</td>
<td>0.376</td>
</tr>
<tr>
<td>300</td>
<td>0.287</td>
<td>0.286</td>
</tr>
<tr>
<td>400</td>
<td>0.232</td>
<td>0.232</td>
</tr>
</tbody>
</table>

The modified RV-coefficient

**Definition** A solution to the problem of the RV-coefficient presents itself by considering the nature of the problem: the numerator of (17) does not tend to zero for random numbers and large $J_1$ and $J_2$. This can be traced back to the diagonal of the matrices $XX'$ and $YY'$. Indeed, if these diagonal elements are ignored (or, equivalently, set to zero), then the problem disappears since, e.g. $\text{Vec}(XX')$ would be a vector of values randomly varying around zero. After using again (7) this would result in an RV-coefficient of nearly zero, as should be the case for the two random matrices. This is then also exactly our proposal for $RV_2$, namely instead of using $XX'$ use $[XX' - \text{diag}(XX')] = XX'$, where $\text{diag}(XX')$ is a matrix containing only the diagonal elements of $XX'$ on its diagonal, and zero's elsewhere. Using the analogous definition for $YY'$ we get

$$RV_2(X, Y) = \frac{\text{Vec}(XX')'\text{Vec}(YY')}{\sqrt{\text{Vec}(XX')'\text{Vec}(XX')\times\text{Vec}(YY')'\text{Vec}(YY')}}$$

(18)

Stated otherwise, ignoring the diagonal elements of $XX'$ and $YY'$ gives a new vector $a$ with $I$ values of 0 and $I(I - 1)$ values $\varepsilon_i^T$. This solves the problem because the numerator of (17) when using $RV_2$ then becomes zero.
Properties The $RV_2$ has different properties than the original RV. The most striking one is that $RV_2$ can become negative. Suppose for example that

$$XX' = \begin{pmatrix} 1 & -0.2 \\ -0.2 & 1 \end{pmatrix},$$  \hspace{1cm} (19)

and

$$YY' = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}$$  \hspace{1cm} (20)

then using (18) gives $RV_2(X,Y) = -1$. If instead

$$XX' = \begin{pmatrix} 1 & 0.2 \\ 0 & 1 \end{pmatrix},$$

then $RV_2(X,Y) = 1$. The interpretation of $RV_2 = -1$ is that the association between the rows of $X$ is proportional to the association between the rows of $Y$ but with a negative sign (equivalent to a negative Pearson correlation).

The $RV_2$ depends only on the cross-products $XX'$ and $YY'$, thus the $RV_2$ is also orientation independent. The $RV_2$ has values in-between $-1$ and $1$. This follows immediately from the Cauchy–Schwarz inequality applied to the vectors in (18).

Examples

Simulated examples

Two simulation examples will be used to illustrate the working of the modified RV-coefficient. The first example addresses the (too) large values of the original RV-coefficient. This example follows closely the gene-expression dataset in which the problem was initially encountered. Two datasets $X$ of size $(I \times 130)$ and $Y$ of size $(I \times 113)$ were generated 100 times with standard normal distributed numbers. The number of samples was increased from 20 to 400 with steps of 20. For each simulation run, the modified RV-coefficient was calculated. The results are shown in Figure 3.

Figure 3. Simulation of the modified RV-coefficient ($RV_2$) with different numbers of samples ($J_1=130, J_2=113$). Plotted is the mean, minimum and maximum for 100 repeats each of the $RV_2$ as a function of the number of samples.
The second example shows the working of the RV-coefficients for the case that the amount of overlap between \( X \) and \( Y \) gradually increases. Two matrices \( X \) and \( Y \) were simulated both of size \((10 \times 100)\) with random numbers drawn from a \( N(0,1) \) distribution and this was repeated 100 times. Gradually, columns of are exchanged with those of in steps of 10\%, 20\%,..., Hence, the amount of overlap between and increases. Figure 4 shows the original and modified RV-coefficient. Indeed, the modified RV-coefficient gradually increases whereas the original RV-coefficient already has high values from the start.

Summarizing, Figures 3 and 4 show that the modified-RV coefficient has the desired behavior: (i) on average it equals zero for not related matrices, (ii) for larger sample sizes its variability decreases and (iii) it increases with an increasing amount of overlap.

**Figure 4.** Simulation of the original and modified RV-coefficient with different amounts of overlap. Plotted are the mean, minimum and maximum for 100 repeats RV’s as a function of the amount of overlap.

**Gene-expression example**

The gene-expression example is taken from the paper of Kleeman et al. (2007). In short, gene-expression profiles were measured in livers of female ApoE*3L transgenic mice (E3L mice) from three diet groups (\( n = 5 \) mice per group): control diet (Ct, no cholesterol), low cholesterol diet (Low) and high cholesterol diet (High). Diets were consumed for 10 weeks. RNA from livers was analyzed using Affymetrix whole-genome mouse array MOE430-2.0. Subsets of genes used for matrix correlation were selected based on functional annotation of genes in biological processes cholesterol metabolism (C, \( J = 71 \)) and inflammation (I, \( J = 66 \)), vascular development (V, \( J = 69 \)) and amino-acid metabolism (A, \( J = 91 \)). Interest focussed on comparison within the different treatment groups, therefore, the RV-coefficients were calculated for the VA blocks and the IC blocks of gene-expressions within each treatment group. All gene-expression values were expressed as deviations from the average Ct group ones without further preprocessing.

Table 2 shows that the modified RV-coefficient is always lower that the original RV-coefficient as it should be. The modified RV-coefficients are more reasonable from a biological perspective. All four selected biological processes were enriched in the selection 404 Matrix correlations of genes differentially expressed in response to cholesterol feeding.
From these, the more pronounced response was found on cholesterol metabolism (both in response to low and high cholesterol) and inflammation (high cholesterol) (Kleeman et al., 2007). The dose-dependent gene-expression responses are likely to result in increased correlation between the matrices. Contrary, it is not reasonable to have high correlations between the groups of genes for the control animals (Ct). Hence, the values of the modified RV-coefficient for the control groups are more reasonable than the original RV-coefficients. Also from a statistical point of view, the modified RV-coefficients are better than the original ones. This will be explained for the numbers in the first row of Table 2. Consensus-principal component analysis (CPCA) is an alternative method to probe similarities between matrices (Smilde et al., 2003). Using PCA on the V and A matrices individually gives explained variances of 70.0% and 66.9% for two principal components, respectively. When using CPCA, the two CPCA components explain 57.4% and 64.9% in each block, respectively. The drop in explained variances per matrix (especially for the V block) means that there is some overlap between the matrices but also differences. This agrees nicely with the much lower value of 0.57 instead of 0.84. Note that CPCA is used here only to judge the performance of the modified RV-coefficient. This method does not give an alternative measure to the RV-coefficient but shows qualitatively the same behavior as the modified RV-coefficient supporting the credibility of the latter.

Table 2. The original and modified RV-values of the gene-expression data (for abbreviations: see text)

<table>
<thead>
<tr>
<th>Case</th>
<th>Original RV</th>
<th>Modified RV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ct, VA</td>
<td>0.84</td>
<td>0.57</td>
</tr>
<tr>
<td>Low, VA</td>
<td>0.94</td>
<td>0.87</td>
</tr>
<tr>
<td>High, VA</td>
<td>0.91</td>
<td>0.85</td>
</tr>
<tr>
<td>Ct, IC</td>
<td>0.52</td>
<td>0.27</td>
</tr>
<tr>
<td>Low, IC</td>
<td>0.55</td>
<td>0.18</td>
</tr>
<tr>
<td>High, IC</td>
<td>0.83</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Metabolomics example

The modified RV-coefficient was applied to a metabolomics dataset (Smilde et al., 2005a). Metabolites were measured in Escherichia coli as a model system. The metabolites were measured using two analytical chemical methods, namely gas-chromatographymass spectrometry (GC-MS) and liquid-chromatography-mass spectrometry (LC-MS). This generated two datasets with dimensions 28×12553 (GC-MS) and 28×2532 (LC-MS) which clearly fit into the framework of our modified RV-coefficient. The original RV-coefficient was 0.79 and the modified RV-coefficient was 0.71. Hence, the difference was not large in this case. The CPCA analysis performed in the original publication (Smilde et al., 2005a) showed that both matrices had overlap, but also a substantial nonoverlapping part. Although both types of RV-coefficients did not differ much in this case, the example is shown to illustrate that the modified RV-coefficient gives also a reasonable value in this case.

In Smilde et al. (2005a), a truncation was used prior to calculating the RV-value, i.e. the RV-value was calculated using the first principal components of both matrices. Simulations (results not shown) have pointed out that this approach suffers from the same problems as the RV itself and this approach is therefore not recommended.
Conclusion

It is often convenient to obtain insight into the relationships between blocks of functional genomics data e.g. as a first step in a data fusion strategy. The modified RV-coefficient is a matrix correlation giving such an insight with a single number between −1 and 1. This number can easily be calculated and interpreted in the same way as an ordinary correlation coefficient. The modified RV-coefficient is theoretically motivated and tested with simulations and real data. The results show that this correlation coefficient is reliable and can be used in bioinformatics practice. This coefficient can also easily be combined with permutation testing for assessing significance (Kazi-Aoual et al., 1995) or with bootstrapping to obtain confidence intervals but that is beyond the scope of this article.

References


Kleemann R et al. (2007) Increased dietary cholesterol-induced atherosclerosis is associated with liver inflammation: Identification of novel regulatory pathways and transcriptional regulators involved in switch from metabolic adaptation to inflammatory state. Genome Biology. 8, R200.


van den Berg R et al. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics, 7, 142.

**Appendix**

**Notation**

| (vector) | bold lowercase |
| (matrix) | bold uppercase |
| = 1,..., | object index |
| = 1,..., | variable index |
| = 1,..., | principal component index |

**RV and orientations**

Changing the orientation of the sample configuration of $X$ and $Y$ can be formalized by using arbitrary orthogonal matrices $Q_1$ and $Q_2$ to rotate $X$ and $Y$, respectively. Upon defining $\bar{X} = XQ_1$ and $\bar{Y} = YQ_2$ and observing that the RV-coefficient can be written to depend only on products $XX'$, (see (7)) it holds that $XQ_1(XQ_1)' = XQ_1Q_2'X' = XX'$ due to the orthogonality property of $Q_1$ (and similarly for $Y$). Hence, $RV(\bar{X}, \bar{Y}) = RV(X, Y)$. The $l \times l$ matrices $XX'$ are called configuration matrices and describe the configuration of the $l$ points (rows of $X$) in their respective row-spaces. The RV-coefficient only measures differences in configurations ($XX'$) and not orientations ($Q_1$).

**RV in Vec notation**

The equality that $tr(A'B) = vec(A)'vec(B)$ can easily been proven by writing both $tr(A'B)$ and $vec(A)'vec(B)$ in terms of the elements of the respective matrices. Using this equality in (3) gives (7).
ANALYZING LONGITUDINAL MICROBIAL METABOLOMICS DATA

Carina M Rubingh, Sabina Bijlsma, Renger H Jellema, Karin M Overkamp, Mariët J. van der Werf, Age K Smilde

TNO Quality of Life, P.O. Box 360, 3700 AJ Zeist, The Netherlands

1Swammerdam Institute for Life Sciences, University of Amsterdam, Nw Achtergracht 166, 1018 WV Amsterdam, The Netherlands

Abstract

A longitudinal experimental design in combination with metabolomics and multi-way data analysis is a powerful approach in the identification of metabolites whose correlation with bio-product formation shows a shift in time. In this paper, a strategy is presented for the analysis of longitudinal microbial metabolomics data, which was performed in order to identify metabolites that are likely inducers of phenylalanine production by *Escherichia coli*. The variation in phenylalanine production as a function of differences in metabolism induced by the different environmental conditions in time was described by a validated multi-way statistical model. Notably, most of the metabolites showing the strongest relations with phenylalanine production seemed to hardly change in time. Apparently, potential bottlenecks in phenylalanine seem to hardly change in the course of a batch fermentation. The approach described in this study is not limited to longitudinal microbial studies but can also be applied to other (biological) studies in which similar longitudinal data need to be analyzed.
Introduction

For centuries micro-organisms are being used for the production of a large number of metabolites and proteins such as bio-ethanol, antibiotics, and detergent enzymes. The onset of a specific microbiological production process is generally regulated by a chain of biomolecular events. The trigger for the onset is a signal, usually a specific metabolite, that initiates via a signaling cascade the expression of the genes involved in the biosynthesis of a specific product (Koetje et al., 2003; Lengeler, 2000; Hellingwerf et al., 1995). At large, a change in the concentration of the signaling molecule concentration is observed minutes-to-hours prior to detectable product formation (Figure 1).

![Figure 1. The trigger for the onset a specific microbiological production process is a signal, usually a specific metabolite, that initiates via a signaling cascade the expression of the genes involved in the biosynthesis of a specific product.](image)

These signaling molecules form, therefore, a vital link in microbial product formation, and knowledge of these molecules can be exploited for further strain and process improvement. However, up to date, very few signaling molecules have been described. In conventional biology, the identification of such effectors would involve the one-by-one testing of large collections of metabolites on enzyme activity and/or gene expression. The comprehensive metabolomics technology recently developed (van der Werf et al., 2007) in combination with longitudinal multivariate statistical approaches makes it possible to identify these highly important signaling molecules in a comprehensive and unbiased manner. In this way, the production-related phenotype (i.e., productivity) is not only related to the (changes in the) metabolome composition at the time of sampling (‘end-stage-approach’), but also to changes in the metabolite composition at prior time points. To demonstrate this new approach, a longitudinal microbial metabolomics study with a phenylalanine-producing Escherichia coli was performed. The obtained data contained many challenges from a biological and analytical point of view, but also from a statistical point of view.
Techniques such as LC-MS and GC-MS (Koek et al., 2006; Coulier et al., 2006) can be used to detect which metabolites in microbial samples are related to metabolic changes. A property of these techniques is that they generate a large amount of data consisting of a large number of correlated variables. This correlation can be twofold to, for instance, analytical correlations (i.e., in case of peak picking (Tikunov et al., 2005) or the derivatization method that is used (Koek et al., 2006), which result in the fact than one metabolite might be represented in the data set by more than one peak/variable), or biological correlations (i.e., metabolites that correlate mutually, for instance, a substrate and product of an enzymatic reaction operating close to equilibrium, or metabolites that give a similar response to a perturbation). As a consequence, the data set contains variables that are not completely independent of each other, which requires appropriate statistical tools for analysis.

The fact that the measurements were taken at several time points made it even more complex: not only the variables are correlated, but there is also a mutual dependency over time. It can be expected that samples which are taken from one fermentation are more related to each other than those taken from another fermentation. Besides that, it could also be that consecutive time points are more related than, for instance, the first measurement and the fifth. This correlation between samples violated the assumption of random sampling which often is required in standard statistical data analysis tools.

Many data analysis techniques, like Dynamic Factor Analysis (DFA; Molenaar, 1985) and Dynamic Principal Component Analysis (DPCA; Ku et al., 1995) are developed to deal with this type of dynamic data. Unfortunately, most of the tools that are addressed in those papers are not appropriate for this particular study, because they assume two-dimensional data (e.g., fermentations and time or variables and time) instead of the three-dimensional data of the present study (fermentations, variables and time).

The analytical techniques that were used as well as the longitudinal aspect of the data implied correlated variables and time points, hence appropriate statistical tools and a well-considered analysis strategy were required. In the present paper, a strategy is given for the analysis of longitudinal microbial metabolomic data. Ideas from other disciplines were carefully selected and used to deal with the complexity of the data. Eventually, the data analysis tools Parallel Factor Analysis (PARAFAC; Smilde et al., 2004; Bro, 1997; Harshman and Lundy, 1994) and multiway Partial Least Squares (nPLS; Smilde, 1997; Bro, 1996), originating from psychometrics and chemometrics, respectively, were applied in order to identify changes in metabolites over time that are linked to phenylalanine production.
**Experimental session**

**Experimental Design and Data Collection**

Escherichia coli NST 74 (ATCC 31884) was obtained from the ATCC (Manassas, VA, USA). Cultures were grown in batch at 30°C in a Bioflo 3000 (New Brunswick Scientific) bioreactor containing 5 L of MMT12E medium with 30 g/L glucose or the equivalent amount on C-mol basis of Na2-succinate·6H2O (45 g/L), as the carbon source. The culture was inoculated with 2 % (v/v) of a preculture grown for 8h on LB medium (200 rpm; 30°C). A constant pH was maintained by automatic titration with 4 M KOH and 6 M HCl. The oxygen tension was maintained at 30% by automatic increase of the stirring speed in the fermentor. Fermentations were performed independently in replicate according to a full experimental design (total 8 different conditions, 16 biological duplicates), varying the C-source (glucose or succinate), the phosphate concentration (high (39.6 mM) or low (4.4 mM)), and the pH (6 or 7).

Samples were taken from the bioreactor every hour for biomass (OD600) and phenylalanine (LC-MS analysis; Smilde et al., 2005) determination. Approximately 12 samples (8 in the logarithmic growth phase and 4 in the stationary growth phase, first sample at an OD600 of ≈ 4) were taken equidistantly from each fermentation (i.e., every 1, 2 or 3h, depending on the growth rate of E. coli under the specific environmental condition of the fermentation) and taken automatically from the bioreactor and immediately quenched at -45°C (Pieterse et al., 2006) using a 221 XL Liquid Handler (Gilson, Middleton, WI, USA). The residence time of the culture fluid in the sample tubing of samples collected in this way was <2.5s. In Table 1, more details concerning the sampling are given.

The intracellular metabolites were extracted from the samples by chloroform extraction (Ruijter and Visser, 1996). At appropriate time points, different quality and internal standards were added. After extraction, the sample was divided in two portions. The LC-MS sample was deproteinized using a Centristart I filter from Sartorius. Subsequently, both the GC- and LC-MS samples were lyophilized. The samples were analyzed in duplicate using the OS-GC-MS method (Koek et al., 2006) and the IP-LC-MS method (Coulier et al., 2006), respectively, which together analyze over 95% of microbial metabolites (Van der Werf et al., 2007). GC-MS-data were preprocessed using TNO-DECO (homemade software, based on multivariate curve resolution alternating least-squares (MCR-ALS) algorithms; Esteban et al., 2000; Tauler et al., 1993), which was used for peak deconvolution, whereas IMPRESS/E-QUEST/WINLIN (home-made software; van de Greef et al., 2004; Vogels et al., 1996) were used to align and peak-pick the LC-MS data. The GC data set contained 580 metabolites and the LC data set contained 633 peaks.
Data Preprocessing

Several additional data preprocessing steps were performed in order to rule out differences between samples which might be due to factors other than the ones enforced by the design. LC and GC data of each fermentation sample were corrected for the recovery of the Internal Standard (IS) for injection and were adjusted for biomass to normalize for the total amount of E. coli biomass from which these samples were obtained. The amount of zeros in the data set, which could be due to a measurement below the detection limit, peak picking error or deconvolution error, was reduced by combining the duplicate measurements. If both analytical duplicates had a zero value or if both had a non-zero value, measurements were averaged, whereas the single value was taken if only one of the duplicates was above zero (Bijlsma et al., 2006).

Table 1. Description of fermentations: design and sampling

<table>
<thead>
<tr>
<th>Fermentation</th>
<th>Carbon Source</th>
<th>Phosphate concentration</th>
<th>pH</th>
<th>OD/Phe [start:by:end]</th>
<th>Quench [start:by:end]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Glucose</td>
<td>High</td>
<td>6</td>
<td>[0:1:44]</td>
<td>[18:2:42]</td>
</tr>
<tr>
<td>2</td>
<td>Glucose</td>
<td>High</td>
<td>6</td>
<td>[0:1:41]</td>
<td>[17:2:39]</td>
</tr>
<tr>
<td>3</td>
<td>Glucose</td>
<td>Low</td>
<td>6</td>
<td>[0:1:48]</td>
<td>[14:2:36]</td>
</tr>
<tr>
<td>4</td>
<td>Glucose</td>
<td>Low</td>
<td>6</td>
<td>[0:1:47]</td>
<td>[14:2:36]</td>
</tr>
<tr>
<td>5</td>
<td>Glucose</td>
<td>High</td>
<td>7</td>
<td>[0:1:25]</td>
<td>[10:1:20]</td>
</tr>
<tr>
<td>6</td>
<td>Glucose</td>
<td>High</td>
<td>7</td>
<td>[0:1:25]</td>
<td>[10:1:20]</td>
</tr>
<tr>
<td>7</td>
<td>Glucose</td>
<td>Low</td>
<td>7</td>
<td>[0:1:31]</td>
<td>[9:2:31]</td>
</tr>
<tr>
<td>8</td>
<td>Glucose</td>
<td>Low</td>
<td>7</td>
<td>[0:1:31]</td>
<td>[9:2:31]</td>
</tr>
<tr>
<td>9</td>
<td>Succinate</td>
<td>High</td>
<td>6</td>
<td>[0:1:5:66; 66:1:93]</td>
<td>[71:2:93]</td>
</tr>
<tr>
<td>10</td>
<td>Succinate</td>
<td>High</td>
<td>6</td>
<td>[0:1:5:66; 66:1:93]</td>
<td>[68:2:96]</td>
</tr>
<tr>
<td>11</td>
<td>Succinate</td>
<td>Low</td>
<td>6</td>
<td>[0:1:79]</td>
<td>[42:3:75]</td>
</tr>
<tr>
<td>12</td>
<td>Succinate</td>
<td>Low</td>
<td>6</td>
<td>[0:1:73]</td>
<td>[40:3:73]</td>
</tr>
<tr>
<td>13</td>
<td>Succinate</td>
<td>High</td>
<td>7</td>
<td>[0:1:59]</td>
<td>[31:2:53]</td>
</tr>
<tr>
<td>14</td>
<td>Succinate</td>
<td>High</td>
<td>7</td>
<td>[0:1:59]</td>
<td>[32:2:54]</td>
</tr>
<tr>
<td>15</td>
<td>Succinate</td>
<td>Low</td>
<td>7</td>
<td>[0:1:45]</td>
<td>[23:2:45]</td>
</tr>
<tr>
<td>16</td>
<td>Succinate</td>
<td>Low</td>
<td>7</td>
<td>[0:1:48]</td>
<td>[26:2:48]</td>
</tr>
</tbody>
</table>

*Sampling started at 0 hrs, with a sampling frequency of 1 hr and ended after 44 hrs*

Some of the LC and GC peaks contained a substantial number of zeros, which made these peaks noninformative and which could disturb the statistical analysis. The noninformative peaks were removed by proceeding with the ‘xx%-rule’ (Bijlsma et al., 2006): 1) consider all measurements of one fermentation to belong to one group; 2) count the number of non-zero values per peak within each group; 3) calculate the percentage of non-zero values per peak for each group (xx%); and 4) retain only those peaks which have a certain percentage of non-zero values for at least one of the groups. For the LC data set, a 10%-rule was applied, which meant that only those LC-peaks which were present in at least 10% of the samples from one of the fermentations, remained in the data set. After applying the 10%-rule, 86 peaks remained. Applying this rule using a higher percentage resulted into a data set with
too few variables. The zero-values that were left, were replaced by a threshold value of half times the minimal value in the complete data set unequal to zero.

Since the number of zeros was much lower in the GC data set compared to the LC data set, the GC data was cleaned up by applying an 80%-rule, which meant that only those GC-peaks which were present in at least 80% of the samples that were taken for one of the fermentations, were remained in the data set. Since applying this rule might imply the removal of inducing variables which are only present at the first few time points, an additional inclusion criteria was defined. Also variables for which a value was present for at least one experimental condition in both biological duplicates for the first until the third or for the second until the fourth measurement were included. After applying the adjusted 80%-rule, 431 peaks were left in the GC data set, of which eight were included by applying the additional inclusion criterion. The zero-values that remained in the GC data were not replaced by a threshold value, since this value was almost equal to zero, which makes the replacement of zeros not necessary for this platform.

As a final preprocessing step, succinate and glucose peaks were removed from the GC data set, because they were highly overloaded and noninformative peaks which reflected the experimental design. They could dominate and disturb the data analysis and may lead to trivial solutions. The same was the case for phenylalanine, the product formed in this microbial fermentation, that was present in both GC and LC data sets, hence they were removed as well. This resulted in a GC data set that contained 411 peaks in total and an LC data set of 83 peaks. By combining the LC- and the GC-data for data-analysis, which led to a fused data set of 494 variables, the information on the metabolome of the E. coli was covered comprehensively.

Mathematical Session

Two-way Analysis versus Three-way Analysis

Basic multivariate data analysis tools like Principal Component Analysis (PCA; Vandeginste et al., 1998; Massart et al., 1997; Martens and Naes, 1989; Joliffe, 1986; Dillon and Goldstein, 1984) and Partial Least Squares (PLS; Martens and Naes, 1989; Geladi and Kowalski, 1986) could be used to analyze the fused data set of size $I \times JK$, where $I =$ fermentation, $J =$ metabolite and $K =$ time. However, these so-called two-way analysis techniques have the disadvantage that the information on time $K$ and on the variables $J$ is intertwined.

To overcome this deficit, three-way data analysis tools were used, which makes it possible to separate the time factor from the metabolites, resulting in clear, simple and better interpretable models. The two-way data set was translated into a 3-way matrix of size $I \times J \times K$, where $I =$ fermentation (1st mode), $J =$ metabolite (2nd mode) and $K =$ time (3rd mode). This matrix reflects the three-way structure of the data. PARAFAC (Bro, 1998; Bro, 1997; Harshman and Lundy, 1994) and nPLS (Smilde, 1997; Bro, 1996) were used to analyze the data.
PARAFAC

PARAFAC is the multi-way generalization of two-way PCA, a well-established statistical method for the multivariate interpretation of complex data. It decomposes the data into a few factors, which are sets of scores and loadings that describe the information in a more compact way than the original data set (Smilde et al., 2004; Bro, 1997; Harshman and Lundy, 1994). In this particular case, the data was decomposed into a matrix of scores that represented the fermentations (size $I \times C$, where $C =$ number of components), a matrix of loadings that described the peak information (size $J \times C$), and a loadings-matrix of size $K \times C$ that contained the information about the relative time. Statistical diagnostics were used to check the stability and reliability of the PARAFAC models (Bro, 1997). The scores and loadings were evaluated graphically and were used in an explorative way in order to detect outliers or any other unexpected pattern in the data set.

nPLS

Multiway-PLS or nPLS is the multi-way generalization of the two-way regression method PLS in which a regression model is build between dependent variables (Y-block) and independent variables (X-block), of which at least one of them is multiway (Smilde, 1997; Bro, 1996). The two blocks have one mode in common, which is the sample mode $I$. nPLS gives a regression matrix that can be used for the interpretation of the model. In the present study, nPLS was used to identify the metabolites that induce phenylalanine production.

Model Validation

The PARAFAC results were validated using the jackknife procedure, in order to identify the standard errors of the model parameters and to detect outliers in the data set. The impact of each fermentation on the PARAFAC-model was assessed by leaving out the data of one fermentation (size $1 \times J \times K$) and by building a PARAFAC model without this fermentation. This was repeated until all fermentations were left out once. The results of the jackknife procedure were evaluated graphically (Riu and Bro, 2002).

Two types of cross-validation were used to determine the optimal number of latent variables (LVs) and to validate the nPLS model: “leave-one-fermentation-out” cross-validation (LOFO-CV) and “leave-one-design-point-out” cross-validation (LODO-CV). Using LOFO-CV, data of one fermentation (size $1 \times J \times K$) was left out in the first cross-validation step, an nPLS model was built, and the Y-variable of the fermentation that was left out was predicted. This was repeated until all 16 fermentations were left out once. The optimal number of LVs was determined based on the minimum value of the root-mean-square error of cross-validation (RMSECV).

Using LODO-CV, data of two fermentations with the same design condition (size $2 \times J \times K$) were left out in the first cross-validation step, an nPLS model was built, and the Y-variable of the fermentations that were left out were predicted. This was repeated until all 8 design points were left out once. The optimal number of LVs was determined based on the
minimum value of the RMSECV. LODO-CV implies that extrapolation might be needed to predict the two fermentations that were left out, hence this way of CV might be too strict for modeling.

The predictivity of the nPLS-model was determined by the R-square of cross-validation ($R^2_{cv}$), which described the percentage of variation of $Y$ that is explained by the cross-validated predictions of $Y$. The closer the $R^2_{cv}$ to 1, the more of the variation of the maximal specific productivity is explained by the prediction, the better the model.

**Centering and Scaling**

Scaling and centering in two-way analysis is quite straightforward. However, in multi-way analysis it is more complicated, because centering across or scaling within a certain mode might disturb prior centering and/or scaling steps (Smilde et al. 2004; Kiers and Van Mechelen 2001; Harshman and Lundy 1984).

In the present study, the fused data set was centered across the fermentation mode $I$ and followed by range-scaling within the metabolite-mode $J$. The centering step was performed to remove constants between the fermentations, whereas range-scaling within the metabolite mode resulted in metabolite concentrations that were relative to the biological range of that metabolite (Van den Berg et al., 2006). By performing the scaling step after the centering step, the prior centering remained unaffected. The centering and scaling was performed within each step of the cross-validation procedure.

All analyses were performed using Matlab Version 7.0.4 R14 (The Mathworks, Inc.) and the n-way toolbox version 2.11 (Anderson and Bro, 2000).

**Results and discussion**

**Study-specific Data Preprocessing**

Because of the variation in analyses and sample workup, the LC and GC peaks and the phenylalanine-production curves showed a spiky pattern over time. To resolve this problem, all peaks and phenylalanine curves were smoothed using a smoothing algorithm that is based on penalized least-squares (Eilers, 2003). The degree of smoothing depends on the value of the penalty (lambda) and the derivative that is used. It was found that smoothing using the second derivative and a penalty of 5 gave the best results for most of the metabolites, both GC and LC.

An additional complexity was the fact that there were differences in run time between fermentations due to differences in growth rate of *E. coli* under the different environmental conditions. When the profiles are aligned, a parsimonious and simple model will be allowed. The differences in run time, varying between 25 and 100h, needed to be removed. In the analysis of batch processes, several approaches like dynamic time warping (Westerhuis et al., 1999; Kassidas et al., 1998) or the use of an indicator variable (Neogi and Schlags, 1998; Kourt et al., 1996) are developed to deal with irregularities such as varying time intervals.
and differences in batch lengths in the preprocessing phase, whereas others solve the problem during the modeling phases (Wold et al., 1998). Since these methods were not fully appropriate for the present data, another approach was chosen.

Differences in run time between the fermentations due to differences in growth rate under the different environmental conditions were removed by using relative time points instead of the original time. However, it was not possible to translate the original time points directly into a relative time by defining the first measurement as 0% of elapsed time and the last measurement as 100% of the elapsed time, since these time points did not represent the same biological stage of growth between fermentations. Some of the fermentations had a long lag phase, whereas others showed an immediate start of the growth after inoculation. Therefore, a new start and end point of elapsed fermentation time were defined. The time point at which 20% of the maximal biomass was reached was considered to be biologically comparable between fermentations and was defined as the beginning of the curve (0% of elapsed time). The fourth sample that was quenched in the stationary phase was set to be the end of the time curve (100% of elapsed time). All time points in between the new starting and end point were translated into relative time by linear interpolation and all time points that were outside this region were discarded. The metabolite-data were interpolated using the new relative time points and the data at the percentiles of the relative time were used in the data-analysis.

Summarizing, all preprocessing steps led to smooth equidistant time curves of equal length for the GC data set, the LC data set and the phenylalanine data set.

**PARAFAC**

A two-component PARAFAC-model was built, which explained 25% of the total variance in the fused data set. The model was validated using the jackknife procedure and there were no major model violations and no strongly deviating points (results not shown). The original design can be recognized in the score plot of the fermentations (Figure 2). There was a clear separation between the samples that were grown on glucose and the samples which had succinate as carbon source (Figure 2(a)). Although not as clear as the separation between carbon sources, also a division into high and low phosphate was observed, which was most clear for the succinate group (Figure 2(b)). Also a minor division into the two pH groups within the glucose and within the succinate group was found (Figure 2(c)).

The loading plot of the metabolites can be used to identify the variables, which contributed to the variance that is seen in the fermentation score plot. The metabolites with a high positive loading for the first factor were mainly present in the succinate fermentations and the metabolites with a high negative loading for the first factor were mainly present in the fermentations that were grown on glucose. Since the identification of metabolites that contribute to the separation between the glucose and succinate fermentations was not of major interest in this study, this loading plot is not shown.
In Figure 2(d), the loading plot of the third mode, the relative time points, is given. One thing attracts the attention immediately, namely, the shape of the loadings. Successive time points are placed on a smooth curve, which it is not that surprising in this particular case. It can be expected that adjacent time points are more related to each other than to the time points further down. The plot of the loadings also reveals that there were two separate processes over time, indicating distinct phases in the fermentation process (I and II in Figure 2(d)). The first PARAFAC-factor describes mainly the information in the last part of the time curve (p50 - p100), which corresponds to the stationary phase, whereas the second factor is more concerned with the first part of the time curve (p0 - p50), which describes the logarithmic phase.

**nPLS**

The maximal specific productivity was calculated as the maximum slope of the smoothed phenylalanine-production curve divided by the biomass at the time point having the maximal slope. This resulted in one single value per fermentation that was most relevant for describing the phenylalanine-production curve and which was used as Y-block (size 1 x 1) for nPLS modeling.

**Statistical Interpretation** Because of the wide range of maximal specific productivity, nPLS models were built using the maximal specific productivity as Y, as well as the log-
transformed (natural log or ln-transformation) maximal specific productivity. These two nPLS models were analyzed in detail using two types of cross-validation. The fused data set of LC and GC data was defined as X-block. The results of the nPLS-models are summarized in Table 2. As expected beforehand, the models based on LOFO-CV performed better than the ones based on LODO-CV. A LOFO cross-validation might be considered to be optimistically biased because a fermentation with the same design condition was still present in the data set. In case of analytical duplicates, this would be unwanted. However, the duplicates that were present in the data set were biological instead of analytical duplicates. Since the biological variation is much larger than the analytical variation for this type of data (Van der Berg et al., 2006), this approach was justified. Leaving out both biological duplicates per cross-validation step, as was done using LODO-CV, meant that one design-condition was left out completely. It was expected that the predictivity of these models was worse compared to the models based on LOFO-CV, because the two samples of the design point that were left out were predicted by extrapolation.

Table 2. Summary of nPLS results (CV = cross-validation; LOFO = leave-one-fermentation-out, LODO = leave-one-design point-out; #LVs = number of latent variables; Ln = natural logarithm).

<table>
<thead>
<tr>
<th>Model</th>
<th>Y</th>
<th>CV</th>
<th># LVs</th>
<th>% Explained Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
<td>Y</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a</td>
<td>Y</td>
<td>LOFO</td>
<td>3</td>
<td>27</td>
</tr>
<tr>
<td>1b</td>
<td>Y</td>
<td>LODO</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>2a</td>
<td>Ln(Y)</td>
<td>LOFO</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>2b</td>
<td>Ln(Y)</td>
<td>LODO</td>
<td>2</td>
<td>22</td>
</tr>
</tbody>
</table>

Although the $R^2_{cv}$ of the model with LOFO-CV and using the original scale of the maximal specific productivity was higher compared to the model using the log(Y), the nPLS model using the log(Y) as Y-block was statistically a better model compared to the model using the maximal specific productivity on the original scale. Although the $R^2_{cv}$ can be used as a model-validation statistic, it is not an absolute criterion, since it is sensitive to influential samples, which are easily identified in a scatter plot between the predicted and the measured values. The two highest values of the maximum specific productivity affected the $R^2_{cv}$ positively by their location compared to the other levels. Therefore, the model based on log(Y) using LOFO-CV will be used to explain the interpretation of the nPLS-models.

Three plots, which are presented in Figure 3, were used to evaluate and interpret the nPLS-model in full detail. For all models, similar figures were made and evaluated. The RMSECV was used to determine the number of LVs and is shown in Figure 3(a). The minimal RMSECV was found for two LVs, hence the number of LVs was determined to be two. The model explained 22% of the variance in the X-block and 93% of the variance in the Y-block. The prediction of the maximal specific productivity based on the cross-validation versus the measured value of log(Y) is plotted Figure 3(b) and was used to calculate the $R^2_{cv}$. For this particular model, the $R^2_{cv}$ is equal to 0.59, which means that 59% of the original variance of log(Y) is explained by the prediction in the cross-validation.
Finally, a plot was made of the loadings of the relative time points (Figure 3(c)). Like in the PARAFAC results, two separate processes over time can be identified. The loadings of the first LV change over the range of p0 - p60, whereas there is only a little change in the loadings of the second LV for the range of p60 - p100. This is in line with what is seen in the growth curve of the fermentations, which has a logarithmic growth phase and a stationary growth phase. Since induction precedes production, also an nPLS model was built using only the p0 - p60 data (results not shown). As expected, the results of this model were similar to the ones using the full range of time, indicating that the main changes in metabolites that are related to the phenylalanine productivity did take place in the logarithmic growth phase.

**Biological Interpretation of the Results** To identify the metabolites that correlated with the maximal specific productivity, the regression matrix (size 494 x 11), with a regression coefficient for each variable at each time point, was assessed. The absolute values of the regression vector were sorted per time point in a descending order and the top 20 of metabolites, being the 20 variables with the highest absolute regression coefficient for that particular time point, was determined. The ranking of metabolites for model 1a is presented in Table 3 and the annotation of the metabolites is given in Table 4. The higher the rank of a
metabolite, the more it contributes to relation with the maximal specific productivity. Each variable that had entered the top 20 for at least one of the time points was used for identification and biological interpretation.

Table 3. Changes in ranking of metabolites in the nPLS model over time.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Elapsed time (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

Assessing the ranking of the metabolites for all time points at ones gives information on their contribution over time, as is illustrated in Table 3. Strikingly, the rank of the metabolites seem to hardly change in time for most metabolites. For instance, G01 (3,5-dihydroxypentoanoic acid, tentatively identified) shows high correlation with the maximal specific productivity for each time point. Of the 20 metabolites showing the highest contribution to the phenylalanine production, 13 metabolites changed less than 3 places in rank in time, while 5 metabolites changed 4-7 places in rank. Only two metabolites that appeared in the top 20, that is, G12 (unknown) and G13 (aspartic acid), showed a much larger drop in rank in time, that is, 18 and 22 places, respectively. Although several of the highest ranking metabolites were previously identified in an independent end-stage study with the same organism15, of only one of them, i.e. tyrosine (G05 in Table 3), a direct link with phenylalanine production is obvious.
Table 4. Annotation of metabolites.

<table>
<thead>
<tr>
<th>Code</th>
<th>Annotation</th>
<th>Code</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>G01</td>
<td>3,5-dihydroxypentanoic acid (putative)</td>
<td>G13</td>
<td>aspartic acid</td>
</tr>
<tr>
<td>G02</td>
<td>Urea</td>
<td>G14</td>
<td>Unknown</td>
</tr>
<tr>
<td>G03</td>
<td>Unknown</td>
<td>G15</td>
<td>arabinose (or isomer)</td>
</tr>
<tr>
<td>G04</td>
<td>Dissaccharide</td>
<td>G16</td>
<td>Putrescine</td>
</tr>
<tr>
<td>G05</td>
<td>Tyrosine</td>
<td>L01</td>
<td>Unknown</td>
</tr>
<tr>
<td>G06</td>
<td>Unknown</td>
<td>L02</td>
<td>unknown</td>
</tr>
<tr>
<td>G07</td>
<td>Unknown</td>
<td>L03</td>
<td>unknown</td>
</tr>
<tr>
<td>G08</td>
<td>Ribose (or isomer)</td>
<td>L04</td>
<td>UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminoheptanedioate-D-alanyl-D-alanine</td>
</tr>
<tr>
<td>G09</td>
<td>Dissaccharide</td>
<td>L05</td>
<td>AMP</td>
</tr>
<tr>
<td>G10</td>
<td>Unknown</td>
<td>L06</td>
<td>Unknown</td>
</tr>
<tr>
<td>G11</td>
<td>Dissaccharide</td>
<td>L07</td>
<td>Unknown</td>
</tr>
<tr>
<td>G12</td>
<td>Unknown</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Visualization** Statistical relevance does not automatically imply biological relevance. Therefore, the regression vectors of all models were used for the biological identification and interpretation, hence by using Y-information and log(Y)-information linear as well as non-linear correlations could be covered.

In the top of metabolites that were important for the nPLS models, both GC and LC peaks appeared. A metabolite that appeared in the top of all the models was tentatively identified as 3,5-dihydroxypentanoic acid (G01 in Table 3). This metabolite and the result of model 2a will be used for visualization purposes.

The interpretation of a regression vector based on a two-way analysis technique can be complicated. A regression coefficient is calculated for each metabolite, reflecting the relation between the Y-variable and a certain metabolite, given the other metabolites. The interpretation of a regression vector that is obtained in a three-way analysis will be even more complicated, because it is not a single vector that needs to be interpreted. The information about the relation between the phenylalanine-production and the metabolism was described by 11 regression vectors, one for each relative time point. In Figure 4, the course of the regression coefficient of 3,5-dihydroxypentanoic acid (tentatively identified) is presented for each relative time point. It shows an increasing positive correlation between the maximal specific phenylalanine productivity and the metabolite concentration over time (all coefficients are positive) and it also shows that this correlation remains the same in the second half of the curve (the coefficients remain more or less the same after 50% of the time was elapsed).
Regression coefficients of 3,5-dihydroxypentanoic acid based on the nPLS-model based on log(maximal specific phenylalanine productivity) and the fused metabolite data, using leave-one-fermentation-out cross-validation.

To illustrate the correlations that were found, in Figure 5, a scatter plot is made of the maximal specific productivity versus the smoothed and aligned concentration of intracellular 3,5-dihydroxypentanoic acid (tentatively identified) per relative time point. The least-squares regression line is also given. All regression lines have positive slopes, indicating that there is a positive relation between the two variables, which was also seen in the plot of the regression coefficients (Figure 4). At the first time point, there is no correlation; a random cloud of data points is seen. The slope of the regression line changes in the first five plots (p0 until p40) and the correlation increases. After p50, they remain similar, which indicates that the relation between the maximal specific phenylalanine productivity and 3,5-dihydroxypentanoic acid (tentatively identified) does not really change after 50% of the time has elapsed. This is the same conclusion as was drawn from the regression coefficients in Figure 4. The random cloud of data points corresponds to the lowest regression coefficient and the more time has been elapsed, the higher the correlation, the higher the regression coefficient. It must be kept in mind that the regression coefficients in Figure 4 are based on a model in which all 494 metabolites were included. So, the coefficients reflects the relation between the maximal specific productivity and 3,5-dihydroxypentanoic acid (tentatively identified), given the presence of the other metabolites. In Figure 5, these other metabolites are not taken into account.
CONCLUSION

A longitudinal experimental design in combination with metabolomics and multi-way data analysis is a powerful approach in the identification of metabolites whose correlation with bio-product formation shows a shift in time. With the use of the multi-way multivariate statistical tools, the time dimension is uncoupled from the variable dimension, allowing the identification of such regulatory metabolites. However, this conceptually new approach is very demanding and many biological, analytical, and statistical challenges had to be overcome. For instance, several fermentations were performed according to a careful set up full factorial experimental design using eight different environmental conditions, and more than 10 samples per fermentation were taken in time. The analysis of the metabolome data that were obtained in order to investigate the metabolites that are involved in phenylalanine production asked for appropriate statistical tools in order to account for analytical variation, sample workup, differences in run-time, correlated variables and mutual dependency over time. A statistical protocol consisting of several carefully selected and validated steps was set up and used to analyze the longitudinal metabolomics data. The strategy that was followed is an example of how to tackle these kinds of problems. The approach described in
this study is not limited to longitudinal microbial studies but can also be applied to other (biological) studies in which similar longitudinal issues need to be addressed.

Notwithstanding all biological, analytical, and statistical challenges in the generation and analysis of this complex data set, valid models were obtained and several of the metabolites that were identified to show the strongest relation with phenylalanine production seemed also biologically relevant. With this approach, for the first time, insight was generated in the changes in time of potential bottlenecks in the course of a batch fermentation process. Surprisingly, with the phenylalanine production process described in this manuscript, hardly any metabolite showed strong changes in rank in time. Although the absolute contribution to the maximal specific phenylalanine productivity of the metabolite did change over time, the internal differences between metabolites was not affected by time. Apparently, potential bottlenecks in phenylalanine seem to hardly change in the course of a batch fermentation.

Acknowledgements

Bas Muilwijk and Leon Coulier are thanked for their contribution and helpful comments on pre-processing, analysis and the identification of the metabolites.

References


DISCOVERY OF SUBTLE EFFECTS IN A HUMAN INTERVENTION TRIAL THROUGH MULTILEVEL MODELING

Carina M Rubingh, Marjan J van Erk, Suzan Wopereis, Trinette van Vliet

Elwin R Verheij, Nicole HP Cnubben, Ben van Ommen, Jan van der Greef

Henk FJ Hendriks, Age K Smilde

TNO Quality of Life, Business Unit Analytical Sciences, P.O. Box 360, 3700 AJ Zeist, The Netherlands

1Currently at: CCMO, the Hague, The Netherlands

2Biosystems Data Analysis, SILS, University of Amsterdam, Amsterdam, The Netherlands

Accepted for publication in Chemometrics and Intelligent Laboratory Systems
Abstract

Many benefits can be gained if multi-factorial diseases with a high incidence and prevalence are better understood. Sophisticated approaches like multilevel analyses are needed to discover subtle differences between healthy people and people at the onset of disease in these types of studies. Multilevel analysis generates different sub-models for each level of variation. For instance, within and between subject variation can be split and analyzed separately if the two factors are orthogonal (i.e., not confounded). In the present paper, the benefits of a multilevel approach in multi-way analysis (nPLS-DA) will be described for the analysis of metabolomics data of a double blinded, randomized, parallel intervention trial with twenty slightly overweight men, whom received a diclofenac or placebo treatment for nine days. Blood samples were taken on multiple time points on 5 treatment days.

The cross-validated error rate for classifying subjects in the correct treatment group for the multilevel nPLS-DA was compared with the error rate from the ordinary nPLS-DA. 42.1% of the subjects were misclassified using ordinary nPLS-DA, whereas only 5% were misclassified using the multilevel approach. Metabolites which contributed in different ways to treatment group differences could be determined and used for biological interpretation.

The multilevel multi-way technique turned out to be a much stronger tool for modeling differences between treatment groups than the ordinary method. The metabolites that contributed most to treatment differences were not only statistically, but also biologically relevant. The multilevel approach found the effects that were better interpretable, whereas the ordinary nPLS-DA failed to do so. The methodology that was described in this paper is not only limited to human intervention studies, but can be used also for studies with a similar data structure. The multilevel approach is able to investigate effects on all levels of variation of every well designed study, hence improving the interpretability of the results.

Abbreviations

CV         Cross Validation
GC-MS      Gas Chromatography – Mass Spectroscopy
IS         Internal Standard
LC-MS      Liquid Chromatography – Mass Spectroscopy
LV         Latent Variable
nPLS       Multi-way Partial Least Squares
nPLS-DA    Multi-way Partial Least Squares Discriminant Analysis
OGTT       Oral Glucose Tolerance Test
PLS        Partial Least Squares
PLS-DA     Partial Least Squares Discriminant Analysis
Introduction

Many benefits can be gained if multifactorial diseases with a high incidence and prevalence are better understood. For instance, metabolic syndrome, cardiovascular diseases, obesity and diabetes type 2 as well as underlying factors such as insulin resistance, cause serious health problems. Cure and prevention is still difficult because the underlying causes are not completely understood. Therefore, studies are performed to obtain insight into molecular mechanisms of diseases in order to cure and/or prevent disease and hence to improve health status. For instance, nowadays it is thought that low-grade inflammatory status, often seen in overweight subjects, plays an important role in the development of insulin resistance (Hu et al., 2004; Spranger et al., 2003).

Techniques such as liquid chromatography mass spectrometry (LC-MS) and gas chromatography mass spectrometry (GC-MS) (Koek et al., 2006; Coulier et al., 2006), among others, are used to obtain system level information (van der Greef et al., 2007). The techniques, often employed as metabolomics tools, can ideally be used to detect metabolic aspects that are related to specific phenotypes of a disease. A property of these techniques is the generation of large amounts of data consisting of many correlated variables. In this huge amount of data, it is difficult to identify subtle intervention differences. Consequently, a deliberate experimental design and subsequent data analysis is needed in studies where small treatment effects can be expected. Such a design is focused on ruling out, as much as possible, all sources of variation other than those caused by the intervention. To increase the power of the study, repeated measurements within subjects over time can be taken or the study can be set up using a cross over study design. Hence, the treatment effects are estimated using changes within subjects rather than between subjects, which often show more variability. It will also limit the number of subjects that needs to be studied.

Differences within a subject (intra-individual) caused by an intervention are often smaller than the differences between subjects (inter-individual). Therefore, it will be difficult to detect small differences within a subject if total variance, being the sum of within and between subject variation, is taken into account. Basic multivariate data analysis tools like Partial Least Squares (PLS) (Martens and Naes, 1989; Geladi and Kowalski, 1986) and Partial Least Squares Discriminant Analysis (PLS-DA) (Barker and Rayens, 2003) do not distinguish inter- from intra- subject variation and thus are not ideal to be used. More sophisticated approaches like multilevel analyses are needed to take the full experimental design into account.

The basic idea of multilevel analysis is that different sub-models for each level of variation are generated, similar to analysis of variance (ANOVA). For instance, within and between subject variation can be split and analyzed separately if the two factors are orthogonal (i.e., not confounded) Multilevel data analysis has proven its value already in the field of metabolomics (Jansen et al., 2005) and psychometrics (Timmermans, 2006). Recently, the
use of a multilevel multivariate discriminant analysis of a metabolic experiment with a crossover design showed major advantages compared to the traditional data analysis approach (Van Velzen et al., 2008). In the present paper, the benefits of a multilevel approach in multi-way analysis will be described for the analysis of metabolomics data from a human intervention trial.

**Methods**

**Study design**

A human intervention trial was performed to gain more insight into the association between inflammatory status and insulin sensitivity in slightly overweight men. The study was designed to identify genes, proteins and metabolites responding to a diclofenac treatment as compared to a placebo treatment. Diclofenac, a non-steroidal anti-inflammatory compound, was chosen as anti-inflammatory model compound. A challenge test, the Oral Glucose Tolerance Test (OGTT), was used to determine changes related to glucose metabolism as a consequence of diclofenac treatment. The question of interest was to determine differences between the two treatment groups in their response to the challenge test after nine days of treatment compared to their response at baseline.

Twenty slightly overweight men (BMI range: 26.1 – 30.9 kg/m²) participated in the double blinded, randomized, parallel intervention trial. Ten subjects received a placebo treatment and ten subjects received diclofenac. One subject in the diclofenac treatment group dropped out. Blood samples were taken at day 0 and after 2, 4, 7 and 9 days of treatment. An OGTT using 75g glucose was performed on day 0 and day 9 during which blood was sampled at eight different time points, namely 0, 15, 30, 45, 60, 90, 120 and 180 minutes after the glucose intake. Metabolites were measured for each day and each time point, whereas the genes and proteins were measured at a selection of these. More details about the study design and data collection can be found elsewhere (Wopereis et al., 2009).

**Data set**

Four metabolomics platforms were used, namely LC-MS lipids, LC-MS free fatty acids, LC-MS polar and GC-MS global. Since the emphasis of this paper is on the analysis strategy, only the results of one of these platforms, namely the LC-MS polar data set, are presented. However, the analysis approach was applied to all platforms and results of all platforms can be found in Wopereis et al. (2009).

LC-MS polar data were corrected for the recovery of the Internal Standard (IS) for injection. Batch to batch differences were removed by synchronizing medians of quality control (QC) samples per batch. Duplicate measurements were combined into a single measurement (Bijlsma et al., 2006). When both analytical duplicates had a zero value or a non-zero value, measurements were averaged, whereas the single value was taken in case only one of the duplicates was above zero. Data were additionally cleaned up by removing glucose-related
peaks and IS-isotopes, since these could disturb the data analysis and may lead to trivial solutions. Finally, 120 peaks were included in the LC polar data set. The data set was of size $I \times (J \times K \times M)$, in which $I = 19$ subjects, $J = 120$ metabolites, $K = 8$ time points, and $M = 2$ measurement days.

**Multilevel multi-way regression**

The challenge test was used as a 'systems read-out'-parameter: the hypothesis was that the resilience of a system will be demonstrated and possibly quantified especially after stressing or perturbing a homeostatic metabolic situation. To determine differences between the two treatment groups in their response to the challenge test on day 9 compared to the day 0 response, the question of interest was stated as a multi-way regression problem. For all $I$ subjects, $J$ metabolites were measured at $K$ different time points at $M$ days. A multi-way regression problem is concerned with finding a model which predicts the value of $y$ from the data block $X$. One way of doing this is multi-way version of PLS (Martens and Naes, 1989; Geladi and Kowalski, 1986), called nPLS (Smilde et al., 2004; Smilde, 1997; Bro, 1996). In the present study, the metabolic response ($X$) is related to treatment groups, hence $y$ is not a continuous parameter as in regular regression, but a dichotomous vector containing the treatment group membership. Therefore, the model is a multi-way version of PLS-DA (Barker and Ryans, 2003), called nPLS-DA.

The following model is used:

$$T = XV$$

$$X = TG(W^M \otimes W^K \otimes W^J)' + E_X$$

$$y = TB + e_y$$

$$\text{max } \text{cov}(t_c, y^{(c-1)}); c=1, \ldots, C$$

$$w_c^M, w_c^K, w_c^J$$

where $V$ is a matrix of weighing coefficients which can be written in terms of $W$, $G$ is the core array, $B$ is the regression matrix for regressing $y$ on $T$, and $E_X$ and $e_y$ are the residuals of the model for $X$ and $y$, respectively (Smilde et al., 2004).

This model can be used to relate the metabolic response to the challenge test ($size I \times J \times K \times M$) to treatment class membership ($size I \times 1$). However, this means that both inter- and intra-individual variation is taken into account. A multilevel approach (Jansen et al., 2005; Timmermans, 2006) can be used to split the variance into a between subject (inter-individual) and a within subject (intra-individual) part, hence the metabolic changes can be investigated at different levels of variation. Since the interest is in intra-individual differences specifically, the inter-individual variation can be removed by subtracting the day 0 data from the day 9 data. This can be best illustrated using a one way ANOVA model. For
simplicity reasons, an example is given to test for treatment effects over a certain number of days at a specific time point:

\[ x_{ijk} = \mu + \alpha_i + \tau_k + \delta_j + (\tau\delta)_{kj} + (\alpha\tau)_{ik} + \epsilon_{ijk} \]  

(2)

where \( x_{ijk} \) = measurement for subject \( i \) at day \( j \) for treatment \( k \), \( \mu \) = the overall mean, \( \alpha_i \) = effect of subject \( i \), \( \tau_k \) = effect of treatment \( k \), \( \delta_j \) = effect of day \( j \), \( (\tau\delta)_{kj} \) = treatment x day interaction, \( (\alpha\tau)_{ik} \) = subject x treatment interaction, and \( \epsilon_{ijk} \) = residual error. If there are, for instance, measurements taken at two different days \( (j = 2) \), and \( x_{i2k} \) is subtracted from \( x_{i1k} \) to test for treatment effects over the two days, all terms that are independent of \( j \) are dropped out, including the effect of each individual subject \( \alpha_i \). The model that is left is:

\[ d_{ik} = \mu_k + \epsilon_{ik} \]  

(3)

where \( d_{ik} \) = the change in response for subject \( i \) for treatment \( k \), \( \mu_k \) = mean change in response for treatment \( k \), and \( \epsilon_{ik} \) = residual error. This residual error takes only the changes within a subject into account.

A multilevel multi-way model was created which regresses parameter \( y \) containing the treatment group membership to the changes in metabolic response between day 0 and day 9, \( X_9 - X_0 \) (size \( I \times J \times K \)). The model was adapted as follows:

\[
T = (X_9 - X_0) V \\
(X_9 - X_0) = TG(W^M \otimes W^K \otimes W^J)' + E_{X0-X9} \\
y = TB + e_y \\
\max \ cov(t_c; y^{(c-1)}); c=1, ..., C \\
w_{c}^{M}, w_{c}^{K}, w_{c}^{J}
\]  

(4)

Note that by using \( X_9 - X_0 \) instead of \( X \) the sets of parameters \( T, V, B, W, E \) and \( e \) in (4) are not the same as in (1). Especially \( W^M \) is different as \( M \) is 1 in (4) and 2 in (1) whereas the dimension of \( T, W^K \) and \( W^J \) is the same. The model given in (4) handles variation between two time points by subtraction. However, the method can be generalized for data with more time points then two. The creation of the \( X \)-block that was used for multilevel nPLS-DA modeling is illustrated in Figure 1. First of all, a 3-way matrix \( X_0 \) of size 19 x 120 x 8 was created out of a 19 x 960 matrix. This matrix contained the metabolic data of day 0, determined at eight different time points for each subject. A matrix \( X_9 \) of the same size was also created, containing similar information for the day 9 measurements. Finally, the \( X_0 \) matrix was subtracted from the \( X_9 \) matrix and this \( X \)-block was used for data analysis. In this way an additive treatment effect will be more clear. If the treatment effect is suspected to non additive, e.g. a multiplicative change, logarithmic transformation of data prior to subtraction can be considered to improve the results.
Centering and scaling

Data ($X_9 - X_0$) were centered across subjects and followed by auto-scaling within metabolites. The centering step was performed to remove constants, whereas the scaling to unit variance within the metabolite mode resulted in metabolite concentrations that were relative to the variation of that metabolite. By performing the scaling step after the centering step, the prior centering remained unaffected (Smilde et al., 2004; Kiers and van Mechelen, 2001; Harshman and Lundy, 1984).

**Figure 1.** The creation of the $X$-block that was used for nPLS-DA modeling.

**Model validation**

To determine the optimal number of latent variables (LVs) and to validate the multilevel nPLS-DA model, a “leave-one-subject-out” cross-validation (CV) was used (Martens and Naes, 1989). In the first CV-step, data of one subject (size $1 \times J \times K$) was left out, a multilevel nPLS-DA model was built, and the class membership of the subject who was left out was predicted. This was repeated until all 19 subjects were left out once. The error rate of the model was determined by the difference between the original class membership and the predicted one by CV. The optimal number of LVs was determined based on the minimum value of this error rate. The final fit of the model was made using the optimal number of LVs.
The nPLS-DA models were optimized by performing variable selection based on a jack-knife approach. An nPLS-DA model was made for each CV-step using data without the subject who was left out in that CV-step and using the same number of LVs that was used for the final model. This resulted in 19 sets of regression matrices of size \( J \times K \), of which the standard deviation was used to determine the relative standard deviations (RSD’s) of each regression coefficient. Only those variables which had RSD of less than 100% for all time points were included in a new data set, which was used to build a second nPLS-DA model. Components that contributed to treatment differences were identified based on absolute regression coefficients of this second model (Martens and Martens, 2001).

A permutation test was performed to test whether the treatment differences were indeed true differences. One thousand dichotomous \( y \) vectors were randomly created using the same proportion of zeros and ones as the vector that was used for modeling. For each random vector, a multilevel nPLS-DA model was made using the same “leave-one-subject-out” cross-validation approach and the cross-validated error rate was calculated. The same variables were used in the permutation test as were used in the corresponding nPLS-DA model. So, the permutation test for the original model contained all variables and the permutation test for the optimized model contained only those variables which had an RSD of less than 100% for all time points. A permutated null distribution was made of all thousand error rates and compared to the error rate for the original model in order to calculate significance of treatment differences.

**Performance**

To assess the performance of the multilevel multi-way model, also an 'ordinary' multi-way analysis was done. A 4-way nPLS-DA model (referred to as ordinary nPLS-DA in the sequel), as described in (1), was defined as the 'ordinary alternative'. The four dimensional data set (size \( I \times J \times K \times M \)) was used as \( X \)-block and the treatment class membership was used as \( y \)-vector. The error rate based on cross-validation for both models was compared. The error rate of the ordinary multi-way model was obtained using a same cross-validation procedure as was used for the multilevel approach. Also this 'ordinary model' was optimized based on jack-knifing the regression coefficients and a permutation test was performed.

**Software**

All analyses were performed using Matlab Version 7.3 2000b (The Mathworks, Inc.) and the n-way toolbox version 2.11 (Andersson and Bro, 2000).

**Results and Discussion**

**Multilevel Multi-way analysis**

A minimal cross-validated error rate of 31.5% was found for the multilevel nPLS-DA model relating the treatment group membership to the changes in metabolic response between
day 0 and day 9. Five LVs were needed for this model. The relatively high number of LVs compared to the total number of subjects illustrates the complexity of the data. As could be expected, it was not possible to describe all metabolic changes in only two or three dimensions.

The model was optimized by using a jack-knife approach. If a subject is left out and the regression coefficient changes a lot, this will result in a relatively high RSD for that particular variable. A variable with a high RSD was considered to be unstable, hence unreliable to use in explaining the differences in response between the placebo and the diclofenac group. After variable selection, a new model was made based on a subset of 31 variables. This model had a cross-validated error rate of 5% and was using 5 LVs. It appeared that variables that where most contributing to the model based on the original 120 variables were maintained after variable selection. So, essentially the same information could be described using fewer variables, illustrating the fact that many variables were unimportant for the model. The error rate of 5% meant that the treatment group membership was correctly predicted for 18 out of 19 subjects using these 31 variables. The optimized model will be used for the interpretation of the results from the multilevel multi-way models.

In Figure 2, the results of the permutation test are visualized. The vertical line represents the cross-validated error rate of the nPLS-DA model that was made, whereas the histogram represents the distribution of error rates based on permuted classes. In Figure 2a the results of the overall multilevel nPLS-DA model is given, and in Figure 2b, the results of the optimized multilevel nPLS-DA is given. The results for the overall model is very moderate (p=0.47), but the treatment differences become more clear after optimization of the model (p=0.006).

![Figure 2. Permutation test results for the original multilevel nPLS-DA model (a), the optimized multilevel nPLS-DA model (b), the original ordinary nPLS-DA model (c) and the optimized ordinary nPLS-DA model (d).](image-url)
The multi-way regression model resulted in a regression matrix of size $J^* \times K$, in which $J^*$ is the number of variables after variable selection. To determine the variables which contributed most to treatment differences, the regression coefficients were sorted by their absolute value in descending order per time point $K$. For each time point, the first ten variables were selected and used as a starting point for biological interpretation. The selected variables are presented in Table 1. The contribution of each variable to the treatment effect can be followed over time by investigating its appearance in the list of parameters that contribute most to the differences between treatments. Some metabolites were important over the whole range of time, whereas others were contributing only for a period of time. The variables which appeared in the top 10 for only one time point were initially considered to be coincidently related to the treatment. Variable 'Isoleucine + Leucine (unresolved)' (V01 in next paragraphs) and 'Glycine' (V02 in next paragraphs) will be used to illustrate further interpretation.

Table 1. Top10-ranking of metabolites which contributed most to treatment differences based on their absolute regression coefficient at time point $K$ (light grey shade: a metabolite that contributes to the response differences between treatments at each time point; dark grey shade: a metabolite that contributes to the response differences between treatments only for a period of time).
V01 is an example of a metabolite that contributes to the response differences between treatments at each measurement point, as is illustrated by the light-grey shade in Table 1. This means that the response of this metabolite between day 0 and day 9 differed during the whole time course in subjects treated with diclofenac compared to the placebo group. This effect is illustrated in Figure 3, in which the mean difference between day 9 and day 0 response for V01 is plotted per treatment group. The placebo group had at fasting state (t0) a mean change of about zero between day 9 and day 0, whereas at the same time point the diclofenac group had a mean decrease of 2.5 units. The difference between treatment groups fluctuates between 1 and 2.5 units, depending of the time point, but it remains quite stable over time. In Figure 4, the regression coefficient of this variable is plotted against the time. The same conclusion towards this metabolite can be drawn from this figure.

**Figure 3.** Mean change in metabolic response to the challenge test between day nine and day zero for subjects on placebo and diclofenac treatment, for V01 'Isoleucine + Leucine (unresolved)', a variable that contributes to treatment differences over the whole time course (error bars are based on standard errors).

**Figure 4.** Regression coefficients over time of a multilevel nPLS-DA model for V01 'Isoleucine + Leucine (unresolved)', a variable that contributes to treatment differences at each time point of the time course.
Variable V02 was only seen in the top 10 of contributing variables at 90 minutes and later of the OGTT, as illustrated by the dark-grey shade in Table 1. This metabolite was ranked 13 at t90 and therefore not included in Table 1 for this time point. The ranking at t0, t15, t30 and t45 was 21, 18, 29 and 26, respectively. So, after one hour this metabolite differed in response to the challenge test between day 9 and day 0 in subjects treated with diclofenac compared to the placebo group. This effect is illustrated in Figure 5: the differences in response are more or less the same up to 45 minutes and around zero, whereas they deviate from t60 and later. In Figure 6, the regression coefficient of this variable is plotted against the time. There is no significant contribution to treatment differences over the first 60 minutes of the curve. Only after an hour, this variable becomes more important.

**Figure 5.** Mean change in metabolic response to the challenge test between day nine and day zero for subjects on placebo and diclofenac treatment, for V02 ‘Glycine’, a variable that contributes to treatment differences in the second part of the time course (error bars are based on standard errors).

**Figure 6.** Regression coefficients over time of a multilevel nPLS-DA model for V02 ‘Glycine’, a variable that contributes only to treatment differences in the second part of the time course.
For the interpretation of the results of this type of modeling, it must be kept in mind that the regression coefficients, which were used to rank the metabolites, are based on a model in which other metabolites were also included. So, each coefficient reflects the relation between the treatment group and that particular metabolite, given the presence of the other metabolites that were used in that particular model. In Figure 4 and 6 the other metabolites are not taken into account, hence these are univariate illustrations of multivariate results.

**Multilevel approach versus ordinary nPLS-DA**

The error rate based on cross-validation for the multilevel nPLS-DA was compared with the error rate from the ordinary nPLS-DA, before and after variable selection. In total, 47.5% of the subjects were misclassified using ordinary nPLS-DA: 6 out of 10 subjects receiving placebo treatment were classified in the diclofenac group and 3 out of 9 subjects on diclofenac treatment were classified as receiving placebo treatment. The percentage of misclassified subjects is higher compared to the multilevel nPLS-DA, which had an error rate of 31.5% before variable selection. Similar results were found after variable selection. The error rate of ordinary nPLS-DA after variable selection was 42.1%, whereas this error rate was 5% for multilevel nPLS-DA. Also the results of the permutation test are worse compared to the multilevel model, which is illustrated in Figure 2. In Figure 2c the results of the overall ordinary nPLS-DA model are given, and Figure 2d shows the results of the optimized ordinary nPLS-DA. Differences between the original and the optimized model are less clear compared to the multilevel variant. Having a p-value of 0.72 and 0.95, for the overall ordinary nPLS-DA and the optimized ordinary nPLS-DA respectively, it is clear that no difference between treatments could be identified.

Between subject variation is often much larger than within subject variation and in the ordinary nPLS-DA both inter- and intra-individual variation are entangled. The between subject variation is too large to detect the subtle differences within a subject, resulting in a much higher error rate. The multilevel approach splits the variation into an inter- and intra-individual part and, in this particular case, focusing on the intra-individual differences only, much better results were obtained.

Also for the 4-way analysis, the regression vector provides information on the contribution of each metabolite to the discrimination between treatment groups. V02, which was of any importance only after 1 hour based on the multilevel approach, appeared also in the top of the 4-way analysis. V02 was ranked around place 5 for each time point and for both days. However, V01 did not appear in the top of important metabolites at all. For some time points, the regression coefficient for V01 was even equal to zero, meaning that it had no contribution at all to the treatment difference.
Biological validation

Diclofenac is known to inhibit and activate several enzymes and transporters among which the inhibition of the enzyme aminopeptidase N (CD13) (Boelsterli, 2003; Ware et al., 1998). CD13 is a broad specificity aminopeptidase that cleaves specifically the N-terminal bound neutral amino acids from oligopeptides. Especially essential neutral amino acids, like L-isoleucine, L-leucine, L-methionine, L-threonine, L-phenylalanine, L-valine and L-tryptophan are expected to show lower plasma concentration in diclofenac treated subjects, whereas most of the basic, acidic and non-essential neutral plasma amino acids, among which L-glycine, are expected not to show this concentration difference. Multiple metabolic intermediates of glutathione metabolism showed time-dependent suppression in response to the oral glucose tolerance test, among which glycine, but also 5-oxoproline and glutamic acid. The glutathione synthesis pathway is insulin sensitive and the difference in response suggests that diclofenac treatment may alter insulin signaling in overweight men (for more details see Wopereis et al., 2009).

Variable VO1 and VO2 were identified as Isoleucine + Leucine (unresolved) and Glycine, respectively. In the multilevel approach, Isoleucine + Leucine (unresolved) was found to be of high importance for explaining differences between the two treatment groups. Glycine appeared in the top 10 only after 1 hour. In ordinary nPLS-DA, Glycine was of importance at each time point, but Isoleucine + Leucine (unresolved) was of no importance at all. Given the effect of diclofenac on CD13 and its effects on amino acids, it can be concluded that the multilevel approach found the effects that were expected, whereas the ordinary nPLS-DA failed to do so.

Multilevel nPLS-DA revealed various metabolites from the same pathway that where contributing to treatment differences, which also endorses to the strength of the methodology. Findings that were found for the LC global platform were also confirmed by the GC-MS platform. Since an in-depth exploration of the biological aspects of the study are beyond scope of the present paper, these results are not presented in more detail. In Wopereis et al. (2009), the biological interpretation is discussed in full detail.

Conclusions

In many (nutritional related) -omics studies, effects on subjects are subtle and hidden in the data. For some study designs it is possible to discover these small differences by using multilevel modeling.

The multilevel multi-way technique turned out to be a much stronger tool for modeling differences between treatment groups than the ordinary method. Taking into account the multilevel structure of the data, the modeling results can be improved. By splitting the variation into an inter- and intra-individual part, it is possible to focus on different variation sources in the data. In the present study, the between subject variation was left out, so that metabolites that contributed to the subtle differences between treatments in response to
the challenge test could be identified. The multilevel approach found the effects that were better interpretable, whereas the ordinary nPLS-DA failed to do so.

The methodology that was described in this paper is not limited to human intervention studies only, but can also be used for studies with similar data structures. The multilevel approach improves the interpretability of the results by taking into account the various levels of variation in a given design.

References


METABOLIC PROFILING OF THE RESPONSE TO AN ORAL GLUCOSE TOLERANCE TEST DETECTS SUBTLE METABOLIC CHANGES

Suzan Wopereis¹, Carina M Rubingh¹, Marjan J van Erk¹, Elwin R Verheij¹, Trinette van Vliet¹,², Nicole HP Cnubben¹, Age K Smilde¹,³, Jan van der Greef¹, Ben van Ommen¹, Henk FJ Hendriks¹

¹Department Quality of Life, TNO, Zeist, The Netherlands
²Currently at: CCMO, the Hague, The Netherlands
³Biosystems Data Analysis, SIRS, University of Amsterdam, Amsterdam, The Netherlands

Abstract

Background: The prevalence of overweight is increasing globally and has become a serious health problem. Low-grade chronic inflammation in overweight subjects is thought to play an important role in disease development. Novel tools to understand these processes are needed. Metabolic profiling is one such tool that can provide novel insights into the impact of treatments on metabolism.

Methodology: To study the metabolic changes induced by a mild anti-inflammatory drug intervention, plasma metabolic profiling was applied in overweight human volunteers with elevated levels of the inflammatory plasma marker C-reactive protein. Liquid and gas chromatography mass spectrometric methods were used to detect high and low abundant plasma metabolites both in fasted conditions and during an oral glucose tolerance test. This is based on the concept that the resilience of the system can be assessed after perturbing a homeostatic situation.

Conclusions: Metabolic changes were subtle and were only detected using metabolic profiling in combination with an oral glucose tolerance test. The repeated measurements during the oral glucose tolerance test increased statistical power, but the metabolic perturbation also revealed metabolites that respond differentially to the oral glucose tolerance test. Specifically, multiple metabolic intermediates of the glutathione synthesis pathway showed time-dependent suppression in response to the glucose challenge test. The fact that this is an insulin sensitive pathway suggests that inflammatory modulation may alter insulin signaling in overweight men.
Introduction

The low-grade inflammatory state often seen in overweight subjects is thought to play an important role in lifestyle associated disease development. This inflammatory state has been associated with cardiovascular diseases (Hansson, 2005), diabetes, insulin resistance (Dandona et al., 2004) and cancer (Fogarty et al., 2008). Since the early 1990s (Bonassi et al., 2001), considerable effort has been made to discover and validate biomarkers with diagnostic or prognostic utility for lifestyle associated diseases (Musaas and Haynes, 2001; Khuseyinova and Koenig, 2006; Rajaram et al., 2004). Metabolites such as cholesterol, fasting glucose and homocysteine have long been used as biomarkers. Genomic – based technologies such as metabolic profiling provide a new means to explore the combination of multiple metabolites as a biomarker, which may allow for more precise outcome predictions. Alternatively, such a biomarker may provide a more comprehensive insight into pathophysiological processes not previously attainable with traditional biomarkers (He, 2006; Ginsburg and Haga, 2006; Wang et al., 2006). These markers should respond to nutritional and pharmaceutical interventions in order to be evaluated.

The main focus of the present study was to demonstrate and quantify the consequence of using diclofenac to reduce inflammation and its effect on metabolism. Subsequently, the study was geared to identify multiple metabolites to be used as a potential biomarker. Diclofenac acts as a non-selective inhibitor of the enzymes cyclooxygenase-1 and -2. Cyclooxygenases catalyze among other things the formation of prostaglandins that act as messenger molecules in inflammation. Metabolic profiling has been shown to be a valuable tool to quantify nutritional metabolic homeostasis and disease mechanisms associated with metabolic stress and metabolic syndrome (Kaddurah-Daouk et al., 2008; Naylor et al., 2008; Yetukuri et al., 2007). Liquid and gas chromatography mass spectrometric methods (LC-MS and GC-MS) were used to detect high and low abundant metabolites in plasma to obtain a comprehensive picture of metabolic changes induced by a mild anti-inflammatory drug intervention. A total of 343 plasma metabolites were quantified, of which 204 could be identified, spanning diverse chemical classes (Supplement Table S1). The metabolic profiling approach was not only applied in fasting (homeostatic) conditions, but also at multiple time points during an oral glucose tolerance test (OGTT). This approach is based on the concept that the resilience of the system can be assessed after challenging or perturbing a homeostatic situation. Plasma metabolic profiling combined with a glucose challenge has already been successfully used to differentiate between healthy individuals and individuals with an impaired glucose tolerance (Shaham et al., 2008). Applying a metabolic perturbation and metabolic profiling may help identify a set of metabolites that predict differences in the responses between treatment groups to the oral glucose tolerance test. A similar set of metabolites might then provide novel insight into the interplay between metabolic and inflammatory processes and provide candidate biomarkers to be applied in (intervention) studies aimed at lifestyle associated diseases.
Materials and Methods

Ethics statement

The study was approved by the Medical Ethics Committee of the University Medical Centre of Utrecht (May 17, 2005). In total, fifty subjects gave written informed consent after being informed about the study, both verbally and in writing.

Subjects and study design

The study was conducted at TNO Quality of Life (Zeist, the Netherlands). Overweight and mildly obese men (Body Mass Index (BMI) between 25.1 and 34.0 kg/m²) were recruited from a pool of volunteers. All fifty subjects completed a questionnaire on medical history and were physically examined. Blood and urine were collected after an overnight fast for routine analysis. In addition, plasma hsCRP levels were determined.

Subjects met the following inclusion and exclusion criteria. Smokers, subjects who reported that they were trying to lose weight or who were on a medically prescribed diet and subjects with allergy or hypersensitivity for non-steroidal anti-inflammatory drugs were excluded from participation. Additionally, subjects who were on medication that may have interfered with parameters to be measured or with diclofenac treatment or subjects who, based on anamnesis, were not suitable to receive diclofenac treatment (history of current gastrointestinal diseases including bleeding, ulcer or perforation, history of stroke, history of current significant haematological disorders, any significant hepatic, renal or cardiovascular disease, asthma) or subjects with a history of medical or surgical events that may have affected the study outcomes were not included. Based on these criteria, 25 subjects were eligible. Of the 25 eligible subjects, the 5 subjects with the lowest CRP values were not included in the study. Levels of hsCRP of the included subjects ranged from 0.41 – 9.72 mg/L (see also Table 1).

The study was designed as a double blind, randomized, parallel trial, in which subjects were treated with diclofenac (n=10) or placebo (n=10). Randomization of subjects to treatment groups was restricted by hsCRP, body mass index (BMI), fasting glucose and age. The result is a homogeneous division of these parameters over the two treatment groups at the start of the study (see Table 1). Subjects consumed one capsule (placebo or 50 mg diclofenac) approximately one hour before breakfast, lunch and dinner for 9 days. Subjects were instructed to keep to their habitual diet during the study. One person dropped out on the first day of the study for study unrelated reasons. Nineteen men completed the study. Their subject characteristics are presented in Table 1. Prostaglandin E2 concentrations were used as a readout for diclofenac treatment and showed a significant reduction in subjects treated with diclofenac (p = 0.02). Prostaglandin E2 concentrations were unchanged in subjects treated with placebo demonstrating a modulation of the inflammatory status in diclofenac treated subjects (Table 2).
Table 1. Demographic data of subjects that completed the study (n=19) at screening; mean ± SD (range)

<table>
<thead>
<tr>
<th></th>
<th>All (n=19)</th>
<th>Placebo treatment (n=10)</th>
<th>Diclofenac treatment (n=9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>43 ± 15</td>
<td>41 ± 16 (19 – 60)</td>
<td>45 ± 15 (21 – 58)</td>
</tr>
<tr>
<td>Body weight (kg)</td>
<td>93.5 ± 8.0</td>
<td>93.5 ± 9.3 (81.1 – 105.2)</td>
<td>93.5 ± 6.9 (85.2 – 104.4)</td>
</tr>
<tr>
<td>Height (m)</td>
<td>1.82 ± 0.08</td>
<td>1.82 ± 0.10 (1.69 – 1.96)</td>
<td>1.83 ± 0.07 (1.70 – 1.92)</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>28.1 ± 1.2</td>
<td>28.1 ± 1.0 (26.7 – 29.3)</td>
<td>28.1 ± 1.5 (26.1 – 30.9)</td>
</tr>
<tr>
<td>hs-CRP (mg/L)</td>
<td>2.22 ± 2.33</td>
<td>2.08 ± 1.88 (0.41 – 6.35)</td>
<td>2.37 ± 2.87 (0.64 – 9.72)</td>
</tr>
<tr>
<td>Fasting glucose (mmol/L)</td>
<td>6.0 ± 0.5</td>
<td>5.9 ± 0.5 (5.2 – 7.1)</td>
<td>6.0 ± 0.6 (5.0 – 6.8)</td>
</tr>
<tr>
<td>Fasting insulin (mU/L)</td>
<td>13.4 ± 8.1</td>
<td>13.4 ± 8.6 (5.1 – 26.8)</td>
<td>13.3 ± 8.1 (3.3 – 26.6)</td>
</tr>
</tbody>
</table>

Blood samples were taken after an overnight fast on days 0, 2, 4, 7 and 9. Subjects underwent an oral glucose tolerance test (OGTT) on day 0 and day 9. Blood samples were taken just before (0 minutes) and 15, 30, 45, 60, 90, 120 and 180 minutes after the administration of the glucose solution (75 grams). Samples were analyzed for glucose and insulin for which the incremental area under the response curves (AUC) was calculated. Table 2 shows the characteristics of these parameters. No significant changes were observed between the treatments. Figure 1 shows study design and time points at which metabolic profiling measurements were done.

Figure 1. Overview of study design, time points at which metabolome was measured and multivariate data analyses. To determine metabolites that were modulated by the diclofenac treatment the following multivariate data comparisons were performed to identify metabolites that were modulated by the diclofenac treatment: a) PLS-DA on metabolic profiling data from day 9 subtracted by metabolic profiling data from day 0, on fasted plasma samples; b) n-PLS-DA on metabolic profiling data from day 0, 2, 4, 7 and 9, on fasted plasma samples; c) n-PLS-DA on metabolic profiling data from day 9 subtracted by metabolic profiling data from day 0, using the fasted plasma samples and the samples after glucose administration, thus metabolic profiling data on 0, 15, 30, 45, 60, 90, 120 and 180 minutes after glucose administration. The multivariate data comparisons from a-c were performed per metabolite platform, thus multivariate models were created for GC-MS global, LC-MS polar, LC-MS lipids and LC-MS free fatty acids data.
Table 2. Characteristics of prostaglandin E2 and OGTT parameters (insulin and glucose) measured at start and end of treatments. AUC = area under the curve.

<table>
<thead>
<tr>
<th></th>
<th>Placebo Mean (sd)</th>
<th>Diclofenac Mean (sd)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Day 0</td>
<td>Day 9</td>
</tr>
<tr>
<td>PGE2 [pg/mL]</td>
<td>56.5 (7.7)</td>
<td>60.8 (11.0)</td>
</tr>
<tr>
<td>AUC glucose [mmol*min/L]</td>
<td>409 (238)</td>
<td>306 (197)</td>
</tr>
<tr>
<td>AUC insulin [mU*min/L]</td>
<td>8986 (5356)</td>
<td>8935 (4821)</td>
</tr>
</tbody>
</table>

Plasma PGE2, insulin and glucose measurements

PGE2 was determined using the Prostaglandin E2 \(^{[125]}\) Biotrak assay system (Amersham Biosciences, UK) with modifications. In short, PGE2 in samples was derivatized to the methyl oximate derivative. The resulting solution was further diluted (final dilution 5 times) in PBS and partly assayed. The assay consists of incubation of the oximated sample PGE2, the \(^{125}\)I-labelled PGE2, and a PGE2 specific antibody. After incubation, the Amerlex-M reagent is added and the free and bound \(^{125}\)I labeled PGE2 separated using centrifugation. The resulting bound radioactivity in the pellet is determined using a gamma-counter. Serum glucose concentrations were measured by using a commercial test kit (Roche diagnostics GmbH, Mannheim, Germany) on a Hitachi 911 automatic analyzer (Hitachi Instrument Division, Ibaraki-ken, Japan), with intra-assay CVs that ranged between 0.7% and 0.9%, depending on the concentration. Serum insulin concentrations were measured using an AIA-600 Immunoassay Analyzer (Tosoh Corporation, Toyama, Japan), with intra-assay CVs that ranged between 4.3% and 5.8%, depending on the concentration.

Metabolic profiling measurements

**LC-MS of lipids and fatty acids** Plasma lipids and free fatty acids (FFA) were analyzed with electrospray LC-MS (Bijlsma et al., 2006; Verhoeckx et al., 2004). The instrument used was a Thermo LTQ equipped with a Thermo Surveyor HPLC pump. Data were acquired by scanning the instrument form m/z 300 to 1200 at a scan rate of approximately 2 scans/s. The FFA LC-MS platform employs the same HPLC conditions as the lipid method except for the gradient. Detection of FFA is performed in negative ion mode, and lipids are measured in positive ion mode. Taken together the two methods can measure approximately 200 different identified lipids and FFA.

In summary, 10 µl of plasma was extracted with 300 µl of isopropanol containing several internal standards (IS: C17:0 lyso-phosphatidylcholine, di-C12:0 phosphatidylcholine, tri-C17:0 glycerol ester, C17:0 cholesterol ester and heptadecanoic acid (C17:0)). Each extract was injected three times (10 µl), once for the LC-MS FFA platform and two times for the LC-MS lipid platform. Furthermore, a quality control (QC) sample was prepared by pooling of plasma from all subjects. The pool was divided into 10 µl aliquots that were extracted the same as the study samples. The QC samples were placed at regular intervals in the analysis sequence (one QC after every 10 samples). The QC samples served two purposes. The first is a regular quality control sample to monitor the LC-MS response in time. After the response
has been characterized, the QC samples were used as standards of unknown composition to calibrate the data (van der Greef et al., 2007).

In the plasma samples, the 6 dominant lipid classes observed with these two methods are the Lyso-phosphatidylcholines (IS used: C17:0 lyso-phosphatidylcholine), Phosphatidylcholines (IS used: di-C12:0 phosphatidylcholine), Sphingomyelines (IS used: di-C12:0 phosphatidylcholine), Cholesterolesters (IS used: C17:0 cholesterol ester), Triglycerides (IS used: tri-C17:0 glycerol ester), and free fatty acids (IS used: C17:0 FFA). In addition to these lipids, the extracts also contain minor lipids, but these were either not detected (concentration too low relative to very abundant lipids like phosphatidylcholines and triglycerides) or they were not included in data processing. The LC-MS lipid and LC-MS FFA data were processed using the LC-Quan software (Thermo).

**LC-MS polar** Polar plasma metabolites were analyzed using LC-MS after derivatization (butylation). The metabolites were extracted from 10 µl plasma with 200 µl methanol containing internal standards (deuterated amino acids). After the methanol evaporated, the extract was dissolved in 100 µl n-butanol containing 4 M/l hydrochloric acid and heated to 65 °C for 60 min. After freeze drying the extracts were dissolved in 100 µl 0.1% formic acid in water, and 10 µl was injected for LC-MS analysis using a Thermo LTQ equipped with an ESI interface and a Surveyor HPLC system. QC samples, prepared from pooled plasma, were analyzed after every 6th study sample. The mass spectrometer was operated in positive ion mode and data were acquired by scanning from m/z 125 to 1000 at approximately 2 scans/s. The HPLC method consisted of an Intersil ODS 3 column (100 x 3 mm id) in combination with an acetonitrile gradient (5 to 80% in 20 min at a flow of 0.3 ml/min) in 0.1% formic acid.

Data were processed with TNO comprehensive peak picking software (IMPRESS; Clish et al., 2004; van der Greef et al., 2004) to find consistent features in the LC-MS files. These features, after de-isotoping, were used for data processing with Thermo LC-Quan software.

**GC-MS global** The GC-MS method used for analyzing a broad range of metabolites was identical to the method reported for microbial metabolic profiling (Koek et al., 2006), except for the sample type. Plasma samples (100 µl) were extracted with methanol and after evaporation the metabolites were derivatized (oximation and silylation). QC samples, prepared from pooled plasma, were analyzed after every 10th study sample.

**Performance of metabolic profiling platforms** The performance of the applied metabolic profiling platforms is assessed through the frequent analysis of the QC sample (van der Greef et al., 2007). This QC sample, prepared by pooling selected study samples, represents the full biochemical diversity of the study samples and allows the calculation of the analytical precision for all metabolites measured. The QC sample data is also used to correct systematic errors (e.g. batch to batch response differences) by a single point calibration model. Typically, this procedure offers excellent precision for a large majority of metabolites (e.g. 50% of the metabolites have an RSD of less than 10%, 75% with an RSD less than 20%). Metabolites with very high imprecision e.g. RSD > 50%, were removed from the data unless
large differences between treatment groups were observed. Furthermore, method performance was carefully monitored using multiple internal standards (5 to 10 depending on method, including analogues, 2H and 13C labeled metabolites) and duplicate analysis of samples. The metabolite data used for statistical data analysis in this study met all of our quality requirements.

**Preprocessing of metabolic profiling data** Data for each subject were corrected for the recovery of the IS for injection. Batch to batch differences in data were removed by synchronizing medians of QC-samples per batch. For all platforms, duplicate measurements were combined into a single measurement. When both analytical duplicates had a zero value or when both had a non-zero value, measurements were averaged. The single value was taken when only one of the duplicates was above zero (Bijlsma et al., 2006). To avoid trivial results, data were additionally optimized by removing glucose-related metabolites and IS-isotopes in the LC-MS polar data and two glucose metabolites in the GC-MS global data set. The correlation between these glucose peaks and glucose measured by a commercial test kit (Roche Diagnostics GmbH, Mannheim, Germany) on a Hitachi 911 automatic analyzer (Hitachi Instrument Division, Ibaraki-ken, Japan) was 0.97 and 0.98 respectively. Finally, the LC-MS FFA data set contained 14 metabolites, the LC-MS Lipids data set consisted of 61 metabolites, 120 metabolites were included in the LC-MS polar data set and the GC-MS data set contained 137 metabolites.

**Multivariate analysis of metabolic profiling data**

**Two-way analysis: PLS-DA** Partial Least Squares Discriminant analysis (PLS-DA; Barker and Rayens; 2003) was used to identify metabolites that differ in their change between day 0 and day 9 in fasted conditions between treatment groups (Figure 1, analysis a). In PLS-DA, a Y-variable containing class membership information is correlated to a data matrix (X-block). The subjects who received the placebo treatment were assigned to class ‘0’ and the subjects who received diclofenac were assigned to class ‘1’. Since the interest was in intra-individual differences between day 0 and day 9, the X-block was defined for each metabolite platform by subtracting the day 0 values from the day 9 values, which removed differences in baseline.

**Two-way analysis: model validation and optimization** Cross-validation was used to validate the PLS-DA models, using a ‘leave-one-out’ cross-validation scheme (Martens and Naes, 1989). Data of one subject were left out in the first cross-validation step, a PLS-DA model was built, and the treatment class membership of the subject that was left out was predicted. This was repeated until all 19 subjects were left out once. The error rate of the model was determined by comparing the original class membership and the predicted one. The optimal number of LVs was determined based on the minimum value of this error rate. The final fit of the model was made using this number of optimal LVs.

PLS-DA models for which an error rate was found below 35% were optimized by performing metabolite selection based on a jackknife approach (Martens and Naes, 1989). Data of one
subject were left out and a PLS-DA model was made using the same number of LVs that was used for the final model. This was repeated until all 19 subjects were left out once. This resulted in 19 sets of regression coefficients, of which the standard deviation was used to determine the relative standard deviations (RSDs) of each regression coefficient. Only those metabolites that had a RSD of less than 50% were included in a new data set. This set was used to build a second PLS-DA model. Metabolites that contributed to treatment differences were identified based on absolute regression coefficients of this second model.

**Three-way analysis: n-PLS-DA**

To identify metabolites that differed in changes over time between the treatment groups, it was necessary to discriminate between the time and the metabolite information. Basic multivariate data analysis tools like Principal Component Analysis (PCA; Dillon and Goldstein, 1984; Joliffe, 1986; Massart et al., 1997; Vandeginste et al., 1998) and Partial Least Squares Discrimant analysis (PLS-DA; Barker and Rayens, 2003) are not sufficient to analyze the data sets, since these methods do not separate the time factor from the metabolites. Therefore, the multi-way generalization of these two-way techniques, nPLS-DA, (Bro, 1996; Smilde, 1997) was used for the analyses. A so called 3-way matrix was created, having size $19 \times J \times T$ where $J$ is equal to the number of metabolites of a particular platform and $T$ is equal to the number of time points, which were either the days 0, 2, 4, 7, and 9 (Figure 1, analysis b) or the time points after glucose administration on day 0 and 9 (Figure 1, analysis c). In order to focus the analysis on changes over time within a subject, the day 0 data were subtracted from the day 9 data.

The GC-MS global and LC-MS polar data sets were centered across subjects and followed by scaling within the metabolite-mode $J$, whereas the LC-MS lipids and fatty acids data sets were only centered across subjects. The centering step was performed to remove constants between the subjects, whereas scaling within the metabolite mode resulted in standardized metabolites. By performing the scaling step after the centering step, the prior centering remained unaffected (Harhsman and Lundy, 1994; Kiers and van Mechelen, 2001; Smilde et al., 2004).

**Three-way analysis: model validation and optimization**

Cross-validation was used to validate the nPLS-DA models, using a ‘leave-one-subject-out’ cross-validation scheme. Data of one subject (all measurements for all metabolites for one subject) were left out in the first cross-validation step, an nPLS-DA model was built using data of the remaining subjects, and the treatment class membership of the subject that was left out was predicted. This was repeated until all 19 subjects were left out once. The error rate of the model was determined by comparing the original class membership and the predicted one. The optimal number of LVs of the nPLS-DA model was determined based on the minimum value of this error rate.

In order to optimize the nPLS-DA models, metabolite selection has been performed using a jackknife approach. Data of one subject (all measurements for all metabolites for one subject) were left out and an nPLS-DA model was made using the same number of LVs that
was used for the final model. This was repeated until all 19 subjects were left out once. This resulted in 19 sets of regression coefficients, of which the standard deviation was used to determine the RSD's of each regression coefficient for each metabolite and each time point. A second nPLS-DA model was build using only those metabolites which showed relatively constant regression coefficients over time.

A permutation test was performed to test whether the treatment differences were indeed true differences similar as described by Bijlsma et al. (2006). Therefore, the Y-variable containing class membership information was randomized a 10000 times. For each random vector, a multilevel nPLS-DA model was made using the same (optimal) number of LVs as determined previously. For every nPLS-DA model built, a sum of squares between/sum of squares within ratio (B/W) was calculated for the class assignment predictions. These distributions of random class assignments can be plotted in a histogram and compared to ratio for the original model. The model is classified as ‘bad’ if the B/W of the model is plotted in the lower half of the B/W distribution of random class assignments; the model is classified as ‘moderate’ if the B/W of the model is plotted in the upper half of the B/W distribution of random class assignments; the model is classified as ‘good’ if the B/W of the model is larger than the B/W distribution of random class assignments.

All analyses were performed using Matlab Version 7.0.4 R14 (The Mathworks, Inc.) and the n-way toolbox version 2.11 (Andersson and Bro, 2000).

Annotation and Identification of metabolites

The nPLS-DA model resulted in a regression matrix of size $J^* \times K$, in which $J^*$ is the number of metabolites after variable selection. To determine the variables which contributed most to treatment differences, the regression coefficients were sorted by their absolute value in descending order per time point. Since the regression coefficients decreased gradually between the highest and the lowest value due to the use of autoscaled data, there was no sharp cutoff. Therefore, for each time point the first ten peaks with unknown identity were selected and used for metabolite identification.

Metabolites were annotated using an in-house metabolite database containing retention time information, MS spectra (EI for GC-MS data), MS/MS spectra (LC-MS) and accurate mass data (LC-MS) of reference substances. The confidence of identification is 100% unless indicated otherwise. Accurate mass MS and MS/MS data of reference substances and metabolites in the study samples were acquired using Thermo LTQ-FT and Thermo LTQ-Orbitrap instruments.

Biological interpretation of metabolic profiling data

For each metabolite, the mean treatment effect difference between day 9 and day 0 during the OGTT time course was calculated as follows:

$$\bar{x}_m = \sum_{t=1}^{T}(100 \times ((\bar{x}9_m - \bar{x}0_m)_t - (\bar{y}9_m - \bar{y}0_m)_t)/\bar{y}0_{mt})/T$$

(1)
Where

\[ \bar{x}_m = \text{mean treatment effect for metabolite } m \text{ in differences between day 9 and day 0 (\%)} \]

\[ \bar{x}_9^0 = \text{mean intensity for metabolite } m \text{ on day 9 for the diclofenac treatment group} \]

\[ \bar{x}_0^0 = \text{mean intensity for metabolite } m \text{ on day 0 for the diclofenac treatment group} \]

\[ \bar{y}_9 = \text{mean intensity for metabolite } m \text{ on day 9 for the placebo group} \]

\[ \bar{y}_0 = \text{mean intensity for metabolite } m \text{ on day 0 for the placebo group} \]

\[ T = \text{total number of time points (t=1 ... 8)} \]

Ranges in terms of percentages were calculated per metabolite as the minimum and maximum value of the treatment effects calculated per time point. The minimum and maximum value of treatment effects were calculated per time point per metabolite and these values were used to determine the ranges of treatment effects in terms of percentages.

Detailed pathway and biological network analysis was performed in Metacore version 4.3 (GeneGo Inc., St. Joseph, MI, USA). Only curated interactions were used for biological network analysis. The following metabolites were not available in Metacore and therefore not used for pathway and network analysis: 1,2-diglyceride (C36:2), 2,3,4-trihydroxybutanoic acid and 2-amino-2-methyl butanoic acid, 1-aminocyclopentanecarboxylic acid. Pathway maps were edited in Mapeditor (GeneGo Inc., St. Joseph, MI, USA) version 2.1.0.

**Results**

**Statistical selection of relevant metabolites for diclofenac treatment**

Metabolites that changed due to diclofenac treatment were identified using various multivariate comparisons. Table 3 shows the results of these multivariate analyses of the different datasets derived from the four metabolic profiling platforms.

Table 3. Multivariate data analysis of various metabolic profiling datasets. The results of the different multivariate models are expressed as error rates.

<table>
<thead>
<tr>
<th>Method</th>
<th>GC-MS global</th>
<th>LC-MS polar</th>
<th>LC-MS Lipids</th>
<th>LC-MS FFA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS-DA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day 9 vs Day 0</td>
<td>42.0%</td>
<td>53.0%</td>
<td>37.0%</td>
<td>37.0%</td>
</tr>
<tr>
<td>n-PLS-DA</td>
<td>Day 0, 2, 4, 7 and 9</td>
<td>57.0%</td>
<td>42.0%</td>
<td>52.5%</td>
</tr>
<tr>
<td>n-PLS-DA</td>
<td>Day 9 vs Day 0; 0-15-30-45-60-90-120-180 min</td>
<td>31.5%</td>
<td>31.5%</td>
<td>21.0%</td>
</tr>
<tr>
<td>n-PLS-DA after metabolite selection</td>
<td># = 77</td>
<td># = 31</td>
<td># = 25</td>
<td>NA</td>
</tr>
<tr>
<td>Day 9 vs Day 0; 0-15-30-45-60-90-120-180 min</td>
<td>10.5%</td>
<td>5.0%</td>
<td>16.0%</td>
<td>NA</td>
</tr>
</tbody>
</table>

*Metabolite selection was only applied if the error rate of the original model of the complete dataset was below 35% and the dataset contained more than 50 metabolites. NA, not analyzed.*
A comparison of the fasting state metabolomes between the subjects treated with placebo and diclofenac on day 9 compared to day 0 resulted in PLS-DA models (Figure 1, analysis a) with high error rates. This indicated that there was no significant difference in fasted plasma samples.

Expanding the n-PLS-DA models with the metabolic profiling data from the plasma samples taken in fasted conditions on several intermediate days (Figure 1, analysis b) during the intervention for the various metabolic profiling platforms also resulted in high error rates. This confirms that differences in metabolic changes could not be detected between subjects treated with placebo and diclofenac in the fasted (homeostatic) condition.

Metabolic perturbation by the OGTT improved the metabolic profiling-based differentiation between the treatments. The n-PLS-DA models on plasma samples taken at 8 point time course after the glucose administration to subjects on day 9 vs day 0 (Figure 1, analysis c) resulted in improved error rates. This was true for all metabolic profiling platforms as compared to PLS-DA and n-PLS-DA models based on plasma in fasted conditions, between the control and anti-inflammatory treatment. The n-PLS-DA models for metabolome data from GC-MS global, LC-MS polar and LC-MS lipids were analyzed further with metabolite selection methods. The n-PLS-DA models after metabolite selection for GC-MS global and LC-MS polar metabolome data resulted in models with error rates of 10.5% and 5.0% respectively, indicating that the GC-MS global model and the LC-MS polar model only misclassified respectively 2 and 1 persons of the total 19. The model of the LC-MS lipids-dataset resulted in a model with an error rate of 16%, indicating that 3 out of 19 total subjects were misclassified. Permutation tests were performed on these 3 models to validate the significance of the treatment differences. The results of the GC-MS global and LC-MS polar permutation test were ‘good’, in contrast to the results for the LC-MS lipids model which were ‘moderate’.

Overall, significant metabolic changes due to the treatment could be detected only in the metabolic profiling data of the OGTT time course. Only metabolome data from the models with ‘good’ results for the permutation test, thus the LC-MS polar and GC-MS global models (after metabolite selection), were used for further interpretation in the statistical analysis.

**Metabolite identification**

The metabolites with the highest absolute regression coefficients per time point (thus 0, 15, 30, 45, 60, 90, 120 and 180 minutes) were selected from the LC-MS polar and GC-MS global n-PLS-DA models as being most discriminative between subjects treated with diclofenac and placebo. The intersection of the regression coefficients per time point resulted in a total of 15 unique metabolites from the GC-MS global dataset and in a total of 24 unique metabolites from the LC-MS polar dataset for metabolite identification. Ultimately, 69% of the selected metabolites could be identified (14 out of 15 in the GC-MS global dataset; 13 out of 24 in the LC-MS polar dataset). Table 4 lists the most discriminating metabolites that could be identified. Only the identified metabolites were used for further interpretation.
Analysis of metabolite response during OGTT time course

The metabolite response was tracked by plotting the mean difference between day 9 and day 0 per time point per treatment group. In general, two different metabolite challenge test responses were distinguished as illustrated in Figure 2. Most of the selected metabolites (81%, Table 4) showed a difference in offset that is constant during the OGTT time course (Figure 2A, response type A). In other words, these metabolites are discriminating between the treated and untreated subjects independent of time during the OGTT time course. This indicates that only minor differences exist between the treatment groups and that these differences can only be identified by repeated measurements. Indeed, time independent PLS-DA analysis yielded a similar error rate (11%).

Some of the selected metabolites (19%), however, showed only a contribution to treatment differences in the second part of the OGTT time course (Figure 2B, response type B). This indicates that these 4 metabolites only differed between treatments when challenging the metabolic situation, leading to alterations in dynamic response to the perturbation. Indeed, time independent PLS-DA analysis increased the error rate of the LC-MS polar model (26%).

Furthermore, Table 4 shows the semi-quantified treatment effects, expressed as the mean change (in %) over time. The majority of the metabolites that had a constant contribution to treatment differences over time (categorized as response type A) showed a decreased concentration in plasma in response to diclofenac treatment compared to subjects treated with placebo. Several amino acids (n=6), organic acids (n=7), carbohydrates (n=2) and fatty acids & lipids (n=3) were categorized with a response type A. Some metabolites were identified as being discriminating between the treatments in data from both analytical platforms (isoleucine, 1-aminocyclopentanecarboxylic acid and 4-hydroxyproline), validating their contribution to the differences between the treatment groups.

All metabolites that specifically showed a dynamic response to the perturbation (categorized as response type B), showed higher concentrations in the diclofenac treated group. Based on plots of mean changes over time, it appeared that mean concentrations of diclofenac treated subjects remained constant, whereas mean concentrations of placebo treated subjects dropped during a specific phase of the time course (Figure 2b). The amino acids glycine, aspartic acid and glutamic acid and the organic acid 5-oxoproline were categorized with a response type B.
Table 4. Overview of most discriminating metabolites, their treatment effect and their metabolite response in the OGTT time course. Per analytical platform the metabolites are ranked according to their importance to the model, thus uric acid and isoleucine contributed most in the discrimination between the treatment groups.\(^1\)

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>Treatment effect (%)</th>
<th>Mean</th>
<th>Range</th>
<th>Response type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uric acid</td>
<td>-12 -18 1</td>
<td>GC-MS</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>1,2-diglyceride (C36:2)</td>
<td>-18 -28 -9</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Proline</td>
<td>9 6 13</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>-18 -22 -13</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>1-aminocyclopentanecarboxylic acid</td>
<td>-2 -5 2</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Threonine</td>
<td>-13 -15 -11</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>4-Hydroxyproline</td>
<td>-59 -72 -51</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>2,3,4-Trihydroxybutanoic acid</td>
<td>15 9 24</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Aminoacidic acid</td>
<td>-26 -43 -16</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Arabitol, ribitol, or xylitol</td>
<td>10 5 13</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Ornithine</td>
<td>-12 -20 2</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Mannose or galactose</td>
<td>10 6 14</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Palmitoleic acid (C16:1)</td>
<td>23 -2 55</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Palmitic acid (C16:0)</td>
<td>6 -13 24</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>-23 -34 -15</td>
<td>LC-MS</td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Glycine</td>
<td>8</td>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>2-Amino-2-methyl butanoic acid</td>
<td>-4 -9 -1</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>5-Oxoproline</td>
<td>14</td>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>1-Aminocyclopentanecarboxylic acid</td>
<td>-69 -91 -35</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>4-Hydroxyproline</td>
<td>-52 -61 -38</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Isoleucine &amp; Leucine (not resolved)</td>
<td>-8 -13 -1</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Hippuric acid</td>
<td>63 48 75</td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>5-Oxoproline (acetonitrile adduct)</td>
<td>8</td>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>9</td>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>4</td>
<td></td>
<td></td>
<td>B</td>
</tr>
<tr>
<td>Citric acid</td>
<td>-10 -18 6</td>
<td></td>
<td></td>
<td>A</td>
</tr>
</tbody>
</table>

\(^1\) Only identified metabolites are shown in Table 4. The column ‘response type’ refers to Figure 2; metabolites with response type A showed a treatment difference over the whole time course and metabolites with response type B showed a treatment difference only in the second part of the time course (a response of the metabolite in the placebo group, whereas no change in the diclofenac treated group). For metabolites with response type B the mean is calculated over the time points with a treatment difference.
Figure 2. OGTT time course mean metabolite response with standard error on day 9 corrected for concentrations on day 0 for subjects on placebo and diclofenac treatment. A) for the metabolite isoleucine that contributes to treatment differences over the whole time course and B) for the metabolite glycine that contributes to treatment differences only in the second part of the time course. Legend to Figure 2: Dashed line: Diclofenac treated subjects; Solid line: Placebo treated subjects.

Discussion

A primary goal of research into lifestyle associated diseases is to optimize health so that the onset of disease can be prevented or delayed. In identifying the key changes involved in the development of lifestyle associated diseases, experimental approaches have to deal with large inter-individual variety and the robustness of homeostasis. The current study deliberately recruited healthy overweight men with slightly increased inflammation parameters and successfully applied a relatively mild anti-inflammatory intervention so that only subtle changes were to be expected. The study authors aimed to demonstrate that an experimental design using metabolic profiling in concert with a challenge test is a good
strategy for the unraveling of biomarkers in intervention studies where only subtle changes are to be expected.

The conventional metabolic profiling approach of measuring blood samples in fasting conditions – even in a time course – and the classical biomarkers (i.e. glucose, insulin, sialic acid, HOMA index, and adiponectin; van Erk et al, in prep) were not able to reveal changes in response to diclofenac treatment. In this study, the subtle metabolic changes resulting from diclofenac treatment could only be determined using an OGTT time course. This can have at least two reasons. Firstly, most of the treatment differences became significant by repeated confirmation of subtle homeostatic alterations in metabolite concentrations (metabolites with response type A) without dealing with any day-to-day variations like in the 'long-term' fasted time course. Secondly, by perturbing a homeostatic metabolic situation, metabolite differences with a dynamic response to the oral glucose tolerance test became visible (response type B metabolites).

Diclofenac is known to inhibit and activate several enzymes and transporters (Boelsterli, 2003; Choi et al., 2005; Mano et al., 2005; Masubuchi et al., 2002; Peretz et al., 2005; Uchaipichat et al., 2004; Voilley et al., 2001; Wade et al., 1997; Ware et al., 1998). CD13 is a broad specificity aminopeptidase that cleaves specifically the N-terminal bound neutral amino acids from oligopeptides. The inhibition of the enzyme aminopeptidase N (CD13) by diclofenac corresponds to the lower plasma concentrations of several neutral amino acids such as L-isoleucine, L-threonine, and L-leucine. Such consistent lowering is reflected in a type A response.

In the current study, the diclofenac intervention applied was successfully shown by significantly reduced concentrations of PGE2 (see Materials and Methods). Metabolic profiling revealed that the diclofenac treatment resulted in lower plasma levels of uric acid. Elevated serum uric acid levels are positively associated with metabolic syndrome, insulin resistance and diabetes type II (Becker and Jolly, 2006; Hayden and Tyagi, 2004) and has been proposed as risk factor for hypertension and cardiovascular diseases (Baker et al., 2005; Johnson et al., 2005). Subsequently, elevated levels of uric acid are associated with inflammation and oxidative stress (Becker and Jolly, 2006; Hayden and Tyagi, 2004; Nakagawa et al., 2006). The current results suggest that inhibition of cyclooxygenase mediated inflammation (shown by significantly reduced concentrations of PGE2) could be associated with reduced concentrations of uric acid and therefore might lead to a reduction of risk on several metabolic diseases. However, this needs to be further explored.

Most interestingly in this study are the metabolites that showed a differential response between the treatments groups to the OGTT (metabolites identified with a response type B). All metabolites with response type B showed the largest difference between the treatments at time point 90 and/or 120 minutes after intake of glucose. Insulin peaks at an average of 65 minutes after glucose intake (Figure 3). This suggests that differences in response to the OGTT may be attributed to the action of insulin.
Of the 26 peaks that were most discriminative in the nPLS-DA models (Table 3), five peaks were found with this response type B profile of which two were annotated as 5-oxoproline and the others as the amino acids glycine, aspartic acid and glutamic acid. Three of these - 5-oxoproline, glycine and glutamic acid - are known to be involved in the glutathione synthesis pathway (Figure 4). Therefore, the response of other intermediates in the glutathione synthesis pathway was also studied. It appeared that glutathione and cysteinylglycine showed a similar dynamic response to the OGTT as the other type B responders (Figure 5), with the exception of plasma cysteine. Figure 4 provides an overview of the glutathione synthesis pathway and its relationship to glucose and insulin.

Higher clearance of plasma L-5-oxoproline is known in case of lower GSH synthesis (Yu et al., 2002). GSH synthesis is predominantly regulated by activity of γ-glutamylcysteine synthetase (GCS) and availability of the rate-limiting substrate cysteine (Townsend et al., 2003). Interestingly, it is known that insulin action increases and glucose decreases the regulation of GSH synthesis by GCS (Townsend et al., 2003; Lu, 2000; Wu et al., 2004).

In this study, the average plasma glutathione concentrations declined to their minimum concentration at 60 minutes after glucose intake and increased again at 90 to 120 minutes after glucose intake (Figure 6). In the current study, the control group showed significant lower concentrations of glutathione synthesis pathway intermediates at 90 to 120 minutes after glucose intake compared to diclofenac treated subjects. This might indicate that diclofenac treatment resulted in a higher GSH synthesis response after the glucose bolus, which might be related to altered insulin signaling with diclofenac treatment. It has been shown earlier that a selective inhibition of cyclooxygenase-2 results in increased insulin sensitivity in overweight or obese subjects (Gonzalez et al., 2005). However, no differences were found in classical insulin sensitivity indexes (HOMA index, ISIcomp, MCRest and Gutt-index) between diclofenac and placebo treated subjects in this study. A possible explanation
is that the combination of multiple metabolites as biomarker in concert with an oral glucose tolerance test allows for an earlier detection of changes in insulin sensitivity, however this is speculation at this stage and should be further explored.

Figure 4. Glutathione synthesis pathway and its connection to glucose and insulin. High levels of glucose inhibit and high levels of insulin activate glutathione synthesis via the enzyme γ-glutamylcysteine synthetase. Legend to Figure 4: Connection arrows with color red represent inhibition and color green represent activation. Purple hexagons represent metabolites; purple hexagons with white star represent metabolites measured with one of the metabolic profiling platforms; orange symbols represent enzymes. Red arrows upwards indicate that higher plasma concentration levels were found in the diclofenac treated group in response to oral glucose tolerance test. Abbreviations: AA, amino acid; Cys-Gly, cysteinylglycine. These figures were created by using MapEditor version 2.1.0 (GeneGo Inc, St Joseph, MI).
Figure 5. Dynamic response of glutathione synthesis pathway intermediates in OGTT time course. A) glutathione mean response with standard error on day 9 corrected for concentrations on day 0 for subjects on placebo and diclofenac treatment. Glutathione showed a treatment difference only in the second part of the time course. B) Cysteinylglycine mean response with standard error on day 9 corrected for concentrations on day 0 for subjects on placebo and diclofenac treatment. Cysteinylglycine showed a treatment difference only in the second part of the time course. Legend to Figure 5: Dashed line: Diclofenac treated subjects; Solid line: Placebo treated subjects.
Day 0 glutathione mean response with standard error is shown in OGTT time course for all subjects. Glutathione showed declined concentrations in the first part of the time course with a minimum concentration at 60 minutes after glucose intake. In the second part of the time course concentrations increase again.

This first exploratory study shows that subtle metabolic changes resulting from an anti-inflammatory treatment could only be determined using a metabolic perturbation test in a well-designed clinical study using metabolic profiling analysis. Differences in dynamic response to the challenge (response type B metabolites) might be derived from insulin regulated processes such as the insulin regulated glutathione synthesis pathway. Our study demonstrates that the use of metabolic profiling in concert with a challenge test may open new avenues for biomarker discovery that could be useful in developing preventive strategies for lifestyle associated diseases.

Acknowledgements

The authors express their gratitude to the volunteers that participated in the study and to the people from the Metabolic Research Unit and laboratories that assisted in the organization of the study.

The authors are member of the European Nutrigenomics Organisation (www.nugo.org) and the Dutch Nutrigenomics Consortium (http://www.nutrigenomicsconsortium.nl).

References


Joliffe IT (1986) Principal Component Analysis. Springer Verlag: New York, USA


Supplement Table S1

Overview of metabolites analyzed with the 4 different metabolic profiling platforms:

<table>
<thead>
<tr>
<th>GCMS global</th>
<th>LCMS polar</th>
<th>LCMS lipids</th>
<th>LCMS Free Fatty Acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbreviations: LPC, lysophosphatidylcholine; PC, phosphatidylcholine; SPM, sphingomyeline; ChE, cholesterolester; TG, triglyceride; FA, fatty acid.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>Metabolite ID</th>
<th>HMDB ID</th>
<th>Metabolite</th>
<th>Metabolite ID</th>
<th>HMDB ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lactic acid</td>
<td>HMDB00190</td>
<td>1</td>
<td>Leucine</td>
<td>HMDB00687</td>
<td></td>
</tr>
<tr>
<td>Alanine</td>
<td>HMDB00164</td>
<td>2</td>
<td>Unknown 65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pyruvic acid</td>
<td>HMDB00243</td>
<td>3</td>
<td>Citrulline</td>
<td>HMDB00904</td>
<td></td>
</tr>
<tr>
<td>Benzoylglycerol</td>
<td>HMDB0112</td>
<td>4</td>
<td>Unknown 66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3-Hydroxybutanoic acid</td>
<td>HMDB00357</td>
<td>5</td>
<td>Cysteinylglycine</td>
<td>HMDB00078</td>
<td></td>
</tr>
<tr>
<td>Phosphoric acid monomethyl ester</td>
<td>HMDB00378</td>
<td>6</td>
<td>Hippuric acid</td>
<td>HMDB00774</td>
<td></td>
</tr>
<tr>
<td>Unknown 7</td>
<td></td>
<td></td>
<td>Tyrosine</td>
<td>HMDB00158</td>
<td></td>
</tr>
<tr>
<td>Valine</td>
<td>HMDB00088</td>
<td>7</td>
<td>Aspartic acid</td>
<td>HMDB00483</td>
<td></td>
</tr>
<tr>
<td>4-Methyl-2-oxovaleric acid</td>
<td>HMDB00149</td>
<td>8</td>
<td>unknown 67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leucine</td>
<td>HMDB00687</td>
<td>9</td>
<td>unknown 68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycerol</td>
<td>HMDB00139</td>
<td>10</td>
<td>unknown 70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phospholipid</td>
<td>HMDB00142</td>
<td>11</td>
<td>Glucose</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proline</td>
<td>HMDB00162</td>
<td>12</td>
<td>Glutamic acid</td>
<td>HMDB00148</td>
<td></td>
</tr>
<tr>
<td>Isoleucine</td>
<td>HMDB00172</td>
<td>13</td>
<td>Tryptophan</td>
<td>HMDB00929</td>
<td></td>
</tr>
<tr>
<td>Glucose</td>
<td>HMDB00123</td>
<td>14</td>
<td>unknown 72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lactic acid</td>
<td>HMDB00190</td>
<td>15</td>
<td>unknown 73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-Aminocyclopropanecarboxylic acid</td>
<td>HMDB00254</td>
<td>16</td>
<td>unknown 74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lactic acid</td>
<td>HMDB00190</td>
<td>17</td>
<td>unknown 75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-Aminoadipate</td>
<td>HMDB00254</td>
<td>18</td>
<td>unknown 76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycerol</td>
<td>HMDB00139</td>
<td>19</td>
<td>unknown 77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Citrulline</td>
<td>HMDB00357</td>
<td>20</td>
<td>unknown 78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-Aminoadipate</td>
<td>HMDB00357</td>
<td>21</td>
<td>unknown 79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isoleucine</td>
<td>HMDB00172</td>
<td>22</td>
<td>unknown 80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>HMDB00148</td>
<td>23</td>
<td>unknown 81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glucose</td>
<td>HMDB00123</td>
<td>24</td>
<td>unknown 82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-Aminocyclopropanecarboxylic acid</td>
<td>HMDB00254</td>
<td>25</td>
<td>unknown 83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glucose</td>
<td>HMDB00123</td>
<td>26</td>
<td>unknown 84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lactic acid</td>
<td>HMDB00190</td>
<td>27</td>
<td>unknown 85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-Aminoadipate</td>
<td>HMDB00254</td>
<td>28</td>
<td>unknown 86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-Aminocyclopropanecarboxylic acid</td>
<td>HMDB00254</td>
<td>29</td>
<td>unknown 87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glucose</td>
<td>HMDB00123</td>
<td>30</td>
<td>unknown 88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolite ID</td>
<td>HMDB ID</td>
<td>Metabolite ID</td>
<td>HMDB ID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>---------------</td>
<td>--------------</td>
<td>---------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>202</td>
<td>glucose and/or other hexose HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td>203</td>
<td>glucose and/or other hexose HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td>204</td>
<td>glucose and/or other hexose HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>HMDB01147</td>
<td>205</td>
<td>glucose and/or other hexose HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td>206</td>
<td>unknown 83 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>HMDB00156</td>
<td>207</td>
<td>Citrate + Na HMDB030094</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td></td>
<td>208</td>
<td>Citrate + Na HMDB030094</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>HMDB00696</td>
<td>209</td>
<td>unknown 84 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>HMDB02142</td>
<td>210</td>
<td>unknown 85 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td>211</td>
<td>unknown 86 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>HMDB00191</td>
<td>212</td>
<td>unknown 87 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>HMDB00729</td>
<td>213</td>
<td>unknown 88 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>HMDB00562</td>
<td>214</td>
<td>unknown 89 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>HMDB00254</td>
<td>215</td>
<td>unknown 90 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>HMDB00413</td>
<td>216</td>
<td>unknown 91 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>HMDB00414</td>
<td>217</td>
<td>unknown 92 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td></td>
<td>218</td>
<td>unknown 93 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td></td>
<td>219</td>
<td>unknown 94 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>HMDB00159</td>
<td>220</td>
<td>unknown 95 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>HMDB00329</td>
<td>221</td>
<td>unknown 96 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>HMDB00208</td>
<td>222</td>
<td>unknown 97 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td></td>
<td>223</td>
<td>unknown 98 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td></td>
<td>224</td>
<td>unknown 99 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>HMDB00518</td>
<td>225</td>
<td>unknown 100 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td></td>
<td>226</td>
<td>unknown 101 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td></td>
<td>227</td>
<td>unknown 102 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td></td>
<td>228</td>
<td>unknown 103 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td></td>
<td>229</td>
<td>unknown 104 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>HMDB00510</td>
<td>230</td>
<td>unknown 105 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
<td>231</td>
<td>Ni-acetylglucosamine HMDB00210</td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>HMDB00098</td>
<td>232</td>
<td>unknown 106 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>HMDB00666</td>
<td>233</td>
<td>unknown 107 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>HMDB00517</td>
<td>234</td>
<td>unknown 108 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>HMDB00283</td>
<td>235</td>
<td>unknown 109 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td></td>
<td>236</td>
<td>unknown 110 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>66</td>
<td></td>
<td>237</td>
<td>unknown 111 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>HMDB00723</td>
<td>238</td>
<td>unknown 112 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>HMDB00142</td>
<td>239</td>
<td>unknown 113 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>HMDB00157</td>
<td>240</td>
<td>unknown 114 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>HMDB00479</td>
<td>241</td>
<td>Glutathione HMDB00125</td>
<td></td>
<td></td>
</tr>
<tr>
<td>71</td>
<td>HMDB00214</td>
<td>242</td>
<td>unknown 115 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>HMDB00094</td>
<td>243</td>
<td>Ni-acetylneuraminic HMDB00230</td>
<td></td>
<td></td>
</tr>
<tr>
<td>73</td>
<td></td>
<td>244</td>
<td>unknown 116 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>74</td>
<td></td>
<td>245</td>
<td>Tryptophan + glucose (this is a by-product) HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>HMDB00177</td>
<td>246</td>
<td>unknown 117 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>76</td>
<td>HMDB00182</td>
<td>247</td>
<td>unknown 118 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>HMDB00118</td>
<td>248</td>
<td>unknown 119 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>HMDB00460</td>
<td>249</td>
<td>unknown 120 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>HMDB00169</td>
<td>250</td>
<td>unknown 121 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>HMDB00122</td>
<td>251</td>
<td>unknown 122 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>HMDB00169</td>
<td>252</td>
<td>unknown 123 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>82</td>
<td></td>
<td>253</td>
<td>unknown 124 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>HMDB00122</td>
<td>254</td>
<td>unknown 125 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>84</td>
<td></td>
<td>255</td>
<td>unknown 126 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>HMDB00229</td>
<td>256</td>
<td>unknown 127 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>86</td>
<td>HMDB00226</td>
<td>257</td>
<td>unknown 128 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>87</td>
<td></td>
<td>258</td>
<td>unknown 129 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>88</td>
<td></td>
<td>259</td>
<td>unknown 130 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>89</td>
<td>HMDB00131</td>
<td>260</td>
<td>unknown 131 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90</td>
<td></td>
<td>261</td>
<td>unknown 132 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>91</td>
<td></td>
<td>262</td>
<td>unknown 133 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>92</td>
<td>HMDB00211</td>
<td>263</td>
<td>unknown 134 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>93</td>
<td>HMDB00289</td>
<td>264</td>
<td>unknown 135 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>94</td>
<td>HMDB00673</td>
<td>265</td>
<td>unknown 136 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>HMDB00207</td>
<td>266</td>
<td>unknown 137 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>HMDB00228</td>
<td>267</td>
<td>unknown 138 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>97</td>
<td>HMDB00929</td>
<td>268</td>
<td>unknown 139 HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>98</td>
<td>HMDB00827</td>
<td>269</td>
<td>C16:0_LPC HMDB00122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolite ID</td>
<td>Metabolite ID</td>
<td>Metabolite ID</td>
<td>Metabolite ID</td>
<td>Metabolite ID</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>100</td>
<td>101</td>
<td>102</td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>unknown 30</td>
<td>unknown 31</td>
<td>unknown 32</td>
<td>unknown uridine</td>
<td>unknown 33</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HMDB00767</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>104</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>unknown 34</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>unknown 35</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>unknown 36</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C16:0 Glycerol ester</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C18:0 Glycerol ester</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>C18:1 Glycerol ester</td>
<td></td>
</tr>
<tr>
<td>111</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maltose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>112</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unknown 37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>113</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unknown 38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>114</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vitamin E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>115</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cholesterol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>116</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethyldrocholesterol</td>
<td>HMDB00908</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>117</td>
<td></td>
</tr>
<tr>
<td>unknown 39</td>
<td>118</td>
<td>unknown 40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>119</td>
<td>unknown 41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>120</td>
<td>unknown 42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>121</td>
<td>unknown 43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>122</td>
<td>unknown 44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>123</td>
<td>unknown 45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>124</td>
<td>unknown 46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>unknown 47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>126</td>
<td>C18:0 Sphingomyelaine</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>127</td>
<td>unknown 48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>128</td>
<td>unknown 49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>129</td>
<td>unknown 50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>130</td>
<td>unknown 51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>131</td>
<td>unknown 52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>132</td>
<td>unknown 53</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>133</td>
<td>unknown 54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>134</td>
<td>unknown 55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>135</td>
<td>1,2-diacylglyceride (C16:2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>136</td>
<td>unknown 56</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>137</td>
<td>1,3-diacylglyceride</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>138</td>
<td>unknown 57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>139</td>
<td>unknown 58</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>140</td>
<td>Isoleucine+Leucine (not resolved)</td>
<td>HMDB00172 or HMDB00687</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycine</td>
<td>HMDB00129</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alanine</td>
<td>HMDB00161</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-Aminobutanoic acid</td>
<td>HMDB00112</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serine</td>
<td>HMDB00187</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unknown 59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proline</td>
<td>HMDB00163</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-unamo-2-methyl butanoic acid</td>
<td>HMDB01109</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valine</td>
<td>HMDB00883</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threonine</td>
<td>HMDB00167</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cysteine</td>
<td>HMDB00574</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>151</td>
<td>unknown 60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>152</td>
<td>unknown 61</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>153</td>
<td>2-Aminocyclohexanecarboxylic acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>154</td>
<td>1-Oxo-proline</td>
<td>HMDB00267</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>155</td>
<td>2-Aminocyclohexanecarboxylic acid</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>156</td>
<td>4-Hydroxyproline</td>
<td>HMDB00725</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>157</td>
<td>Creatine</td>
<td>HMDB00054</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>158</td>
<td>Leucine &amp; Isoleucine (not resolved)</td>
<td>HMDB00172 or HMDB00687</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homocysteine</td>
<td>HMDB00042</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>unknown 62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allylamine</td>
<td>HMDB01123</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glutamate</td>
<td>HMDB01339</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tryptophan</td>
<td>HMDB00929</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methionine</td>
<td>HMDB00646</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>164</td>
<td>unknown 63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolite</td>
<td>Metabolite ID</td>
<td>HMDB ID</td>
<td>Metabolite</td>
<td>Metabolite ID</td>
<td>HMDB ID</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------</td>
<td>---------</td>
<td>------------------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>166</td>
<td>unknown 64</td>
<td></td>
<td>338</td>
<td>C20:1_FA</td>
<td></td>
</tr>
<tr>
<td>167</td>
<td>Carnitine</td>
<td>HMDB000562</td>
<td>339</td>
<td>C20:2_FA</td>
<td></td>
</tr>
<tr>
<td>168</td>
<td>Phenylalanine</td>
<td>HMDB001159</td>
<td>340</td>
<td>C20:3_FA</td>
<td></td>
</tr>
<tr>
<td>169</td>
<td>N-methylhistidine</td>
<td>HMDB00479</td>
<td>341</td>
<td>C20:4_FA</td>
<td></td>
</tr>
<tr>
<td>170</td>
<td>5-Oxoproline</td>
<td>HMDB00247</td>
<td>342</td>
<td>C22:5_FA</td>
<td></td>
</tr>
<tr>
<td>171</td>
<td>Phenylalanine</td>
<td>HMDB00119</td>
<td>343</td>
<td>C22:6_FA</td>
<td></td>
</tr>
</tbody>
</table>
ASSESSING THE PERFORMANCE OF STATISTICAL VALIDATION TOOLS
FOR MEGAVARIATE METABOLICOMICS DATA

Carina M Rubingh, Sabina Bijlsma, Eduard PPA Derks, Ivana Bobeldijk,
Elwin R Verheij, Sunil Kochhar¹, Age K Smilde

TNO Quality of Life, Business Unit Analytical Sciences, P.O. Box 360, 3700 AJ Zeist, The Netherlands
¹Nestlé Research Center, BioAnalytical Science Department, PO Box 44, CH-1000 Lausanne 26, Switzerland

Metabolomics 2006; 2. 53-61.
Abstract

Statistical model validation tools such as cross-validation, jack-knifing model parameters and permutation tests are meant to obtain an objective assessment of the performance and stability of a statistical model. However, little is known about the performance of these tools for megavariate data sets, having, for instance, a number of variables larger than 10 times the number of subjects. The performance is assessed for megavariate metabolomics data, but the conclusions also carry over to proteomics, transcriptomics and many other research areas.

Partial least squares discriminant analyses models were built for several LC-MS lipidomic training data sets of various numbers of lean and obese subjects. The training data sets were compared on their modeling performance and their predictability using a 10-fold cross-validation, a permutation test, and test data sets. A wide range of cross-validation error rates was found (from 7.5% to 16.3% for the largest trainings set and from 0% to 60% for the smallest training set) and the error rate increased when the number of subjects decreased. The test error rates varied from 5% to 50%.

The smaller the number of subjects compared to the number of variables, the less the outcome of validation tools such as cross-validation, jack-knifing model parameters and permutation tests can be trusted. The result depends crucially on the specific sample of subjects that is used for modeling. The validation tools cannot be used as warning mechanism for problems due to sample size or to representativity of the sampling.
Introduction

Metabolomics studies are performed to investigate responses of biologic systems on environmental influences due to, for instance, toxicological exposure, nutrition or medical treatment. In this field, metabolites in biological samples like plasma or urine are analytically determined using techniques such as nuclear magnetic resonance (NMR; Derome, 1987), liquid or gas chromatography mass spectrometry (LC-MS and GC-MS, respectively; Wilson et al., 2005; Lenz et al., 2004; Lafaye et al., 2003; Plumb et al., 2003; Van der Greef et al., 2003; Fiehn, 2002). These analytical techniques can generate a large amount of data containing information about a large number of correlated variables, which asks for appropriate statistical tools for data analysis. Multivariate data analysis (MVA) is used to analyze the correlated data. MVA can be used to summarize the data by reducing the dimensions of the data, for regularization purposes, for variable selection, etcetera. One of the applications of MVA is to use correlations and trends in the data in order to discriminate between groups (Massart et al., 1997; Vandeginste et al., 1998).

Discriminant analysis (DA) is a MVA method that can be used if the interest is focused on differences between groups of objects or on subgroup structures and can serve two slightly different purposes. If it is used to separate distinct sets of objects or observations, discrimination is the main purpose. If it is used to define classification rules to allocate new objects or observations to previously defined groups, it is used for classification (Vandeginste et al., 1998).

However, the results found in DA can not always be trusted as they are sensitive to chance-correlations and/or to the risk of overfitting. Validation tools like cross-validation (Stone, 1974; Hastie et al., 2001; Martens and Naes, 1989), permutation tests (Mielke and Berry, 2001; Good, 2000; Manly, 1997; Efron and Tibshirani, 1993), jack-knifing model parameters (Efron, 1982; Martens and Martens, 2002) and test data sets are used to address these problems and provide an objective assessment of the performance and stability of a model. These tools are commonly used to validate the results of multivariate data analyses. When multivariate data become megavariate data, the number of variables is even larger and, due to the curse of dimensionality, the chance of false correlations and the risk of overfit is even higher. In the present study, a data set having a number of variables larger than 10 times the number of subjects is defined to be megavariate. The validity of cross-validation for small-sample classification was assessed under low dimensionality (Martens and Dardenne, 1998; Braga-Neto and Dougherty, 2004), but still little is known about how validation tools such as cross-validation, jack-knifing and permutation tests will perform for megavariate data.

Cross-validation is used to choose the optimal model parameters as well as to test the predictability of the statistical model. Cross-validation uses the available data minus a particular part (e.g., 1/k-th part of the total data set) to fit the model and the part that was left out to test the model (Hastie et al., 2001). However, the predictability based on a single cross-validation is biased and often too optimistic because the determination of the model...
meta parameters (e.g. number of latent variables or any regularization term) is based on the
same set as is used to determine the predictability. Hence, still a separate test set is required
to determine the predictability for future data. This problem can be addressed by double
cross-validation, which makes most efficient use of the available data as all objects are used
for model building and validation (Stone, 1974).

The stability of model parameters is assessed by the jack-knife procedure. All available data
minus the data of one (or more) objects is used to fit the model and for each perturbed set,
the parameters estimates can be obtained. A graphical presentation or an evaluation of
(relative) standard deviations of the estimates gives an impression of the stability of the
estimates (Efron, 1998; Martens and Martens, 2000).

A permutation test is used to assess the significance of a classification. The class assignment
can be permuted several times and for each permutation, a model between the data and the
permutated class-assignment can be built. The discrimination between classes of the model
based on the permutated class-assignment is compared to the discrimination of the model
based on the original classification (Mielke and Berry, 2001; Good, 2000; Manly, 1997; Efron
and Tibshirani, 1993).

The classification of a test data set using the model-parameters based on the training data
set, provides information about the generalizability of a model; whether the model is only
applicable for the subjects in the training set or whether it can also be used to predict the
classification of new subjects. All these tools can be used to prevent that conclusions about
the discrimination between classes may be drawn, which can not be statistically supported.

In order to assess the performance of statistical validation tools for megavariate data sets,
several data sets of various sizes, all derived from the same original data set of human LC-MS
lipidomic data, are compared on their modeling performance and their predictability. These
data were obtained from a co-operative metabolomics study of TNO, Nestlé Research Centre
(Lausanne, Switzerland) and the EU NUGENOB project (NUGENOB is the acronym of the
project 'Nutrient-Gene interactions in human obesity – implications for dietary guidelines'
supported by the European Community (Contract no. QLK1-CT-2000-00618), see the web-
site www.nugenob.com; Blaak et al., 2006; Petersen et al., 2005). The main objective of this
metabolomics study was to find biomarkers that characterize differences between high and
low fat burners in lean and obese subjects. A strategy for data preprocessing, data analysis
and validation of statistical models was also developed (Bijlsma et al., 2006). The present
study was performed in order to investigate the effect of decreasing the number of subjects
on the performance of the statistical validation tools. Although metabolomics data were
used for the analyses, the issue also carries over to proteomics and transcriptomics data.
Materials and Methods

Data

General Although real-life data may lead to less distinguishing differences between sets, it was preferred above simulated data because it illustrates the problems researchers have to deal with best. Both biological and analytical variation are present in the data and may be of influence on the results. Data of a co-operative metabolomics study of TNO, Nestlé Research Centre (Lausanne, Switzerland) and the EU NUGENOB project were used. This study involved plasma from 50 lean and 100 obese human subjects, collected at four different time points (t=0, 1, 2, and 3 hours) after a single intake of a fat rich meal. All samples were analysed using four analytical platforms: NMR, GC-MS, LC-MS polar and LC-MS lipid. Details about the study design can be found in Petersen et al. (2005), whereas details about the data and data preprocessing can be found in Bijlsma et al. (2006). The data set used in the present study was based on the LC-MS lipid data measured at t=0 hours, which contained 947 LC-MS peaks (variables).

Base data set The focus was on differences between lean and obese subjects in the LC-MS lipid. In order to avoid confounding of the results due to an unbalanced number of lean and obese subjects, a random selection of 50 out of the 100 obese subjects was made. The created data set of 50 obese and 50 lean subjects was used as base data set (data50:50). Subsets of the data50:50 were used to study the effect of the decrease of the number of subjects on the analysis and validation results.

Data subsets A data set was generated containing the data of 40 lean and 40 obese subjects (data40:40). The inclusion of a subject into the data40:40 data set was based on random selection without replacement. The creation of the data40:40 set was repeated 10 times, each based on a new random selection of the original data50:50 set (base data set). This process was repeated for 10 sets of 30 lean and 30 obese subjects (data30:30), for 10 sets of 20 lean and 20 obese subjects (data20:20), for 10 sets of 10 lean and 10 obese subjects (data10:10) and finally, for 10 sets of 5 lean and 5 obese subjects (data05:05), all based on random selection out of the data50:50 base data set. Although additional information about the subjects such as being a high or low fat burner or the center at which the sample was collected (Petersen et al., 2005), was not used in the statistical analysis, equal representation of these factors over the created subsets was secured.

The subset data sets were used for modeling and were used as so called training data sets. The data of the remaining subjects were used as test data sets. The test data set of the data40:40 set contained the data of the remaining 10 lean and 10 obese subjects, the test data set of the data30:30 set contained the data of the remaining 20 lean and 20 obese subjects, the test data set of the data20:20 set contained the data of the remaining 30 lean and 30 obese subjects, the test data set of the data10:10 set contained the data of the remaining 40 lean and 40 obese subjects, and the test data set of the data05:05 set
contained the data of the remaining 45 lean and 45 obese subjects. As a consequence of this procedure, the size of the test sets differ. To rule out the possible effect of the test data set size, an extra test set, based on a random selection without replacement of the data of 10 obese and 10 lean subjects, was also created for each subset. Summarizing, one base data set was generated as well as 50 training sets (5x10) and 50 test sets (5x10) and 50 extra test sets (5x10). The procedure that was followed to obtain all data sets, is illustrated in Figure 1. This procedure was chosen to mimic reality, in which very few samples are available for data analysis. The data may be unrealistically small for human studies, but was incorporated for illustrative purposes.

![Figure 1. Illustration of the procedure that was followed to obtain the data sets.](Image)

**Statistical Analysis**

Partial least squares discriminant analysis (PLS-DA; Barker and Rayens, 2003; Vong et al., 1988) was used to find a small number of linear combinations of the original variables (called 'latent variables'; LVs), that was predictive for the class membership and that described most of the variability of the LC-MS metabolic profiles. PLS-DA is a linear regression method whereby the multivariate variables (the X-block) corresponding to the observations are related to the class membership (the Y-Block) for each subject. The Y-block contains "1" and "0" only, corresponding to the lean and obese class assignment. It is a classical PLS regression (Martens and Naes, 1989; VandeGinste et al., 1998; Massart et al., 1997; Geladi and Kowalski, 1986) where the response is a categorical one expressing the class membership of a subject. PLS-DA will maximise the covariance between the predicting data
set (X block with LC-MS metabolomic profiles) and the data to be predicted (Y-block with class assignments).

Data were mean-centered before analyses. The center-parameters of the training set were used to transform the corresponding test data set. Details about other aspects of the data pre-processing can be found in Bijlsma et al. (2006). All analyses were performed using Matlab Version 7.0.4 R14 (The Mathworks, Inc.) and the PLS Toolbox Version 3.0.4 (Eigenvector Research, Inc.).

**Statistical Model Validation**

**Cross-validation** The use of a double cross-validation would be preferred (Stone, 1974), because a single cross-validation may lead to bias and overestimation of the true error rate (Hastie et al., 2001). However, the issue of bias is in this case of less importance, because only the error rates are compared and it is assumed that the bias in each model is similar. For this reason and for the fact that in case of very small data sets a double cross validation becomes less appropriate, a single cross-validation was used instead of a double cross-validation. A single 10-fold venetian blind cross-validation based on stratified sampling having the lean and obese class membership as strata, was used to choose the optimal number of LVs as well as to obtain an estimate of the error rate of the PLS-DA model. In the first cross-validation step, 1/10-th of a training data set was left out, under the restriction that the number of lean subjects that was left out was equal to the number of obese subjects that was left out, and data of the remaining subjects were used to build a PLS-DA model. The model was used to predict the class assignment of the “left out” subjects. This was repeated until all subjects were left out once. The number of LVs yielding the lowest percentage of misclassifications (error rate) was chosen as the optimal model. Note that by using a 10-fold cross-validation for data05:05, only 1 subject is left out at each step of the cross-validation. Hence in this case, the 10-fold cross-validation is equal to a "leave-one-out" cross-validation.

**Jack-knife** The stability of the regression coefficients of the PLS-DA models was assessed by jack-knifing (Efron, 1998; Martens and Martens, 2000). In order to be able to use the same data set parts as was used in cross-validation, all available data minus 1/10-th was used to fit the model, instead of leaving-out-one observation per jack-knife step which is a more usual way of jack-knifing. In the first jack-knife step, 1/10-th of a training data set was left out, under the restriction that the number of lean subjects that was left out was equal to the number of obese subjects that was left out, and data of the remaining subjects were used to build a PLS-DA model. This was repeated until all subjects were left out once. The 10 variables having the largest coefficient in the reference model data50:50 were evaluated graphically using Box-and-Whisker-plots.
**Permutation test** Cross-validation can be used to assess the class-predictability of a model. In order to assess the discrimination, an exact or an approximate permutation test can be used (Mielke and Berry, 2001; Good, 2000; Manly, 1997; Efron and Tibshirani, 1993).

The class assignment was permuted in such a way that the ratio between the number of lean (“0”) and obese (“1”) subjects remained equal, and this was done 1000 times with replacement of the class vector. As an exact permutation test would lead to too many combinations, an approximate permutation test was performed on each of the data sets. For the data05:05 subset, only 100 permutations of the Y-block were performed, because the number of possible permutations is much lower than 1000. For each permutation, a PLS-DA model was built between the X-block and the permuted Y-block using the same optimal number of LVs as determined by cross-validation for the model based on the original class assignment. The ratio of the between sum of squares and the within sum of squares (B/W-ratio) for the class assignment prediction of each model was calculated. If the B/W-ratio of the original class assignment is a part of the distribution based on the permuted class assignments, the contrast between the two classes cannot be considered to be significantly different from a statistical point of view. If, on the other hand, the B/W-ratio based on the original class assignment is much higher compared to the ratios based on the permuted class assignments, the differences between the classes are statistically significant. Because exact accuracy percentages are not important in the scope of this paper, the permutation test is evaluated visually according to Figure 2 (Bijlsma et al., 2006).

![Distribution of random class assignments](image)

**Figure 2.** Visual evaluation of the permutation test.

**Predictability** Cross-validation, jack-knifing and the permutation test provide information about the validity of the model based on the information in the training data set. The generalizability suggested by the cross-validation error rate was assessed by the prediction of the class assignment of new subjects, which are in this case defined as the subjects in the test data sets. The class assignment prediction of the subjects in these test data sets was determined based on the model parameters of the corresponding training data set. Hence,
the prediction was based on the same (number of) LVs as was used for the training set. The error rate of the test data set, being the percentage of misclassified subjects, was calculated and was used as measure for the generalizability of the model. Ideally, the test error rates are comparable to the ones found by 10-fold cross-validation.

Results and Discussion

Training sets The results of the PLS-DA model for the data50:50 are presented in Figure 3. As this model is based on all lean and obese subjects, this model is considered to be the reference model. For data50:50, the cross-validation error rate of the model is 11% (0.11 in Figure 3a) based on 11 LVs and is shown in Figure 3a. Figure 3b shows the prediction based on the cross-validation for the lean (first 50; marked as 'o') and the obese (second 50; marked as '*' subjects. The overlap between the two classes shown in this sub-figure corresponds to the error rate of 11%. In Figure 3c the final fit is shown, which is much more optimistic compared to the prediction based on cross-validation. Finally, in Figure 3d the jack-knife results for the 10 largest regression coefficients is given. The results in Figure 3 are similar to the results found by Bijlsma et al. (2006) in the analyses on the data set based on 100 obese and 50 lean subjects.

![Figure 3. PLS-DA results for data50:50: Cross-validation error rate (a), Prediction based on cross-validation (b; o = lean, * = obese), Prediction based on fit (c; o = lean, * = obese), and Jack-knife (d).](image-url)
A summary of the results of the PLS-DA models based on all training sets is given in Table 1. Per data set and per model, the error rate based on the 10-fold cross-validation, the number of used LVs and the evaluation of the permutation test are given. Also the mean and standard deviation of the error rate and the mean number of LVs per data set are presented.

The mean cross-validation error rate and the variance of the error rate both increase if the number of subjects in the data set decreases. The results of the analysis of the data05:05 sets are the most variable, showing a range of error rates from 0% for the 10th selection to 60% for the 4th selection. The results of the 4th and the 10th selection of data05:05 are presented in Figure 4 A and B. The jack-knife results confirm the above described discrepancy between the conclusions based on both sets of data05:05.

The 10 largest regression coefficients found in the reference model of data50:50 were considered to be the most important variables for the discrimination between the two groups and therefore, only these 10 were used to evaluate the jack-knife results. Needless to say, the absolute values presented in Figure 4 A and B are not comparable to the values presented in Figure 3. The coefficients of the 4th selection of data05:05 show a lot of variation and the coefficients of the 10th selection show only little variation but were almost all equal to zero. This finding confirms that it can be expected that both sets were not representative for the total set of 50 lean and 50 obese subjects.

**Table 1.** Summary of PLS-DA results based on all training sets (ER = cross validation error rate in %, LV = number of latent variables, P = evaluation permutation test with e = excellent, g = good, m = moderate, b = bad).

<table>
<thead>
<tr>
<th>Model</th>
<th>4040</th>
<th>3030</th>
<th>2020</th>
<th>1010</th>
<th>0505</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ER</td>
<td>LV</td>
<td>P</td>
<td>ER</td>
<td>LV</td>
</tr>
<tr>
<td>1</td>
<td>12.5</td>
<td>6</td>
<td>e</td>
<td>20.0</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>16.3</td>
<td>7</td>
<td>e</td>
<td>15.0</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>13.8</td>
<td>7</td>
<td>e</td>
<td>18.3</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>12.5</td>
<td>7</td>
<td>e</td>
<td>18.3</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>11.3</td>
<td>6</td>
<td>e</td>
<td>18.3</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>15.0</td>
<td>7</td>
<td>e</td>
<td>8.3</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>15.0</td>
<td>7</td>
<td>e</td>
<td>15.0</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>11.3</td>
<td>6</td>
<td>e</td>
<td>16.7</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>16.3</td>
<td>7</td>
<td>e</td>
<td>16.7</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>7.5</td>
<td>8</td>
<td>e</td>
<td>13.3</td>
<td>6</td>
</tr>
<tr>
<td>MEAN</td>
<td>13.1</td>
<td>7.0</td>
<td>16.0</td>
<td>7.0</td>
<td>18.5</td>
</tr>
<tr>
<td>Std</td>
<td>2.7</td>
<td>3.4</td>
<td>8.3</td>
<td>12.3</td>
<td>18.9</td>
</tr>
</tbody>
</table>

**Test sets** The test data sets were used to determine the generalizability of the models. The number of LVs was based on the number of LVs used for modeling the training data set. The mean and the standard deviation of the test error rate per data set are presented in Table 2 and reveal that the predictability of the models based on small training data sets was worse than the predictability of the models based on larger training data sets.
The test error rates varied from 5% to 30% for the test sets corresponding to data40:40 and from 35% to 50% for the test sets corresponding to data05:05. The mean test error rates in table 2 are similar to the cross-validation error rates of the corresponding training data sets in table 1, except for data10:10. The standard deviations of the test error rates are less variable compared to the cross-validation error rates of the training data sets presented in Table 1.

Although the 10th selection of data05:05 had a much better cross-validation error rate for the training set (0%) compared to the 4th selection (60%), their test error rate based on the corresponding test set is similar (both 50%). The results of the extra test data sets of 10 obese and 10 lean subjects are also presented in Table 2. Although the mean levels of the error rates are similar, the rates are more variable compared to the original test data sets, due to the smaller size of the extra test data sets.

Table 2. Summary of PLS-DA results based on the projection of all test data sets (number of LVs based on corresponding training data sets).

<table>
<thead>
<tr>
<th>Model</th>
<th>4040</th>
<th>3030</th>
<th>2020</th>
<th>1010</th>
<th>0505</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Size testset</td>
<td>Size testset</td>
<td>Size testset</td>
<td>Size testset</td>
<td>Size testset</td>
</tr>
<tr>
<td>1</td>
<td>15.0</td>
<td>15.0</td>
<td>35.0</td>
<td>35.0</td>
<td>40.0</td>
</tr>
<tr>
<td>2</td>
<td>30.0</td>
<td>30.0</td>
<td>27.5</td>
<td>25.0</td>
<td>41.7</td>
</tr>
<tr>
<td>3</td>
<td>30.0</td>
<td>30.0</td>
<td>42.5</td>
<td>40.0</td>
<td>21.3</td>
</tr>
<tr>
<td>4</td>
<td>30.0</td>
<td>30.0</td>
<td>17.5</td>
<td>10.0</td>
<td>13.3</td>
</tr>
<tr>
<td>5</td>
<td>15.0</td>
<td>15.0</td>
<td>22.5</td>
<td>30.0</td>
<td>48.3</td>
</tr>
<tr>
<td>6</td>
<td>20.0</td>
<td>20.0</td>
<td>17.5</td>
<td>10.0</td>
<td>33.3</td>
</tr>
<tr>
<td>7</td>
<td>35.0</td>
<td>35.0</td>
<td>35.0</td>
<td>35.0</td>
<td>50.0</td>
</tr>
<tr>
<td>8</td>
<td>25.0</td>
<td>25.0</td>
<td>35.0</td>
<td>40.0</td>
<td>30.0</td>
</tr>
</tbody>
</table>

Discussion
The results are predominantly driven by the size of the training data set and the selection of the subjects in that data set, which is especially illustrated by the smaller training data sets. The mean cross-validation error rate increases as the number of subjects in the training data set decreases. In itself this is not a spectacular finding. A model based on a larger training data set can be determined more precisely than a model based on a smaller data set. On the other hand, the larger the test data set, the more precise the mean test error rate can be estimated. Ideally, test error rates are of the same order as cross-validation error rates. The test set error rates and the cross-validation error rates were quite similar at a mean level, except for data10:10. However, at individual set level, the cross-validation error rate is in most cases not comparable to the test error rate. This illustrates that the result crucially depends on the specific sample of subjects that was used for modeling.
Figure 4. PLS-DA results for the 4th selection (A) and the 10th selection (B) of data05:05: Cross-validation error rate (a), Prediction based on cross-validation (b; o = lean, * = obese), Prediction based on fit (c; o = lean, * = obese), and Jack-knife (d).
With only a small selection from a total population it is more likely that the selected subjects are not representative for the studied population, because it is possible that only subjects out of the extremes of the population distribution are selected. This study shows that the selection of subjects is crucial for the conclusions that are drawn about the model.

The effect is best seen in the results of data05:05. The 10th selection of data05:05 had a much better cross-validation error rate for the training set compared to the 4th selection. If the 5 lean and 5 obese subjects of the 10th selection were selected as the representatives of the population under study, the conclusion would be that the 2 groups can be separated based on their LC-MS lipidomic profiles, even based on cross-validation results. If the 10 subjects out of set 4 were the subjects selected as the representatives of the population under study, the conclusion would be completely opposite. This means that the conclusions about the model completely depend on the selected 10 subjects. Nevertheless, the error rates based on the corresponding test sets were quite similar. As the predictability of both models was poor, it can be expected that both sets were not representative for the total set of 50 lean and 50 obese subjects. This illustrates how it could go wrong using data sets having considerably less subjects compared to the number of variables and it also shows the risk of drawing (too) optimistic conclusions about the distinction between the two classes, even based on cross-validation results. The size of the test data set did not seem to be an issue, as the results of the extra test data sets of 10 obese and 10 lean subjects were similar to the results based on the original test data sets.

Because different purposes are served, the conclusions about model validity based on cross-validation are not always comparable to the conclusions drawn based on the permutation test. The variation in performance of the permutation test was lower compared to the variation in error rates. The test only assesses the significance of the classification and does not take the predictability into account, which can explain why a model having a high cross-validation error rate can perform well in the permutation test.

All results indicate that cross-validation, jack-knifing and permutation tests are insufficient validation tools for megavarie data sets with only a few samples. The lower the ratio between the number of subjects and the number of variables, the less the validation results can be trusted. Taking only the results of these validation tools into account can be very misleading and may lead to incorrect conclusions. In order to avoid these problems, the number of samples per group should be large enough. In the present study, the turning point seemed to be between the sets having 10 and 20 subjects per group and based on about 950 variables. Unfortunately, it is impossible to translate this into a “golden rule” for all megavarie data sets.

Due to practical or budgetary limitations, it is often impossible to include the number of subjects that would be necessary to avoid the problems presented above. Another way to deal with megavarie data sets is to make the sets “less megavarie” by reducing the number of variables that are used for the statistical data analysis. This can be done, for
instance, based on i) analytical grounds by using a target approach instead of the total screening approach, ii) biological grounds by using a priori variable selection, iii) a selection method using statistical tools (Smilde et al., 2005), iv) grey models, in which prior knowledge about (groups of) variables is taken into account (Gurden et al., 2001; Bijlsma and Smilde, 2000) or v) regularization techniques, like using simplified correlation matrices (Schäfer and Strimmer, 2005). The disadvantage of the third and fifth approach is that the variables are selected using MVA methods which use the full data and similar problems as mentioned above can affect the selection. Using this approach, the bias due to selection should be assessed and corrections should be made (Ambroise and McLachlan, 2002). In case of a small number of subjects compared to the number of variables, contradictory results can be expected. Whether more simple statistical methods, e.g, those ignoring correlations like Nearest Shrunken Centroids (Tibshirani et al., 2003; Tibshirani et al., 2002), can be used to reduce the number of variables, is still under investigation.

The performance is assessed using this specific megavariate metabolomics data, but it is expected that the conclusions will also carry over to many other research areas. It was known that the data represented small differences between obese and lean subjects (Bijlsma et al., 2006). It is possible that the findings would be less dramatic if data that represents larger differences between groups is used.

The present study did not take the variable selection into account and only investigated the influence of the number of samples in the data sets. Future research may reveal the impact of the variable selection on the reliability of the standard statistical validation tools for megavariate data.

**Concluding Remarks**

The lower the number of subjects compared to the number of variables, the less the outcome of validation tools such as cross-validation, jack-knifing and permutation tests can be trusted. The result depends crucially on the specific sample of subjects that is used for modeling. The validation tools can not be used as warning mechanism for problems due to sample size or representativity issues.

**References**


THE COSTS OF COMPLEX MODEL OPTIMIZATION

Carina M Rubingh\(^1\), Harald Martens\(^2\), Hilko van der Voet\(^3\), Age K Smilde\(^4\)

\(^1\)TNO Quality of Life, P.O. Box 360, 3700 AJ Zeist, The Netherlands
\(^2\)NOFIMA MAT, Osloveien 1, NO-1430 Ås, Norway
\(^3\)Biometris, Wageningen University and Research Centre, P.O. Box 100, 6700 AC Wageningen, Netherlands
\(^4\)Swammerdam Institute for Life Sciences, University of Amsterdam, Nw Achtergracht 166, 1018 WV Amsterdam, The Netherlands

Submitted for publication
Abstract

Each data-driven action in data modeling consumes degrees of freedom, whether it concerns estimation of parameters, estimation of meta-parameters or selecting variables. By using a double cross validation approach for degrees of freedom calculation the costs for meta-parameter estimation and variable selection can determined explicitly. The only assumptions are independent and identically distributed errors, which makes the approach applicable to many predictive modeling techniques.
Summary

Each data-driven action in data modeling consumes degrees of freedom, whether it concerns estimation of parameters (such as regression coefficients), estimation of meta-parameters (such as optimal model rank) or selecting variables. By using a double cross validation approach for the degrees of freedom calculation, the costs for meta-parameter estimation and variable selection can be determined explicitly. To illustrate the concept, PLS regression was applied to two real-life metabolomics data sets. It is shown that only a small price needs to be paid to determine the meta-parameter (i.e. the optimal model complexity) whereas variable selection is much more ‘expensive’. Concerning a relatively small data set, almost 35% of the total degrees of freedom could be lost due to variable selection. Considering the fact that in functional genomics research the number of objects is mostly even smaller than the smallest subset that was used for the present paper, the impact of ignoring the degrees of freedom consumed by variable selection becomes very alarming.

Introduction

In the field of classical regression models, relating one property (regressand $y$) to a set of explanatory variables (regressors, $X$) in $n$ samples or objects, degrees of freedom is a very common and frequently used expression. But it is often difficult to understand, and even more difficult to keep track of quantitatively. Input data may have fewer degrees of freedom than expected, and data modeling may consume more degrees of freedom than is realised.

Conceptually, a degree of freedom corresponds to an independent way in which something can vary. Physically, an airplane has three degrees of freedom of forward motion through the 3D air space. Independently of that, it can rotate in different ways, which gives additional degrees of freedom for the pilot. In chemistry, a water molecule can likewise move in three directions, and rotate, and in addition it can vibrate in various ways, adding yet more degrees of freedom.

In statistics, the most basic definition of degrees of freedom is that of a parameter of the chi-square distribution, which is used to describe the distribution of quadratic forms under standard statistical assumptions: For a set of standard normal deviates, i.e. variables with independently, identically distributed elements, the distribution of their squared sum is defined as a chi-square distribution, and its degrees of freedom is equal to the number of summed squared standard normal deviates in the set. Another interpretation of degrees of freedom is the costs that are needed to be paid for parameter estimation. For instance, in full-rank multiple linear regression models, each of the $p$ predictor variable as well as the constant are said to use one degree of freedom leaving $n-p-1$ degrees of freedom for the residual variance of a model estimated from $n$ independent observations.. In model selection criteria, such as Akaike information criterion (AIC; Akaike, 1973) or Mallows’ $C_p$ (Mallows, 1973), the number of degrees of freedom is used as a model complexity measure.
The more complex the model, the more degrees of freedom are needed to estimate the parameters.

The above mentioned applications of degrees of freedom are straightforward for standard linear models, like full-rank ordinary least squares estimation in multiple regression models. The concept of degrees of freedom is in this case also intuitively understandable. Considering degrees of freedom as independent parts of information in a data set, for each data-driven ‘choice’ that is to be made concerning modeling, a price has to be paid in terms of losing a number of these independent parts. The more choices are made, the less independent information is available, thereby lowering the number of degrees of freedom left over.

Full-rank regression modeling requires independent regressors, and that is unnatural for observational data sets; they call for reduced-rank modeling. Partial Least Squares Regression (PLSR) (Wold et al. 1983, Geladi and Kowalski, 1986; Martens and Naes, 1989) is a reduced-rank regression method commonly used in chemometrics due to its versatility and graphical accessibility. However, for PLSR, no formal definition of degrees of freedom exists. But the intuitive concept as stated above still holds. PLSR is an example of a complex modeling technique, for which the price for modeling in terms of degrees of freedom cannot easily be determined as in linear models. Although PLSR is often referred to as linear model due to its linear form of the final predictive equation, it is a method that depends on non-linear estimators for its coefficients. Hence, better methods for estimating the degrees of freedom in PLSR are needed.

Ye (1998) introduced a concept of generalized degrees of freedom consumed (GDF), that was applicable for evaluation of the final model or fits produced by data mining. The concept of GDF is illustrated in Figure 1a. GDF was defined as the sum of the sensitivity of each fitted value to perturbations in the observed response value. In other words, it measures the flexibility of the modeling procedure by evaluating the fit of the model to small changes in the observed values. If a model is very flexible, the higher the sensitivity of the fitted values to the observed values would be. The fitted values would be close to the observed values, and the GDF would be large. The size of the perturbations is defined by a tuning parameter that must be selected and also a choice needs to be made concerning the number of perturbations. GDF not only depends on the modeling procedure, but also on the underlying true model. Unfortunately, computing GDF suffers from practical difficulties, which makes the approach not very attractive.

A practical suggestion for calculating degrees of freedom for complex modeling was given by Van der Voet (1999). Pseudo degrees of freedom (PDF) were used to assess model complexity and its calculation is based on the predictive performance of a model. The concept of PDF is illustrated in Figure 1b. There are several ways to describe the predictive performance of a model, of which assessing the prediction error based on cross validation is one of the possible manners. Van der Voet (1999) defined PDF using ordinary and cross-
validation residuals as a ratio between a measure of model fit to a measure of predictive performance. This approach is very practical, especially because cross-validation is frequently used to determine the final model fit. So, all elements that are needed to calculate PDF are often already available.

Each choice in modeling will take its degrees of freedom. For complex models, not only model parameters needs to be estimated, but also a choice must be made concerning the estimation of the meta-parameter(s). Another aspect in modeling concerns variable selection: a choice needs to be made about which variables are included in the final model. Again, by making this choice on the basis of the data themselves, the total amount of independent parts of information that are available decreases. In other words, degrees of freedom must be paid in order to be able to make the final model selection. The cost for modeling in terms of meta-parameter selection and variable selection can be determined by using an extension of the PDF calculation as suggested by Van der Voet (1999). To illustrate the concept, PLS regression was applied to real-life metabolomics data sets.

In the PLS regression, the fitted model is $\hat{y} = Xb_A$, where $A$ is the number of estimated Latent Variables (LVs) or PLS Components. Here, the meta-parameter is the number of Latent Variables $A$. It is usually determined based on assessing the mean square error of prediction (MSEP) from the deletion residual error cross-validation (Stone, 1974), in the present context named “Single Cross Validation (SCV)”, in which the $n$ available samples are split into $M$ subsets. The MSEP is a measure of how well the model estimated without sample subset # $m$ predicts the true data in subset # $m$. The $n_m$ samples in each subset are treated as an independent test set for assessing a model developed on the basis of the data in the remaining $n- n_m$ samples. When all samples, in turn, have been treated as independent test samples, the prediction error MSEP is estimated as a function of the number of LVs, $a$: $MSEP_a = \frac{1}{n} \sum_{m=1}^{M} \sum_{i=1}^{n_m} (y_{i(m)} - \hat{y}_{i(m),a})^2$, $a = 0,1,2,...,A,...$, where $i(m)$ is the sample index within subset $m$, and $n = \sum_{m=1}^{M} n_m$. In other words, MSEP is a measure for the generalization error and the lack of fit. The optimal number of LVs is found as that which gives minimum MSEP value, or at the lowest number of LVs for which the MSEP is not significantly larger than that minimum in an appropriate significance test (van de Voet 1994).

By using a Double cross-validation approach (DCV; Stone, 1974; also called cross-model-validation or CMV (Anderssen et al., 2006; Smit et al., 2007), the prediction error is determined independently of the meta-parameter estimation. By using this validation approach to calculate the PDF, the costs for meta-parameter estimation can be determined as is shown in the present paper. The concept is also illustrated in Figure 1c.
The price that needs to be paid for variable selection can also be determined by using a DCV approach for PDF calculation. In functional genomics research (e.g., metabolomics, transcriptomics) or any other field in which the number of available samples is often much lower compared to the number of variables, the impact of selection bias due to variable selection can be significant. Ambroise and McLachlan (2002) described how selection bias can be assessed by a cross-validation or bootstrap external to the selection process. It was demonstrated that it is a necessity to correct for the selection of a subset of variables in order to estimate the true prediction error. The same was demonstrated by Anderssen et al.,
Ye (1998) and Van der Voet (1999) suggested two different ways to calculate the degrees of freedom for complex models. However, the methods do not distinguish the various segments of modeling and the cost of each modeling step cannot be determined explicitly. Ye (1998) calculated degrees of freedom for the whole modeling procedure. By defining variable selection as a part of the modeling procedure, the GDF can be used to correct for bias from model selection. Although Van der Voet (1999) is quite general on predictive models, he did not explicitly partition the total PDF among different parts of the model estimation problem. By using an extension of the PDF calculation as suggested by Van der Voet (1999), the cost of meta-parameter estimation and variable selection can be determined explicitly. The common elements of this extended approach and those of Ye (1998) and Van der Voet (1999) is perturbing data and using fitted and predicted values to calculate approximative degrees of freedom. The only assumptions are independent and identically distributed (iid) errors.

**Materials and Methods**

**Data** – Two real-life metabolomics data sets were used. The first data set originates from a quality research to investigate flavours in tomatoes, which are important targets for plant breeders to improve the quality of fresh tomatoes. Metabolic profiles, AFLP marker scores, sensory trait evaluations, plant and fruit morphology measurements and consumer appreciation assessments were obtained on ripe fruits of 94 tomato varieties. Details concerning the data that was obtained for this quality research can be found in Gavai et al. (2009), Ursum et al. (2008), and van Berloo et al. (2008). Due to missing values, 2 varieties were left out of the analysis. A total of 108 GC compounds (volatiles), 25 LC compounds (non-volatiles), 29 agronomical parameters and one of the sensory measurements on 92 tomato varieties were used for the present paper. This data set will be referred to as ‘the
tomato data set’, using the sensory parameter as phenotype and the GC and LC data in combination with the agronomical parameters as ‘metabolic profile’.

The second data set was a microbial metabolomics data set, which originates from a study that was performed in order to identify metabolites that are likely inducers of phenylalalnine production (Phe) by *E.coli*. Sixteen fermentations under eight different environmental conditions were performed according to a 2 times replicated full factorial $2^3$ experimental design, varying Glucose and Succinate as carbon source, high and low phosphate concentration and pH6 and pH7. Samples were drawn at different time intervals during growth. In total, 194 samples were available. Samples were analysed for Phe production and LC-MS and GC-MS were used for metabolite analysis. Details concerning data collection and pre-processing can be found in Rubingh et al. (2009). The Phe data and GC-MS data containing 411 GC components were used for the present paper. In this paper the focus is on the degrees of freedom for the regressand $\mathbf{y}$ only, and not on the regressors $\mathbf{X}$. The temporal smoothing of Phe for the fermentation data set (Rubingh et al., 2009) reduces the total available degrees of freedom in the input Phe data. However, these initial DF losses are constant and outside the focus of the present paper, and therefore presently ignored. It was known that the Phe production varied a lot between fermentation batches. Two batches had extremely high Phe levels compared to the other batches. Also their variance within the batch was much higher than was seen in the other batches. These two batches could disturb the analysis seriously, hence they were removed before modeling. Finally, 170 samples were included for analysis. This data set will be referred to as ‘the fermentation data set’, using Phe as phenotype and the GC data as metabolic profile.

**Model and model validation** - PLS regression was used to correlate the phenotype to the metabolic data. For each input set, three steps were taken to determine the degrees of freedom used for meta-parameter estimation: (1) a Double Cross Validation, (2) a Single Cross Validation, and (3) a final PLS regression fit to estimate the parameters utilizing all samples. A Double Cross-Validation approach is used to determine the error of modeling independently of the meta-parameter estimation. The meta-parameter of the final model is estimated using the Single Cross-Validation. The final fit of the PLS regression model is based on the meta-parameter estimation of the SCV and the error of the regression model is based on DCV (Smit et al., 2007).

Firstly, a Double Cross Validation (DCV) was performed for each of the two data sets. The analysis of the tomato data was based on a tenfold venetian blind DCV. Ten percent of the samples ($n(1)$) were left out in an outer cross validation loop. The remaining data (n-$n(1)$) were used in an inner loop, in which an 'Inner Single Cross Validation' (ISCV) was used for estimating MSEP and the optimal model rank, $A$. The ISCV consisted here in leaving out, in turn, 10% of the samples ($n(2)$), for testing a model built on the remaining $n - n(1) - n(2)$ samples with respect to MSEP at rank $A(2) = 0, 1, 2,...$. The optimum number of LVs $A(2)$ was determined based on the minimal ISCV prediction error.
The analysis of the fermentation data was based on a 'leave-one-fermentation-out' DCV. For 'leave-one-fermentation-out' Double Cross Validation (DCV), all available time points $n_m(1)$ of the one fermentation batch $m(1)$ at a time were left out in an outer cross validation loop. Data of the remaining $M-1$ fermentations were used in an inner loop, in which an 'Inner Single Cross Validation' (ISCV) was used for estimating MSEP and the optimal model rank, $A$. The ISCV consisted here in leaving out, in turn, each of the $M-1$ remaining fermentation batches $m(2)$ with its $n_m(2)$ samples, for testing a model built on the remaining $n$- $n_m(1) - n_m(2)$ samples with respect to MSEP at rank $a_m(2)=0,1,2,...$. The optimum number of LVs $A_m(2)$ was determined based on the minimal ISCV prediction error.

To prevent over-fitting, the number of LVs that was used to determine each outer loop prediction was chosen as follows: 1) choose the minimal value of the RMSEP (square root of MSEP) based on the inner loop predictions, 2) compare this minimal value to the RMSEP using one LV less compared to the number that was chosen, 3) for a simple approach, if the reduction in RMSEP between these two values was less than 5%, the model with one LV less was chosen. This number of LVs was used to fit a PLS model on the inner loop samples, which was then used to predict the measurements of the samples that were left out in the outer loop. The outer loop predictions were saved and the procedure was repeated until all samples were left out once in the outer loop. As a consequence of this double cross-validation approach, the outer loop predictions are based on different model complexities, so a slightly different set of samples was used for meta-parameter estimation as was used for error prediction (Smit et al., 2007).

Secondly, a Single Cross Validation (SCV) was used for each of the two data sets. Ten percent of the samples were left out per cross-validation step for modeling the tomato data set (a tenfold venetian blind cross-validation approach) and all available measurements per fermentation were left out per cross-validation step for modeling the fermentation data set (a 'leave-one-fermentation-out' cross validation approach). Based on the minimal prediction error for the measurements that were left out, the optimum number of LVs was determined. The same restriction of 5% change in RMSEP as in the ISCV was used to determine the optimum number of LVs for final modeling. As a consequence of this validation approach, the same set of samples was used for meta-parameter estimation as well as for error prediction.

Finally, the number of LVs that was determined in the previous described SCV procedure was used to built a final PLS model using data of all samples, resulting in a set of final fit predictions. The phenotype levels that were predicted using this model are indicated as 'final fitted values'. As a result, three sets of predictions were available, namely based on DCV, based on SCV and based on the final fit.

**Random subsets** - To investigate the effect of the number of samples on the used DFs, random subsets of the original data sets were made. Four levels of selection were defined, using about 30%, 50%, 70% and 90% of the data. The random selection for the fermentation
data set was stratified by fermentation. For each level, hundred random selections from the original data sets were made to be able to assess the variation in the estimation of the degrees of freedom, resulting in 4x 100 subsets.

By using 100 different subsets for various sizes of the data set, the variance in modeling can be assessed. However, also bias might be present. The impact of bias can be reduced by leaving out only a few samples per cross-validation step. But if only one sample is left out, which is the case in a leave-one-out cross-validation approach, the variance component cannot be estimated. Therefore, Hastie et al. (2001) recommended the use a tenfold cross-validation approach, leaving out 10% of the data per cross-validation step. The difference between the DCV and the SCV results, might be influenced by bias. However, since the number of samples that were left out in the outer loop of the DCV and each step of the SCV are comparable, the impact of bias will be similar for both approaches. A tenfold cross-validation was used for the tomato data set. By using leave-one-fermentation-out each cross-validation for the fermentation data set, about 7% of the data was left out per cross-validation step, meaning that it was between a leave-one-sample out and a tenfold cross-validation scheme, hence on the safe side.

**Degrees of freedom** - Van der Voet (1999) proposed a definition for pseudo degrees of freedom (PDF) for PLS models based on predictive performance. The PDF are defined as:

\[
PDF = n \left[ 1 - \sqrt{\frac{MSEP_{fit}}{MSEP_{cv}}} \right],
\]

where \( n \) is the number of objects, \( MSEP_{fit} \) is the mean squared error based on fitted values and \( MSEP_{cv} \) is the mean squared error based on prediction from cross-validation. The PDF are based on an SCV procedure and concern the degrees of freedom that are needed for estimation all model parameters, given the meta-parameter, which is the number of LVs that is needed for modeling. An extension to the calculation by van der Voet was used to investigate the costs for meta-parameter estimation:

\[
PDF_{scv} = n \left[ 1 - \sqrt{\frac{MSEP_{fit}}{MSEP_{scv}}} \right] \quad (2a)
\]

\[
PDF_{dcv} = n \left[ 1 - \sqrt{\frac{MSEP_{fit}}{MSEP_{dcv}}} \right] \quad (2b)
\]

\[
PDF_{mp} = PDF_{dcv} - PDF_{scv}, \quad (2c)
\]

where \( PDF_{scv} \) is equal to (1), \( PDF_{dcv} \) are the PDF in which the predictions are based on predicted values in the outer loop of the DCV, and \( PDF_{mp} \) are the costs for meta-parameter estimation. \( MSEP_{dcv} \) are the predictions based on DCV, \( MSEP_{scv} \) are the predictions based on SCV, and \( MSEP_{fit} \) are the predictions based on the final fit of the model. (2a)-(2c) were used for the original data set as well as for the 400 random subsets. The degree of variability in the PDF estimates between the 400 subsets was also assessed. The \( PDF_{mp} \) is compared with the total degrees of freedom in the data set, which is considered to be equal to the number of objects in the data set, \( n \). This means for the tomato data set: \( n = 92 \), and when using 30%
of the data, n = 28; using 50%, n = 46; using 70%, n = 65; and using 90%, n = 83. For the fermentation data set this means n = 170, and when using about 30% of the data per fermentation time series, n = 57; using 50%, n = 87; using 70%, n = 127; and using 90%, n = 157.

Variable selection – A jackknife approach was used for variable selection, which was performed within both ways of cross-validation. The variable selection in DCV was performed as follows:

1) Define a test set with all observations that are left out in the first outer loop and define a training data set containing all available remaining measurements;
2) Perform an SCV on the training set as described before (ISCV);
3) Determine meta-parameter (# LVs);
4) Calculate the relative standard deviation (RSD) for each variable using the regression coefficients of each step of the inner loop using the number of LVs as determined in 3) and select only those variables with RSD < 50%;
5) Perform a second ISCV on the training set using only the variables which are selected in 4);
6) Determine meta-parameter (# LVs);
7) Use variables from 4) and meta-parameter from 6) to predict the samples which were left out in the outer loop (in 1));
8) Repeat 1) - 7) until all samples are left out once in the outer loop.

The variable selection in SCV was performed is described in 2-6 for DCV. The variables from 5) and the meta-parameter from 6) are used for final modeling to obtain the final model predictions.

The selection procedure was completed for the original data set as well as the 400 subsets. Again, (2a) – (2c) were used to calculate the costs for modeling. Now, PDF$_{mp}$ indicates not only the price for meta-parameter estimation but also for variable selection. PDF$_{scv}$ represent the costs for estimating the parameters of the PLSmodel, given a selection of variables and a number of LVs, whereas the PDF$_{dcv}$ includes the costs of selecting variables and choosing a number of LVs. The difference between PDF$_{scv}$ and PDF$_{dcv}$ is the cost for variable selection and choosing the number of LVs. This measure is compared to the total degrees of freedom in the data set (n) in order to estimate the true impact of the modeling choices.

Model assumptions – As mentioned before, the proposed approach is not only applicable for PLS regression models, but can be generalized to any predictive model assuming iid errors. For the fermentation data set, multiple measurements were taken for each batch, indicating that both within and between batch variation is modelled. By centering the data
per batch, the contribution of between batch variation could be reduced to almost zero. However, this would imply that the errors are not completely independent anymore since the same data was used for correction as was used for modeling. DCV was used instead of SCV, exactly for that purpose, hence it was decided not to correct for between batch differences in the fermentation data set and to model both within and between batch processes in PLS regression.

**Data pre-processing** - Within each inner loop of the double cross-validation procedure, data were scaled to mean zero and unit variance. The inner loop scaling parameters were used to scale data in the outer loop.

**Software** - All analyses were performed using Matlab Version 7.0.4 R14 (The Mathworks, Inc.) and the n-way toolbox version 2.11.

**Results and Discussion**

The price that needs to be paid for meta-parameter estimation and variable selection were determined by using a DCV approach for PDF. In Table 1, the results for the complete data sets are given, for both the tomato and the fermentation data set and for both modeling with and without variable selection. The degrees of freedom based on the model without variable selection indicate the costs for determining the number of LVs. There is only a small difference in degrees of freedom, which means that only a small price is paid to obtain the number of components.

**Table 1.** Costs of meta-parameter estimation and variable selection in PLS-modeling expressed as the absolute number of degrees of freedom (PDF) consumed by the PLS-modeling, calculated using single (SCV) and double (DCV) cross-validation, as well as the difference in degrees of freedom ($\Delta$), which are the cost for modeling.

<table>
<thead>
<tr>
<th></th>
<th>DCV</th>
<th>SCV</th>
<th>$\Delta$</th>
<th>n</th>
<th>$%\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># var</td>
<td>PDF</td>
<td># var</td>
<td>PDF</td>
<td>LVs</td>
</tr>
<tr>
<td>Tomato</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without</td>
<td>162</td>
<td>25.0</td>
<td>162</td>
<td>23.0</td>
<td>3</td>
</tr>
<tr>
<td>With</td>
<td>109-118</td>
<td>26.5</td>
<td>114</td>
<td>21.1</td>
<td>3</td>
</tr>
<tr>
<td>Fermentation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without</td>
<td>408</td>
<td>78.7</td>
<td>408</td>
<td>74.5</td>
<td>2</td>
</tr>
<tr>
<td>With</td>
<td>265-300</td>
<td>79.8</td>
<td>298</td>
<td>63.7</td>
<td>2</td>
</tr>
</tbody>
</table>

1 $PDF_{dcv} = n[1 - \sqrt{(MSEP_{fit} / MSEP_{dcv})}]$

2 $PDF_{scv} = n[1 - \sqrt{(MSEP_{fit} / MSEP_{scv})}]$

3 $\Delta = PDF_{dcv} - PDF_{scv}$

4 $\%\Delta = 100*(PDF_{dcv} - PDF_{scv})/n$

The difference between $PDF_{scv}$ and $PDF_{dcv}$ ($\Delta$ in Table 1) that are calculated using variable selection is a measure for the cost of the variable selection and meta-parameter estimation, i.e. how many degrees of freedom are consumed when optimizing the modeling of y from X by testing and removing X-variables based on their apparent predictive contribution for y. In other words, the $PDF_{scv}$ represent the costs for estimating parameters of the PLS model,
given the selected variables and the chosen number of LVs, whereas the $PDF_{dcv}$ is determined unconditionally of the number of selected variables and chosen number of LVs. The difference between $PDF_{scv}$ and $PDF_{dcv}$ after variable selection ($\Delta$ in Table 1) is larger compared to the difference that was calculated without variable selection, hence variable selection is much more 'expensive' than choosing the number of LVs.

Table 2a. Costs of meta-parameter estimation and variable selection in PLS-modeling ($\Delta$) expressed as summaries (over 100 repeated resamplings per subset) of the absolute number of degrees of freedom (PDF) consumed by the PLS-modeling, calculated using single (SCV) and double (DCV) cross-validation for several sizes of the tomato data set.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Without variable selection</th>
<th>With variable selection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90%</td>
<td>70%</td>
</tr>
<tr>
<td>N</td>
<td>83</td>
<td>65</td>
</tr>
<tr>
<td>$PDF_{scv}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>13.8</td>
<td>2.7</td>
</tr>
<tr>
<td>Median</td>
<td>21.9</td>
<td>20.6</td>
</tr>
<tr>
<td>Max</td>
<td>29.1</td>
<td>36.5</td>
</tr>
<tr>
<td>IQR</td>
<td>4.4</td>
<td>5.8</td>
</tr>
<tr>
<td>$PDF_{dcv}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>15.0</td>
<td>4.4</td>
</tr>
<tr>
<td>Median</td>
<td>21.1</td>
<td>19.9</td>
</tr>
<tr>
<td>Max</td>
<td>27.5</td>
<td>34.0</td>
</tr>
<tr>
<td>IQR</td>
<td>3.0</td>
<td>4.9</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>0.6</td>
<td>1.0</td>
</tr>
<tr>
<td>%$\Delta$</td>
<td>0.7</td>
<td>1.5</td>
</tr>
<tr>
<td>max%$\Delta$</td>
<td>4.1</td>
<td>8.2</td>
</tr>
</tbody>
</table>

$\Delta = \text{median}(PDF_{dcv} - PDF_{scv})$

%$\Delta = 100 \times (\text{median}(PDF_{dcv} - PDF_{scv})) / n$

max%$\Delta = 100 \times (\max(PDF_{dcv} - PDF_{scv})) / n$

In Table 2a and 2b, the results are presented for the subset analysis of the tomato data set and the fermentation data set, respectively. For each level of subset analysis, the pseudo-degrees of freedom (PDF) consumed were estimated. The median of these degrees of freedom estimates over the 100 repeated subsets are shown, as well as their minimal and maximal value. The inter-quartile range (IQR), the interval about the median containing 50% of the data, is also calculated as a measure of variation of the degrees of freedom. Index $n$ gives the number of samples and is considered as the total degrees of freedom in the particular data set, $DF_{total}$. The relative change in DF (%$\Delta$ in Table 2a and 2b) is calculated as percentage of the number of calibration samples used for the four different subset sizes ($n = 83, 65, 46$ or $28$ for the tomato data, and $n=156, 127, 87$ or $57$ for the fermentation data). The relative difference clearly shows the price that is paid for meta-parameter estimations and for variable selection. For the tomato data set, the costs for meta-parameter estimation are between 0.7% and 3.2% and the costs for variable selection and meta-parameter estimation are between 6.1% and 11.8%. These are the percentages of the originally
available parts of independent information that must be paid for the choices that are made. For the fermentation data set, these costs for meta-parameter estimation are between 2.7% and 3.5% and for variable selection are between 13.4% and 15.4% (Figure 2).

Table 2b. Costs of meta-parameter estimation and variable selection in PLS-modeling (Δ) expressed as summaries (over 100 repeated resamplings per subset) of the absolute number of degrees of freedom (PDF) consumed by the PLS-modeling, calculated using single (SCV) and double (DCV) cross-validation for several sizes of the fermentation data set.

<table>
<thead>
<tr>
<th>Without variable selection</th>
<th>With variable selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subset</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>PDF&lt;sub&gt;dcv&lt;/sub&gt; Min</td>
<td>83.2</td>
</tr>
<tr>
<td>Median</td>
<td>89.5</td>
</tr>
<tr>
<td>Max</td>
<td>103.8</td>
</tr>
<tr>
<td>Iqr</td>
<td>3.5</td>
</tr>
<tr>
<td>PDF&lt;sub&gt;scv&lt;/sub&gt; Min</td>
<td>80.6</td>
</tr>
<tr>
<td>Median</td>
<td>85.3</td>
</tr>
<tr>
<td>Max</td>
<td>89.4</td>
</tr>
<tr>
<td>Iqr</td>
<td>2.5</td>
</tr>
</tbody>
</table>

1. Δ<sup>1</sup> = \( \text{median}(\text{PDF}_{dcv} - \text{PDF}_{scv}) \)
2. %Δ<sup>2</sup> = 100*(median(\(\text{PDF}_{dcv} - \text{PDF}_{scv}\)))/n
3. max%Δ<sup>3</sup> = 100*(max(\(\text{PDF}_{dcv} - \text{PDF}_{scv}\)))/n

The costs for meta-parameter estimation and variable selection are relatively higher for smaller data sets compared to the larger data set, which can be explained in terms of model complexity. In Table 3a and 3b, the frequencies of the number of latent variables are given for the various subsets for the tomato data set and the fermentation data set, respectively. More complex models using more LVs are selected for the larger subsets using 70% and 90% of the data: more data means more possibilities to include details in the model. Much more variation in model complexity is seen in the smaller data sets. Apparently, less effort is needed to determine the complexity of the model for the larger data sets compared to the smaller data sets.
Figure 2. Relative costs in terms of pseudo degrees of freedom ($\%\Delta$) for meta-parameter estimation and variable selection for the tomato and the fermentation data set.

Table 3a. Frequency table of model complexity for several sizes of the data set: percentages of chosen number of latent variables in the PLS model applied to the tomato data.

<table>
<thead>
<tr>
<th>Subset</th>
<th>#LV</th>
<th>Without variable selection</th>
<th></th>
<th>With variable selection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>90%</td>
<td>70%</td>
<td>50%</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>27</td>
<td>41</td>
<td>42</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>72</td>
<td>42</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>1</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3b. Frequency table of model complexity for several sizes of the data set: percentages of chosen number of latent variables in the PLS model applied to the fermentation data.

<table>
<thead>
<tr>
<th>Subset</th>
<th>#LV</th>
<th>Without variable selection</th>
<th></th>
<th>With variable selection</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>90%</td>
<td>70%</td>
<td>50%</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>100</td>
<td>100</td>
<td>89</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>
The conclusions based on the fermentation data set are comparable to the ones based on the tomato data set: variable selection is more expensive than meta-parameter estimation and the costs for meta-parameter estimation and variable selection are relatively higher for smaller data sets compared to the larger data set. However, the results for the fermentation data set are more extreme than for the tomato data set. This illustrates an important and treacherous phenomenon, called pseudo-replication. As stated before, the experimental design of the fermentation data has been ignored for modeling. By doing so, the hierarchical structure is ignored and a model was applied assuming \( n \) independent observations. Only for independent observations we can state total DF = \( n \). With correlated observations such as arise in hierarchical designs the total DF should be less than \( n \), at least intuitively (remember that there is no formal definition of DF in such a case). As an extreme example consider the case where all observations in a fermentation batch would be exactly equal. Then the total DF would be 14 (\( \equiv \) number of batches) rather than 170 (\( \equiv \) number of observations). Of course in reality data sets are never that extreme but in general we expect total DF to be less than \( n \) observations in case of dependencies within the larger units. Modeling seems to cost relatively more in terms of DF, but this is due to pseudo-replication, which makes that the initial number of \( n \) DF must be interpreted with great care and must only be used as indicative.

**Figure 3a.** Distribution of estimated pseudo-degrees of freedom (in percentage of number of calibration samples; \( n=83, 65, 46 \) or 28, respectively) lost by PLSR model optimization for the tomato data set, including rank estimation, variable selection and remaining parameter estimation. For visual comparison, straight lines have been added for \( PDF_{dCV} = PDF_{scv} \).
A lot of variation is seen within a subset, which is illustrated in a density plot of $PDF_{scv}$ versus $PDF_{dcv}$ per subset (Figure 3a and 3b). In general, the smaller the data set, the more variation is seen. The density cloud using 90% of the original data set is less fuzzy than the cloud for the 30% subset. It is also seen that the smaller the data set, the more extreme values of degrees of freedom are found: much more islets are seen for the 50% and 30% data set than for the 70% and 90% subsets. For the smallest subset, using 30% of the data, the costs for meta-parameter estimation and variable selection, differ between 14 and 43 degrees of freedom. At its worst, 30% of the apparent total degrees of freedom are lost. However, the results for the 30% data set must be put in perspective to some extent. Resampling assumes that the sampling distribution reflects the population distribution and that is likely not the case for small sample sizes. As seen before in Rubingh et al. (2006), small data sets show a lot of variation, which is also seen in Figure 3. Nevertheless, even for the 50% data set, which shows less variation than the 30% data set, still 20% of the apparent total degrees of freedom can be lost due to modeling.

**Figure 3b.** Distribution of estimated pseudo-degrees of freedom (in percentage of number of calibration samples; $n=156$, $127$, $87$ or $57$, respectively) lost by PLSR model optimization for the fermentation data set, including rank estimation, variable selection and remaining parameter estimation. For visual comparison, straight lines have been added for $PDF_{dcv} = PDF_{scv}$. 
Conclusion

Each data-driven action in data modeling consumes degrees of freedom, whether it concerns estimation of parameters (such as regression coefficients), estimation of meta-parameters (such as optimal model rank) or estimation/choice of variable weights (variable selection). Ye (1998) and Van der Voet (1999) introduced the GDF and PDF, respectively, to estimate the degrees of freedom for complex models. However, both approaches do not distinguish the various segments of modeling, like meta-parameter estimation or variable selection. By using a double cross validation approach for PDF calculation, these costs can be determined explicitly. PLS regression was applied to two real-life metabolomics data sets to illustrate the concept. The idea is not restricted to PLS only, but can be generalized to many predictive modeling techniques which assume iid errors. Unfortunately, equality of variance (homoscedasticity) is not always a realistic assumption in chemical analysis, where high measured values tend to have higher uncertainty than low measured values. If necessary, a proper data transformation, like a log transformation, can be applied to obtain homoscedasticity and the method can be applied. There is also no restriction to the method that is used for variable selection.

Considering the fact that in functional genomics research the number of objects is mostly even smaller than the smallest subset that was used for the present paper, the impact of ignoring the costs of variable selection becomes very alarming. Bearing in mind the direct link between the degrees of freedom and the power, it emphasizes the necessity to include enough objects in, for instance, functional genomics studies, in which variable selection is frequently applied such as biomarker discovery studies.

Acknowledgement

We gratefully acknowledge Mariët van der Werf (formerly TNO Quality of Life, Zeist, The Netherlands) and Karin Overkamp (TNO Quality of Life, Zeist, The Netherlands) for making available the microbial metabolomics data. The tomato data that was used as an example in this paper was (co)financed by the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative / Netherlands Organisation for Scientific Research.

References


OUTRODUCTION

Carina M Rubingh

TNO Quality of Life, Zeist, The Netherlands
Outroduction

Freely translated from Webster’s comprehensive dictionary, real life metabolomics studies would be about studying the true or actual state in which animals and plants exits, including all physical and chemical processes that are constantly taking place in living organisms, which distinguishes them from inorganic substances and from dead organisms. In other words, it is about understanding life.

Life is mysterious. Although it is expectable that it is unlikely that life will ever be understood completely, curiosity makes that the human species is always trying to explain and to understand what is happening. A scientific discipline that can help understanding life is called ‘metabolomics’, which concerns studies that can be used to find answers to questions related to responses of biological systems on environmental influences. All findings in all metabolomics studies that will ever be done, will only cover a very small part in understanding life, but it is a start.

In this light, even though it is a bio-statistical thesis, the chapter with the least statistical contents might be considered as the most important one: a (very) small bit of life is better understood. In chapter 5, the biological context is given for one of the statistical chapters (chapter 4). A method or analysis strategy might be perfect from a statistical point of view, but if the outcomes give no answers to biological questions, it is completely useless. Therefore, it is always important to apply the method to real life data. Of course, this does not mean that simulated data should never be used, since it can help in finding out how the method should work out: if you know exactly what is going in, you know what should come out. However, it can never be the final stage in method development, since real life data, especially from biological samples, will never behave as simulated data.

Since metabolomics studies become more and more complex, the urge for better and more sophisticated data analysis techniques is increasing. For good reason, the meaning of complex is complicated, which is certainly not similar to impossible. The complexity must be used instead of ignored. A well considered study design, and, as a consequence, a well considered analysis plan, makes it possible to obtain new information in the most optimal way. In this thesis, several examples are given of how to deal with complex metabolomics data sets. The complications can occur in various stages of the data analyses, namely during data exploration (chapter 2), data modeling (chapter 3, chapter 4, chapter 5) and during model validation (chapter 6, chapter 7).

One way to explore data is to calculate the correlation between data sets. It is often convenient to have a single simple number characterizing the relationship between pairs of high-dimensional datasets in a comprehensive way. Matrix correlations are such numbers and are appealing since they can be interpreted in the same way as Pearson’s correlations familiar to biologists. There are many measures that can be used to express this correlation. Unfortunately, it turned out that of the common used measures was not appropriate for
high-dimensional data, making this coefficient invalid for many metabolomics studies. The RV coefficient that is described in this thesis is modified in such a way that the use of it is valid for high-dimensional data. The modified RV-coefficient can be used in high-dimensional data analysis studies as an easy measure of common information of two datasets (chapter 2).

A complicating factor in making the right choices in modeling is when samples are measured according to a certain design. A design of experiments can be used to cause variations at different levels and over a broad range. If the interest is in the evolution over time, repeated measurements can be done, which is also a kind a designing the data. A design can increase the information that can be obtained from a data set. However, it also complicates the statistical analysis. Straightforward MVA techniques are not sufficient, because the structure in the data is ignored. Approaches like multilevel analysis or three-way analyses are in that case more appropriate since they take the structure explicitly into account. The use of experimental designs is very common in microbial metabolomic studies. Notwithstanding all biological, analytical and statistical challenges in the generation and analysis of these complex data sets, valid models can be obtained which also result in biologically relevant outcomes. A longitudinal experimental design in combination with metabolomics and multi-way data analysis is a powerful approach in the identification of metabolites whose correlation with bio-product formation shows a shift in time (chapter 3).

The biological relevance of a model is a measure of its validity. A statistical model only makes sense if it answers the (biologically driven) research question of interest. If there is no biological rationale for the statistical model, the model can be considered meaningless or ‘invalid’. By taking the data structure into account in the statistical analysis, even subtle treatment effects can be discovered. By applying a well considered study design, information can be obtained in the optimal way. However, it will have consequences for the data analysis. Straightforward multivariate data analysis techniques are no longer sufficient. Given the design of the human intervention trial that is described in this thesis, the multilevel multi-way technique turned out to be a much stronger tool than the ordinary method. The metabolites that contributed most to treatment differences were not only statistically, but also biologically relevant. The multilevel approach found the effects that were better interpretable, whereas the ordinary approach failed to do so (chapter 4, chapter 5).

A statistical model on metabolomics data can always be made. Due to the high number of variables that are included in the analysis, it will be hard not to find a model with a good fit. Unfortunately, a good fit will not guarantee stable and reliable findings. The model that is found will be due to chance instead of due to true effects. Especially since it is more a rule than an exception that the data sets contain measurements on many variables for only a few samples. Leaving out a few samples can give a totally different outcome. Therefore, a thorough validation is needed to rely on the results. However, an extended validation will
also not bring universal happiness either, as is illustrated in this thesis. It is shown that also a
good standard validation method like cross-validation is sensitive to sampling. The lower the
number of subjects compared to the number of variables, the less the outcome of validation
tools such as cross-validation, jack-knifing and permutation tests can be trusted. The result
depends crucially on the specific sample of subjects that is used for modeling. The validation
tools cannot be used as warning mechanism for problems due to sample size or
representativity of the samples (chapter 6).

Cross-validation is obligatory to find a reliable outcome, but it cannot be applied
indiscriminately. Choices on how to perform the cross-validation procedure must be made in
good consideration. And besides that, more choices must be made in order to find a solid
model. The meta-parameters of the model (i.e. the optimal model complexity) must be
determined and often a selection of most important or most contributing metabolites or
peaks must be made. The cost for these modeling aspects can be expressed as degrees of
freedom. Considering degrees of freedom as independent parts of information in a data set,
for each data-driven ‘choice’ that is to be made concerning modeling, like meta-parameter
estimation or variable selection, a price has to be paid in terms of losing a number of these
independent parts. The more choices are made, the less independent information is
available, thereby lowering the number of degrees of freedom left over. If many choices are
made based on a small data set, hence on a set with little degrees of freedom, the cost
might be disproportionate to the profit. If too much available degrees of freedom are
consumed by the choices that are made for modeling, questions may be raised on the
validity of the outcome. Calculating the costs of complex model optimization gives insight
into this. It was shown that only a small price needed to be paid to determine the meta-
parameter whereas variable selection was much more ‘expensive’. Concerning a relatively
small data set, almost 35% of the total degrees of freedom could be lost due to variable
selection. Considering the fact that in metabolomics research the number of objects is often
very small, the impact of especially variable selection becomes very alarming (chapter 7).

**Bridging**

For this thesis, data of five different studies were used, obtained from different
measurement platforms and analysed in different ways. The challenges that were risen in
these studies, will be recognizable for many researchers. Fortunately, the methods and
solutions that are described in this thesis are directly applicable in all other (biological)
studies in which similar issues need to be addressed. And even if the question of interest is a
bit different and the approaches cannot directly be applied, hopefully, reading this thesis
makes people aware of the fact that it is important to think about the consequences but also
about the possibilities of structured data for validation and analysis purposes.
Divine the future

The end of possibilities of metabolomics studies has not yet been reached, and it is unlikely that it will in the near future. However, it will be sure that current metabolomics will change. What is required from metabolomics in the future may depend on the application area. However, each application area is served by analytical method development, which will make it possible to measure more and more metabolites: not only the 'low hanging fruits', but also the ones with (very) low intensities. The ability of measuring all existing metabolites will not be enough. It is perhaps even more important, that they all can be identified.

A long list of targeted identified metabolites improves the translation into biology, and thus it improves the interpretation of the findings. It will also increase the acceptability of metabolomics studies. In, for instance, microbiological studies it is more common to apply kinetic flux models than to perform 'metabolomics' for production improvement. As long as the sensitivity of measurements is not improved and the number of identified metabolites is not increased, it will be hard to convince the microbiologist that metabolomics studies can increase his knowledge. It is at this point, that a pitfall for statistical analysts may occur: for many statistical analysis, the identification of the metabolites is often not of importance, so the improvement of statistical methods for the analysis of metabolomics data is not pending on metabolite identification, which may cause a gap between what the industry needs (market) and what is statistically possible (knowledge). The risk of improving a statistical method is that it becomes more and more sophisticated, which is for non-statisticians already often synonymous for too difficult and therefore not useful, which will be even more the case when it results in outcomes that cannot be interpreted due a missing identification. So, as long as the results are based on unidentified metabolites, it will remain a challenge to get complex data analyses accepted, even if the method is more appropriate from a statistical point of view.

The more is known about the identification of metabolites, the more metabolomics studies will be accepted, and the broader the application area will become. In, for instance, human health research, it can be expected that the application area of metabolomics will make a shift, which may lead to new demands. Until now, the focus of metabolomics in this research area is mainly on finding causes of diseases. But it will become more important to be able to predict the development of a disease in an individual or to prescribe better targeted medicine for a single person than to investigate general effects in a general population, hence the application of metabolomics must become more personalized. Since it is often said that 'the bigger n, the better', this is a serious challenge for statistics. From a statistical point of view, 'N=1'-studies are considered to be unreliable, but the questions of interest in human health research are seeking for an 'N=1'-answer. These answers can only be found if the statistical analysis becomes more focused, which means that dealing with data in the most appropriate way will not be enough. Including knowledge about biological processes
may complicate the analysis, but it also means that it becomes focussed on relevant effects, only.

The integration of measurements describing effects in various biological elements, like metabolites, proteins and genes, is already a topic and is thought to provide a print-out of a biological system. However, this print-out is incomplete, since psychological and environmental aspects are often neglected. Nevertheless, these aspects can be very influential in identifying a person, as, for instance, is illustrated by the placebo-effect, which is a perfect example of how the spirit can influence biological processes. Psychological or environmental differences may explain yet unknown phenomenons, like why some individuals respond very well to a particular drug whereas the same drug shows no reaction in others. The possibilities to measure these psychological and environmental aspects are there, but the integration with system biology disciplines like metabolomics stay behind.

It is clear that statistics will not be leading in divining the future of metabolomics studies, since it must always be driven by biology. However, every change in the field of metabolomics will ask for its own appropriate statistical solution. The future of metabolomics studies is promising, and while the analysts are working on improving the measurements and extending the metabolite identification, it is up to the statistical analysts to be sure that the gap between the statistical possibilities and the market will not become irreconcilable. All upcoming changes will bring up new challenges for data analysis: MVA techniques like PCA and PLS were sufficient in the early days of metabolomics studies (history); techniques that can deal with structured data are needed for contemporary metabolomics studies (this thesis); and methods that can deal with, for instance, time resolved data, (very) small samples sizes or techniques which can integrate data from different disciplines are upcoming (future). However, as long as the human species is curious, the ultimate goal of metabolomics studies will not change: understanding life, even if it is only for a little bit.

Acknowledgements

The author expresses her gratitude to Mariët van der Werf, Ben van Ommen and Jan van der Greef for their glance into the crystal ball.
SUMMARY

Carina M Rubingh

TNO Quality of Life, Zeist, The Netherlands
Summary

The homo sapiens is a curious species. He wants to know what is happening around him, and even more, he wants to understand why things are happening and how things are happening. Whether his questions are spiritual or down to earth, most questions are related to 'life': the homo sapiens wants to understand life. There is no science discipline that can explain all of these questions and the combination of all of them may explain only a little bit about life. But it is a start in feeding the curiosity of the human species. Questions that are related to responses of biological systems on environmental influences due to, for instance, toxicological exposure, nutrition or medical treatment, can be investigated by metabolomics studies. At the early days of metabolomics studies, multivariate techniques like Principal Component Analysis and Partial Least Squares regression were sufficient for statistical analysis. But nowadays, the metabolomics studies become more and more complicated and the urge for better and more sophisticated data analysis techniques is increasing. Moreover, there is also a growing demand on validation of results: how reliable are our statements about the data; what did they actually learn us? The present thesis is a proof of this.

Modern functional genomics generates high-dimensional datasets. It is often convenient to have a single simple number characterizing the relationship between pairs of such high-dimensional datasets in a comprehensive way. Matrix correlations are such numbers and are appealing since they can be interpreted in the same way as Pearson's correlations familiar to biologists. The high-dimensionality of functional genomics data is, however, problematic for existing matrix correlations. In chapter 2 an improvement of the most promising matrix correlation coefficient (the RV-coefficient) circumventing the problems of high-dimensional data is given. The modified RV-coefficient can be used in high-dimensional data analysis studies as an easy measure of common information of two datasets. This is shown by theoretical arguments, simulations and applications to two real-life examples from functional genomics.

In chapter 3 it is shown that a longitudinal experimental design in combination with metabolomics and multi-way data analysis is a powerful approach in the identification of metabolites whose correlation with bio-product formation shows a shift in time. A strategy is presented for the analysis of longitudinal microbial metabolomics data, which was performed in order to identify metabolites that are likely inducers of phenylalanine production by Escherichia coli. The variation in phenylalanine production as a function of differences in metabolism induced by the different environmental conditions in time was described by a validated multi-way statistical model. Notably, most of the metabolites showing the strongest relations with phenylalanine production seemed to hardly change in time. Apparently, potential bottlenecks in phenylalanine seem to hardly change in the course of a batch fermentation.

Sophisticated approaches like multilevel analyses are needed to discover subtle differences between healthy people and people at the onset of (a multi-factorial) disease, as is
presented in chapter 4. Multilevel analysis generates different sub-models for each level of variation. For instance, within and between subject variation can be split and analyzed separately if the two factors are orthogonal (i.e., not confounded). The benefits of a multilevel approach in multi-way analysis are described for the analysis of metabolomics data of an double blinded, randomized, parallel intervention trial with twenty slightly overweight men, whom received a diclofenac or placebo treatment for nine days. The multilevel multi-way technique turned out to be a much stronger tool for modeling differences between treatment groups than an ordinary method.

The metabolites that contributed most to treatment differences in this human intervention trial were not only statistically, but also biologically relevant as is illustrated in chapter 5. Metabolic changes were subtle and were only detected using metabolic profiling in combination with an oral glucose tolerance test. The repeated measurements during the oral glucose tolerance test increased statistical power, but the metabolic perturbation also revealed metabolites that respond differentially to the oral glucose tolerance test. Specifically, multiple metabolic intermediates of the glutathione synthesis pathway showed time-dependent suppression in response to the glucose challenge test. The fact that this is an insulin sensitive pathway suggests that inflammatory modulation may alter insulin signaling in overweight men.

In chapter 6 the performance of statistical model validation tools such as cross-validation, jack-knifing model parameters and permutation tests, which are meant to obtain an objective assessment of the performance and stability of a statistical model, are assessed for megavariate data sets. The smaller the number of subjects compared to the number of variables, the less the outcome of validation tools such as cross-validation, jack-knifing model parameters and permutation tests can be trusted. The result depends crucially on the specific sample of subjects that was used for modeling. It illustrates that the validation tools cannot be used as warning mechanism for problems due to sample size or to representativity of the sampling.

Each data-driven action in data modeling consumes degrees of freedom, whether it concerns estimation of parameters (such as regression coefficients), estimation of meta-parameters (such as optimal model rank) or selecting variables. In chapter 7, a double cross validation approach for the degrees of freedom calculation is presented to determine the costs for meta-parameter estimation and variable selection explicitly. To illustrate the concept, PLS regression was applied to two real-life metabolomics data sets. It is shown that only a small price needs to be paid to determine the meta-parameter whereas variable selection is much more 'expensive'. Concerning a relatively small data set, almost 35% of the total degrees of freedom could be lost due to variable selection. Considering the fact that in functional genomics research the number of objects is mostly even smaller than the smallest subset that was used for this chapter, the impact of ignoring the degrees of freedom consumed by variable selection becomes very alarming.
SAMENVATTING

Carina M Rubingh

TNO Quality of Life, Zeist, The Netherlands
Samenvatting

De mens is van nature nieuwsgierig. Hij wil niet alleen weten wát er om hem heen gebeurt, maar hij wil ook weten waarom het gebeurt en hoe het komt dat het gebeurt. Of zijn vragen nu spiritueel of rationeel van aard zijn, de meeste van deze vragen zijn gerelateerd aan ‘het leven’: de mens wil het leven begrijpen. Er is geen enkele wetenschap die al deze vragen zou kunnen beantwoorden en alle wetenschappen samen zullen waarschijnlijk maar een klein beetje kunnen verklaren. Maar het is een begin in het bevredigen van de nieuwsgierigheid van de mens. Metabolomics studies kunnen antwoord geven op vragen die gerelateerd zijn aan de reactie van een biologisch systeem op omgevingsinvloeden zoals bijvoorbeeld een toxicologische blootstelling, voeding of een medische behandeling. Toen metabolomics studies net in opkomst kwamen, waren multivariate technieken zoals Principale Componenten Analyse en Partial Least Squares regressie voldoende voor de statistische analyse. Maar tegenwoordig zijn de metabolomics studies veel gecompliceerder, waardoor de vraag naar betere en meer geavanceerde technieken voor data analyse toeneemt. Dit proefschrift is hier een bewijs van.

In functional genomics studies worden hoog-dimensionele data sets gegenereerd. Het is in veel gevallen handig als één getal de samenhang tussen twee hoog-dimensionele data sets kan beschrijven. Dit kan middels matrix correlaties. Deze correlatiemaat is prettig, omdat de interpretatie ervan overeen komt met de Pearson’s correlatie coefficient; een maat die bij velen bekend is. In geval van hoog-dimensionele data ontstaan er echter problemen voor de bestaande matrix correlatie maten. In hoofdstuk 2 wordt een verbetering van de RV-coefficient gegeven. Door een aanpassing in de berekening van de coefficient worden de problemen als gevolg van de hoge dimensies omzeild. Dit wordt aangetoond op basis van theorie, simulaties en de toepassing op twee echte functional genomics voorbeelden.

In hoofdstuk 3 wordt aangetoond dat een longitudinaal experimenteel design in combinatie met metabolomics en een meerweg data analyse een krachtige manier is om metabolierten aan te kunnen tonen, die een verschuiving over de tijd vertonen in de correlatie met de aanmaak van het bio-product. Er wordt een strategie gepresenteerd voor de data analyse van een longitudinaal microbiële metabolomics studie, die uitgevoerd was om metabolierten te identificeren die er voor zorgen dat de fenylalanine productie door *Escherichia coli* wordt geïnduceerd. Een gevalideerd meerweg statistisch model beschrijft de variatie in de productie van fenylalanine als functie van metabole verschillen. Het is opvallend dat de metabolierten die de sterkste samenhang vertoonden met de fenylalanine productie nauwelijks veranderden over de tijd.

Geavanceerde technieken zoals multilevel analyses zijn nodig om subtiele verschillen tussen gezonde personen en personen met een beginnende (multi-factoriële) ziekte te kunnen aantonen. Dit blijkt uit hoofdstuk 4. Multilevel analyses delen de data op naar verschillende submodellen, één model voor elk variatie niveau. Zo kunnen binnen- en tussenpersoonsvariantie gescheiden en apart geanalyseerd worden, mits de twee factoren
orthogonaal zijn. De voordelen van een multilevel aanpak in een meerweg analyse wordt beschreven voor de analyse van metabolomics data uit een dubbel blinde gerandomiseerde parallelle interventiestudie bij 20 mannen met licht overgewicht. Deze mannen kregen een negen-daagse behandeling met diclofenac of een placebo. De multilevel meerweg aanpak bleek veel krachtiger om de verschillen tussen de behandelingen te modeleren dan de minder geavanceerde standaard methode.

Uit hoofdstuk 5 blijkt dat de metaboliemen die een bijdrage leverden aan het verschil tussen de behandelingen uit de humane interventie studie niet alleen statistische relevant, maar ook biologisch relevant waren. De metabole verschillen waren subtiel en konden alleen opgepikt worden door de metabole profilering in combinatie met een orale glucose tolerantie test. De statistische power nam toe door de herhaalde metingen tijdens de glucose tolerantie test en door de metabole pertubatie werden metaboliemen ontdekt die verschillend reageerden op de glucose test. Verschillende metabole intermediairen uit de glutathion synthese vertoonden een tijdsafhankelijke onderdrukking als reactie op de orale glucose test. Het feit dat dit een insuline-gevoelig proces is, doet vermoeden dat de ontstekingsregulering mbv diclofenac in mannen met overgewicht de werking van het hormoon insulinine beïnvloed.

In hoofdstuk 6 wordt de prestatie van kruisvalidatie, jack-knifing en de permutatietest in geval van megavariate data onderzocht. Deze validatie methoden zijn bedoeld om een objectieve beoordeling te kunnen geven van een prestatie en stabilititeit van een statistisch model. Hoe kleiner het aantal objecten ten opzichte van het aantal variabelen, hoe minder de uitkomsten van deze validatie methoden te vertrouwen zijn. In dat geval hangen de resultaten puur af van de specifieke set van objecten die gebruikt was om het model te maken. Dit geeft aan dat de validatie methoden niet gebruikt kunnen worden als alarmerings mechanisme voor problemen als gevolg van de grootte of de representativiteit van de steekproef.

Of het nu het schatten van model parameters, het schatten van meta-parameters of het selecteren van variabelen betreft, elke stap in het modeleren van data die gebaseerd wordt op de data zelf, kost vrijheidsgraden. In hoofdstuk 7 wordt een dubbele kruisvalidatie aanpak gepresenteerd om de kosten van het schatten van de meta-parameter en het selecteren van variabelen expliciet in beeld te brengen. PLS regressie op twee echte metabolomics data sets illustreren het concept. Hieruit blijkt dat er een kleine prijs betaald moet worden voor het schatten van de meta-parameter, terwijl variabele selectie veel ‘duurder’ is. Voor een relatief kleine data set gaat bijna 35% van de vrijheidsgraden verloren aan het selecteren van variabelen. Gegeven het feit dat in functional genomics onderzoek het aantal objecten zelfs nog kleiner is dan de kleinste data set in dit hoofdstuk, maakt het negeren van het verlies aan vrijheidsgraden als gevolg van het selecteren van variabelen erg alarmerend.
NAWOORD

Carina M Rubingh

TNO Quality of Life, Zeist, The Netherlands
Wat een geluk...

.... dat ik zoveel mensen om me heen mag kennen, die me de afgelopen jaren gesteund hebben bij het bereiken van deze mijlpaal.
The publication of this thesis was made possible by the Netherlands Bioinformatics Centre (NBIC)