Crossing borders between biology and data analysis

Robert van den Berg

Crossing borders between biology and data analysis

Crossing borders between biology and data analysis

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 Aula der Universiteit
op woensdag 24 september, te 14:00 uur

door

Robert Anthonius van den Berg

geboren te Woerden

Promotiecommissie

Promotor:	Prof. dr. A. K. Smilde
Co-promotores:	Dr. ir. M. J. van der Werf
	Dr. J. A. Westerhuis
Overige Leden:	Prof. dr. R. J. M. M. Does
	Prof. dr. R. Goodacre
	Prof. dr. K. J. Hellingwerf
	Prof. dr. A. H. C. van Kampen
	Prof. dr. O. P. Kuipers
	Prof. dr. I. Van Mechelen

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Table of contents

1 Introduction

Key factors for successful top down systems biology in biotechnology

Robert A. van den Berg, Age K. Smilde, Johan A. Westerhuis, Mariët J. van der Werf

1 Introduction

Systems biology involves the study of biological systems by approaching them as an integrated and interacting network of genes, proteins, metabolites, and biochemical reactions. The biological system studied can be, for example, a cell, an organ, or a whole organism. By modeling the interacting network, systems biology attempts to identify the underlying mechanisms that influence the behavior and functionality of the biological system.

1.1 Systems biology philosophies

Within systems biology there are different philosophies with regard to the modeling of a biological system [1,2]. In bottom up systems biology [1,3], biological systems are modeled based on knowledge about, for instance, the genome, metabolic networks [4,5], or reaction kinetics [6,7]. Based on the integration of the behavior of the individual components, predictions regarding the behavior of the biological system are made.

Top down systems biology models are developed based on the direct measurements of the response of the studied biological system to experimental conditions to which the biological system is subjected. Unlike bottom up systems biology, top down systems biology does not require mechanistic assumptions regarding the interactions of the studied biomolecules. The system wide response on different biological levels (e.g. transcriptomics, proteomics, or metabolomics) of the biological system to the applied experimental conditions is measured with advanced analytical techniques like, for instance, micro-arrays [8] or GC/LCMS [9,10] methods. Changes of the levels of the measured biomolecules in response to different experimental conditions are modeled with advanced data analysis methods. These data analysis methods search for trends in the behavior of the measured biomolecules related to a certain biological question following a 'guilt by association' approach. An example of a top down systems biology biological question is, for instance: the behavior of which biomolecules is associated with high and low biomass yield.

Furthermore, middle-out systems biology was proposed [2] as a pragmatic approach in which depending on available data top down and bottom up systems biology approaches are combined to explore the biological system. An example of a method that integrates knowledge of pathway topology with the data analysis of transcriptomics data is network component analysis [11]. Furthermore, grey component analysis [12] could be suited for this purpose.

1.2 Advantages of top down systems biology

Currently, most papers published on systems biology research are based on a bottom up systems biology approach. In our opinion, this does not do justice to the advantages of top down systems biology. First, the choice for experimental conditions in top down systems biology studies are not limited by model assumptions and therefore the experimental conditions can be selected to target the biological question as directly as possible. In contrast, bottom up systems biology models often require assumptions that limit the experimental conditions that can be studied. An example of such an assumption in the case of microbial flux balance models is the steady state assumption for the studied biological system; this assumption is rarely met for, e.g. batch or fed-batch based industrial microbiology processes.

Second, there is no *a priori* focus on specific biomolecules that should be related to the biological question. This enables the discovery of previously unknown or unexpected relations between the behavior of the biomolecules and the biological question. This advantage applies as well to genome wide bottom up systems biology models.

Third, whilst bottom up systems biology requires extensive knowledge of the studied organism, top down systems biology does not have this requirement. Hence top down systems biology is more generic in nature and can also be applied to relatively poor characterized strains, such as, recently discovered micro-organisms, or strains obtained after UV-mutagenesis.

1.3 Top down systems biology challenges

Top down systems biology studies typically involve the generation and analysis of large data sets. Consequently, the success rate of a top down systems biology study is highly dependent on the information richness of the data and therefore on the design and execution of the study. To successfully extract information relevant to the biological question, the experimental –omics data needs to contain information relevant to this question. Therefore, experimental conditions have to be carefully selected to induce the phenomena of interest and to capture these phenomena in the samples. Furthermore, the data analysis should be able (i) to extract information relevant to the biological question from the experimental data, and (ii) to statistically validate the extracted information. The information obtained in this way is used to address the biological question and to plan the next step in the study, e.g. biological validation.

In a top down systems biology study, the three factors (i) biological question, (ii) experimental design, and (iii) data analysis represent three crucial strongly interdependent aspects. In this Chapter, we discuss the importance of these different aspects and their mutual dependence as requirements for executing successful top down systems biology studies. We will illustrate our discussion using three different microbiological questions (Table 1).

2 Crucial aspects of top down systems biology

The three different aspects, biological question, experimental design, and data analysis together are the corner stones of top down systems biology. The relations between these

Biological question	Experimental design	Data analysis method
Case I		
What are the differences at the metabolic level between a wild type and an overproducing strain?	Distinguish strain specific behavior from normal variation within a strain by independent repeated experiments with the wild type and the overproducing strain.	Classification method Statistical validation targeted on reliability of metabolite contributions
Case II		
Which biomolecules are associated with bioproduct formation?	Environmental conditions that result in an evenly distributed range of bioproduct yields, titer, or productivity. Biological level (proteome, metabolome)	Regression model that associates the behavior of the biomolecule yield with the -omics data. Statistical validation based on reliability of biomolecule contributions.
Case III		
Which biomolecules are regulated by the same regulatory mechanisms?	Select experimental conditions that induce the regulatory effects of interest. Biological level (transcriptome, proteome, metabolome)	Select data analysis approach based on the behavior expected from biomolecules which are subjected to the same regulatory mechanism.

Table 1 – Examples of biological questions in relation to experimental design and data analysis method considerations. Here, three cases of biological questions with a selection of their specific considerations for experimental design and data analysis method are presented.

aspects are visualized in the top down systems biology research triangle (Figure 1). Analogous to systems biology, this research triangle is an interlinked network of which the individual factors are difficult to separate.

2.1 The biological question

2.1.1 Articulating essential aspects of the biological question

The biological question is the start and the end of a top down systems biology study. It can be very specific: Which biomolecules are related to bioproduct formation (Table I, case II)?; or broad: How do regulatory effects manifest themselves in the behavior of biomolecules (Table I, case III)?. The original biological question, often stated in generic terms, has to be made operational and translated into an experimental design and a data analysis strategy. The biological question is made operational by articulating essential aspects of the research in the biological question and by preventing implicit assumptions about these essential aspects. Often experimental design and data analysis choices follow naturally from the articulation of these essential aspects. For instance, in case III (Table I, Which biomolecules are regulated by the same regulatory mechanisms?) the biological question can be made operational by (i) identifying under which experimental conditions the regulatory mechanisms become activated, i.e. by literature searches or screenings experiments, and (ii) how these regulatory mechanisms are expected to manifest themselves in the behavior of the measured biomolecules. The first will affect the choices for the experimental design, and the latter is important for the data analysis strategy. Making the biological question operational will facilitate the identification of key aspects for all factors of the top down systems biology research triangle.

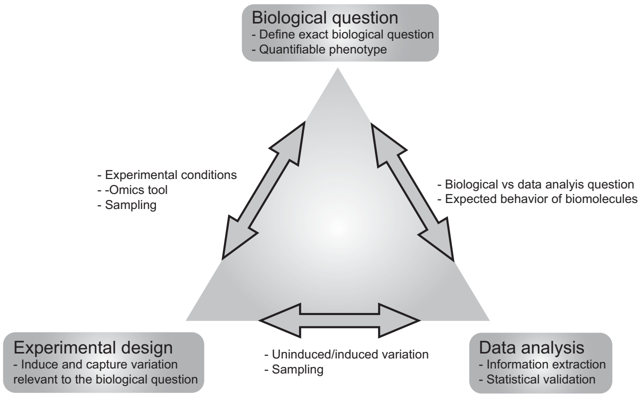

Biological question
- Define exact biological question
- Quantifiable phenotype
- Experimental conditions
- -Omics tool
- Sampling
- Biological vs data analyis question
- Expected behavior of biomolecules
Experimental design
- Induce and capture variation relevant to the biological question
- Uninduced/induced variation
- Sampling
Data analysis
- Information extraction
- Statistical validation

Figure 1 – Top down systems biology research triangle. In top down systems biology three interlinked factors are crucial: (i) the biological question, (ii) experimental design, and (iii) data analysis.

2.1.2 Utilization of a quantifiable phenotype

Often, it is possible to relate the biological question to a quantifiable phenotypic parameter. A quantifiable phenotypic parameter aids the selection of experimental conditions and helps to focus the data analysis, since variation in the behavior of the phenotypic parameter is directly related to the biological question. In case II, the quantifiable phenotype is bioproduct yield. Other quantifiable parameters which can be valuable are, for example, growth rate, or acid/base consumption for pH control. Also qualitative information like strain type, or morphology characteristics can be utilized -especially in the data analysis- by classifying this type of information and applying the class information in the study.

2.2 Experimental design

Top down systems biology is based on the statistical modeling with advanced data analysis tools of experimental –omics data. The goal of experimental design is therefore: to translate the biological question into an experimental procedure which results in –omics data containing information which can be accessed and validated with suited data analysis tools. Whereas statistical validation cannot replace biological validation, it is still very important to validate the performance of the data analysis methods (see section 2.3.2).

Experimental design is not limited to designing traditional factorial designs [13-15] in which a set of experimental parameters is systematically varied, but it involves a range of choices which will be discussed below.

2.2.1 Experimental conditions

Determining the experimental conditions relevant for the biological question is based on biological knowledge from literature and prior knowledge. Sometimes it is unclear what experimental factors, such as nutrients or pH, are important for the induction of variation in the behavior of biomolecules relevant for the biological problem. Then, screenings experiments can be conducted to obtain more information regarding the importance of these experimental parameters. The availability of a quantifiable phenotypic parameter aids the selection of the experimental conditions as the experimental conditions should induce variation in this parameter.

2.2.2 -omics tool

For a top down systems biology study, the biological system can be measured on different levels in the cellular organization, e.g. the metabolome, proteome, or transcriptome. Selection of an –omics tool is therefore an important design consideration and the choice depends heavily on which biochemical level the biological phenomena relevant for the biological question occur, or which biochemical level is studied. In case III, the selection of the -omics tool determines the regulatory mechanisms which can be identified, such as, transcriptional or allosteric regulation. The biological question for case III should therefore specify on what level or levels the relevant regulatory mechanisms should be studied.

2.2.3 Sample collection

The sampling procedure determines when and how (many) samples for the -omics analysis are taken. It should ensure that the biological phenomena relevant to the biological question, e.g. the onset of a regulation event, are captured in the collected samples and that degradation of the sample is prevented [16,17]. Furthermore, the number of samples influences the performance of data analysis methods [18]. Determining a sampling scheme for a certain biological question is finding a trade off between four aspects: (i) sample collection considerations to capture the relevant biological phenomena; (ii) balance between exploring new experimental conditions and firmly establishing a few experimental conditions; (iii) the increased performance of data analysis methods for increased number of samples; and (iv) the costs.

2.2.3.1 Capturing the relevant biological phenomena

Depending on the biological question and the selected –omics tool, it is not necessarily straightforward to capture the biological phenomena of interest in the samples. In the example of case II, it is beforehand unknown which phases during a batch fermentation process contain information related to the bioproduct yield at the end of the process. When this is not known from literature or screenings experiments, the sampling protocol should cover different growth phases and phase transitions. Here, other practical issues can play a role as well. For instance, the sampling volumes can limit the number of samples that can be collected, but also sample work up considerations can influence the sample collection.

2.2.3.2 Exploring new conditions or firmly establishing the most important conditions?

The importance of repeated measurements, or the establishment of the variability of the measured biomolecules in response to a certain experimental condition, depends on the biological question. For instance, in case I it is essential to distinguish differences in the behavior of relevant biomolecules between strains (induced biological variation, Figure 2) from differences which can occur within repeated measurements of the same strain (uninduced biological variation, Figure 2). The characteristics of the uninduced variation is estimated from biologically independent repeated measurements and more repeated measurements increase the reliability of the estimation of the uninduced variation. On the other hand, for case II it could be more beneficial to include new experimental conditions possibly important for inducing variation in bioproduct formation over firmly establishing repeatability of a few variation inducing conditions. Generally speaking, it is good to keep in mind that the induced biological variation (Figure 2) is often larger than the uninduced biological variation and the technical variation [19]. This is mostly due to the selection of experimental conditions which induce variation in the behavior of the relevant biomolecules.

2.2.3.3 Improved performance data analysis methods

The performance of data analysis methods benefit from increased sample numbers. Increased sample numbers improve the reliability of the estimated contributions of the analyzed biomolecules to the modeled behavior. In case I, for example, more samples will improve the reliability of the estimation of the contributions of the measured metabolites to the differences between the wild type and the overproducing strain. While for univariate data analysis (i.e. considering the behavior in different samples of each biomolecule without taking into account interactions between biomolecules) different articles [20-23] are

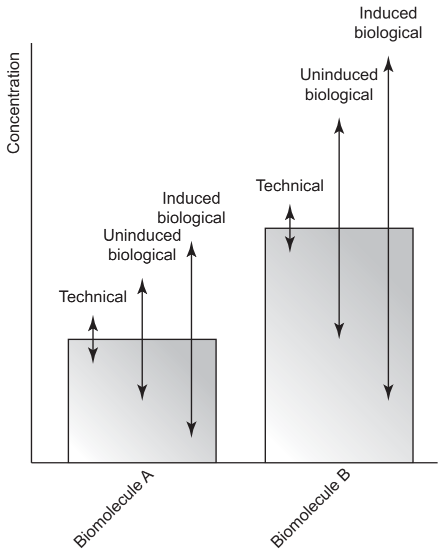


Figure 2 – Different levels of variation present in –omics data. The total variation in the behavior of a biomolecule in a top down systems biology data set is the sum of technical, uninduced biological, and induced biological variation. Technical variation originates from the technical procedure. Uninduced biological variation originates from biological variability between conditions. It can differ from biomolecule to biomolecule (biomolecule A from biomolecule B) as well as from condition to condition. Induced biological variation is the variation induced by the experimental design. Ideally, the induced biological variation of a biomolecule is much larger than its uninduced biological and technical variation.

published which discuss means to estimate sample sizes; for multivariate data analysis (i.e. the analysis of the behavior in samples of biomolecules in relation to the behavior of the other measured biomolecules) it is not yet clear how to determine sample size [18]. The more the better is currently the best recommendation [18].

2.3 Data analysis

A data analysis strategy is based on (i) the biological question, (ii) properties of the data set (e.g. number of experiments/variables, time series, et cetera), and (iii) how information relevant to the biological question is expected to manifest itself in the data set. Furthermore, statistical validation of the data analysis will provide an indication of the significance of the identified effects. There are many data analysis methods available which are suited to address different biological questions (Table 1). For instance, for case I, classification methods such as partial least squares discriminant analysis (PLS-DA) [24] or principal component discriminant analysis (PCDA) [25] can build models based on the differences between the wild type and the overproducing strain. The data resulting from the experiments for case II can be analyzed with regression methods like PLS [26] or principal component regression (PCR) [27] that can relate the behavior of the measured biomolecules to the behavior of the phenotype parameter.

2.3.1 Translation of the biological question into the expected behavior of biomolecules

Different data analysis methods interpret the behavior of the measured biomolecules differently. It is therefore important to translate the biological question into the expected behavior of the biomolecules as perceived by the data analysis method. When factors like the abundance of a biomolecule or the magnitude of the fold change are not important for the biological question, it can be necessary to correct for the influence of these factors [19]. Data pretreatment methods can be applied to emphasize those aspects of the variation that are important for the specific biological question and for the specific properties of the selected data analysis method.

2.3.2 Statistical validation of the data analysis

An important part of the data analysis is the validation of the data analysis results. By validating the results, it can be assessed how well the obtained results compare to chance correlations, or if the found model is too optimistic due to overfitting. Overfitting means that besides induced biological variation also variation unrelated to the experimental design and thus the biological question, is captured in the model. As a result, overfitting reduces the generic applicability of the top down systems biology model and hence should be prevented. Frequently applied data analysis validation strategies in top down systems biology are cross validation, permutation, jackknife, and bootstrapping [18,28-30]. The question what is to be validated determines the validation approach since validating contributions of individual biomolecules to a classification model is different from validating the classification performance of the same model.

Statistical validation aids the interpretation of the data analysis results by providing indications of the limitations of the results of the applied data analysis tool. In this way, statistical validation can guide the selection of biomolecules important for the biological question. It cannot, however, replace biological validation of the leads to answering the biological question found with a top down systems biology approach.

3 Conclusions

Top down systems biology is a potentially suited research approach for many issues in biotechnology, such as, finding targets for strain improvement, medium optimization, or analyzing regulatory questions (e.g. protease induction [31]). Moreover, it does not require extensive knowledge regarding the studied organism; it is flexible with regard to the environmental conditions which can be studied (e.g. no steady state assumption); and the data analysis gives an open view on possible important biomolecules.

Top down systems biology studies, however, require large information-rich data sets for the modeling of the biological systems and for providing answers to research questions. Therefore, it is essential to carefully consider the three corner points of the top down systems biology research triangle: biological question, experimental design, data analysis, and their interdependence. The impact of choices made within one of the three corner points is not limited to the respective corner point itself, but extends to the other points as well.

In our opinion, parts of the top down systems biology research triangle are often neglected, which leads to –omics data sets without a clear biological question, or a clear underlying experimental design. Consequently, it is very difficult to extract biologically relevant information from these data sets and the conducted experiments in turn can be considered a loss of effort and resources. The framework of the top down systems biology research triangle can help improve the setup of top down systems biology studies and improve the success rate of top down systems biology research projects.

In Chapters 2 to 6, different aspects of the relation between a biological question and a data analysis strategy, such as, data pretreatment and selection of the most suited data analysis method, are further explored.

4 References

1. Bruggeman FJ, Westerhoff HV: **The nature of systems biology.** *Trends Microbiol* 2007, **15:**45-50.
2. Noble D: **The rise of computational biology.** *Nature Reviews Molecular Cell Biology* 2002, **3:**459-463.
3. Teusink B, Smid EJ: **Modelling strategies for the industrial exploitation of lactic acid bacteria.** *Nat Rev Micro* 2006, **4:**46-56.
4. Reed JL, Vo TD, Schilling CH, Palsson BO: **An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR).** *Genome Biol* 2003, **4:**R54.
5. Schuetz R, Kuepfer L, Sauer U: **Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli.** *Mol Syst Biol* 2007, **3**.
6. Snoep JL, Bruggeman F, Olivier BG, Westerhoff HV: **Towards building the silicon cell: A modular approach.** *Bio Systems* 2006, **83:**207-216.
7. Jamshidi N, Edwards JS, Fahland T, Church GM, Palsson BO: **Dynamic simulation of the human red blood cell metabolic network.** *Bioinformatics* 2001, **17:**286-287.
8. Quackenbush J: **Microarrays--Guilt by Association.** *Science* 2003, **302:**240-241.
9. Romijn EP, Krijgsveld J, Heck AJR: **Recent liquid chromatographic-(tandem) mass spectrometric applications in proteomics.** *Journal of Chromatography A* 2003, **1000:**589-608.
10. van der Werf MJ, Overkamp KM, Muilwijk B, Coulier L, Hankemeier T: **Microbial metabolomics: Toward a platform with full metabolome coverage.** *Anal Biochem* 2007, **370:**17-25.
11. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP: **Network component analysis: Reconstruction of regulatory signals in biological systems.** *Proc Natl Acad Sci USA* 2003, **100:**15522-15527.
12. Westerhuis JA, Derks EPPA, Hoefsloot HCJ, Smilde AK: **Grey component analysis.** *J Chemom* 2007, **21:**474-485.
13. Kerr MK: **Design Considerations for Efficient and Effective Microarray Studies.** *Biometrics* 2003, **59:**822-828.
14. Vandeginste BGM, Massart DL, Buydens LMC, Jong Sd, Lewi PJ, Smeyers-Verbeke J: *Handbook of chemometrics.* Amsterdam: Elsevier; 1998.
15. Churchill GA: **Fundamentals of experimental design for cDNA microarrays.** *Nat Genet* 2002, **32:**490-495.
16. Pieterse B, Jellema RH, van der Werf MJ: **Quenching of microbial samples for increased reliability of microarray data.** *J Microbiol Methods* 2006, **64:**207-216.
17. de Koning W, van Dam K: **A method for the determination of changes of**

glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral pH. *Anal Biochem* 1992, **204:**118-123.

18. Rubingh CM, Bijlsma S, Derks EPPA, Bobeldijk I, Verheij ER, Kochhar S, Smilde AK: **Assessing the performance of statistical validation tools for megavariate metabolomics data.** *Metabolomics* 2006, **2:**53-61.
19. van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ: **Centering, scaling, and transformations: improving the biological information content of metabolomics data.** *BMC Genomics* 2006, **7**.
20. Tibshirani R: **A simple method for assessing sample sizes in microarray experiments.** *BMC Bioinformatics* 2006, **7:**106.
21. Ferreira JA, Zwinderman A: **Approximate Sample Size Calculations with Microarray Data: An Illustration.** *Statistical Applications in Genetics and Molecular Biology* 2006, **5:**25.
22. Dobbin K, Simon R: **Sample size determination in microarray experiments for class comparison and prognostic classification.** *Biostat* 2005, **6:**27-38.
23. Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus.** *Nat Rev Genet* 2006, **7:**55-65.
24. Barker M, Rayens W: **Partial least squares for discrimination.** *J Chemom* 2003, **17:**166-173.
25. Hoogerbrugge R, Willig SJ, Kistemaker PG: **Discriminant Analysis by Double Stage Principal Component Analysis.** *Anal Chem* 1983, **55:**1710-1712.
26. Geladi P, Kowalski BR: **Partial least-squares regression: a tutorial.** *Anal Chim Acta* 1986, **185:**1-17.
27. Mardia KV, Kent JT, Bibby JM: *Multivariate Analysis.* Academic Press; 1979.
28. Martens H, Martens M: **Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR).** *Food Quality and Preference* 2000, **11:**5-16.
29. Efron B, Tibshirani RJ: *An Introduction to the Bootstrap.* New York: Chapman & Hall; 1993.
30. Westerhuis J, Hoefsloot H, Smit S, Vis D, Smilde A, van Velzen E, van Duijnhoven J, van Dorsten F: **Assessment of PLSDA cross validation.** *Metabolomics* 2008, **4:**81-89.
31. Braaksma M, Punt PJ: ***Aspergillus* as a cell factory for protein production: controlling protease activity in fungal production.** In *The Aspergilli. Genomics, Medical Aspects, Biotechnology, and Research Methods*. Edited by Edited by Goldman GH, Osmani SA. Boca Raton: CRC Press; 2008.

2 Centering, scaling, and transformations: improving the biological information content of metabolomics data

Robert A. van den Berg[1*], Huub C. J. Hoefsloot[2], Johan A. Westerhuis[2], Age K. Smilde[1,2] and Mariët J. van der Werf[1]

Summary

Extracting relevant biological information from large data sets is a major challenge in functional genomics research. Different aspects of the data hamper their biological interpretation. For instance, 5000-fold differences in concentration for different metabolites are present in a metabolomics data set, while these differences are not proportional to the biological relevance of these metabolites. However, data analysis methods are not able to make this distinction. Data pretreatment methods can correct for aspects that hinder the biological interpretation of metabolomics data sets by emphasizing the biological information in the data set and thus improving their biological interpretability.

Different data pretreatment methods, i.e. centering, autoscaling, pareto scaling, range scaling, vast scaling, log transformation, and power transformation, were tested on a real-life metabolomics data set. They were found to greatly affect the outcome of the data analysis and thus the rank of the, from a biological point of view, most important metabolites. Furthermore, the stability of the rank, the influence of technical errors on data analysis, and the preference of data analysis methods for selecting highly abundant metabolites were affected by the data pretreatment method used prior to data analysis.

Different pretreatment methods emphasize different aspects of the data and each pretreatment method has its own merits and drawbacks. The choice for a pretreatment method depends on the biological question to be answered, the properties of the data set and the data analysis method selected. For the explorative analysis of the validation data set used in this study, autoscaling and range scaling performed better than the other pretreatment methods. That is, range scaling and autoscaling were able to remove the dependence of the rank of the metabolites on the average concentration and the magnitude of the fold changes and showed biologically sensible results after PCA (principal component analysis). In conclusion, selecting a proper data pretreatment method is an essential step in the analysis of metabolomics data and greatly affects the metabolites that are identified to be the most important.

This chapter is published as R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde and M. J. van der Werf. BMC Genomics, 2006, 7 (142).

1 Background

Functional genomics approaches are increasingly being used for the elucidation of complex biological questions with applications that range from human health [1] to microbial strain improvement [2]. Functional genomics tools have in common that they aim to measure the complete biomolecule response of an organism to the environmental conditions of interest. While transcriptomics and proteomics aim to measure all mRNA and proteins, respectively, metabolomics aims to measure all metabolites [3,4].

In metabolomics research, there are several steps between the sampling of the biological condition under study and the biological interpretation of the results of the data analysis (Figure 1). First, the biological samples are extracted and prepared for analysis. Subsequently, different data preprocessing steps [3,5] are applied in order to generate 'clean' data in the form of normalized peak areas that reflect the (intracellular) metabolite concentrations. These clean data can be used as the input for data analysis. However, it is important to use an appropriate data pretreatment method before starting data analysis. Data pretreatment methods convert the clean data to a different scale (for instance, relative or logarithmic scale). Hereby, they aim to focus on the relevant (biological) information and to reduce the influence of disturbing factors such as measurement noise. Procedures that can be used for data pretreatment are scaling, centering and transformations.

In this paper, we discuss different properties of metabolomics data, how pretreatment methods influence these properties, and how the effects of the data pretreatment methods can be analyzed. The effect of data pretreatment will be illustrated by the application of eight data pretreatment methods to a metabolomics data set of *Pseudomonas putida* S12 grown on four different carbon sources.

1.1 Properties of metabolome data

In metabolomics experiments, a snapshot of the metabolome is obtained that reflects the cellular state, or phenotype, under the experimental conditions studied [3]. The experiments that resulted in the data set used in this paper were conducted according to an experimental design. In an experimental design, the experimental conditions are purposely chosen to induce variation in the

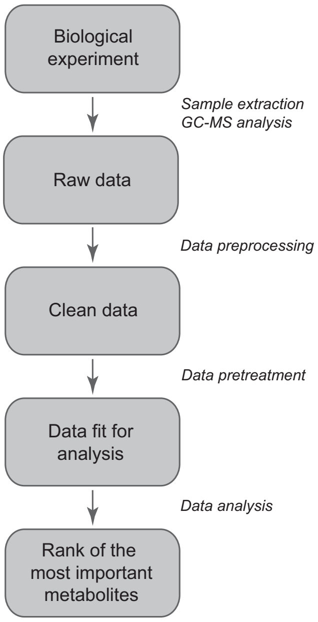


Figure 1 - The different steps between biological sampling and ranking of the most important metabolites.

area of interest. The resulting variation in the metabolome is called induced biological variation.

However, other factors are also present in metabolomics data:

1. Differences in orders of magnitude between measured metabolite concentrations; for example, the average concentration of a signal molecule is much lower than the average concentration of a highly abundant compound like ATP. However, from a biological point of view, metabolites present in high concentrations are not necessarily more important than those present at low concentrations.
2. Differences in the fold changes in metabolite concentration due to the induced variation; the concentrations of metabolites in the central metabolism are generally relatively constant, while the concentrations of metabolites that are present in pathways of the secondary metabolism usually show much larger differences in concentration depending on the environmental conditions.
3. Some metabolites show large fluctuations in concentration under identical experimental conditions. This is called uninduced biological variation.

Besides these biological factors, other effects present in the data set are:

4. Technical variation; this originates from, for instance, sampling, sample work-up and analytical errors.
5. Heteroscedasticity; for data analysis, it is often assumed that the total uninduced variation resulting from biology, sampling, and analytical measurements is symmetric around zero with equal standard deviations. However, this assumption is generally not true. For instance, the standard deviation due to uninduced biological variation depends on the average value of the measurement. This is called heteroscedasticity, and it results in the introduction of additional structure in the data [6,7]. Heteroscedasticity occurs in uninduced biological variation as well as in technical variation.

The variation in the data resulting from a metabolomics experiment is the sum of the induced variation and the total uninduced variation. The total uninduced variation is all the variation originating from uninduced biological variation, sampling, sample work-up, and analytical variation. Data pretreatment focuses on the biologically relevant information by emphasizing different aspects in the clean data, for instance, the metabolite concentration under a growth condition relative to the average concentration, or relative to the biological range of that metabolite. In metabolomics, data pretreatment relates the differences in metabolite concentrations in the different samples to differences in the phenotypes of the cells from which these samples were obtained [3].

1.2 Data pretreatment methods

The choice for a data pretreatment method does not only depend on the biological information to be obtained, but also on the data analysis method chosen since different data analysis methods focus on different aspects of the data. For example, a clustering method focuses on the analysis of (dis)similarities, whereas principal component analysis (PCA) attempts to explain as much variation as possible in as few components as possible. Changing data properties using data pretreatment may therefore enhance the results of a clustering method, while obscuring the results of a PCA analysis.

In this paper, we discuss three classes of data pretreatment methods: (I) centering, (II) scaling and (III) transformations (Table 1).

1.2.1 Class I: Centering

Centering converts all the concentrations to fluctuations around zero instead of around the mean of the metabolite concentrations. Hereby, it adjusts for differences in the offset between high and low abundant metabolites. It is therefore used to focus on the fluctuating part of the data [8,9], and leaves only the relevant variation (being the variation between the samples) for analysis. Centering is applied in combination with all the methods described below.

1.2.2 Class II: Scaling

Scaling methods are data pretreatment approaches that divide each variable by a factor, the scaling factor, which is different for each variable. They aim to adjust for the differences in fold differences between the different metabolites by converting the data into differences in concentration relative to the scaling factor. This often results in the inflation of small values, which can have an undesirable side effect as the influence of the measurement error, that is usually relatively large for small values, is increased as well.

There are two subclasses within scaling. The first class uses a measure of the data dispersion (such as, the standard deviation) as a scaling factor, while the second class uses a size measure (for instance, the mean).

1.2.2.1 Scaling based on data dispersion

Scaling methods tested that use a dispersion measure for scaling were autoscaling [9], pareto scaling [10], range scaling [11], and vast scaling [12] (Table 1). Autoscaling, also called unit, or unit variance scaling, is commonly applied and uses the standard deviation as the scaling factor [9]. After autoscaling, all metabolites have a standard deviation of one and therefore the data is analyzed on the basis of correlations instead of covariances, as is the case with centering.

Pareto scaling [10] is very similar to autoscaling. However, instead of the standard deviation, the square root of the standard deviation is used as the scaling factor. Now, large fold changes are decreased more than small fold changes, thus the large fold changes are less dominant compared to clean data. Furthermore, the data does not become dimensionless as after autoscaling (Table 1).

Vast scaling [12] is an acronym of *va*riable *st*ability scaling and it is an extension of autoscaling. It focuses on stable variables, the variables that do not show strong variation, using the standard deviation and the so-called coefficient of variation (cv) as scaling factors (Table 1). The cv is defined as the ratio of the standard deviation and the mean: $\frac{s_i}{\bar{x}_i}$. The use of the cv results in a higher importance for metabolites with a small relative standard deviation and a lower importance for metabolites with a large relative standard deviation. Vast scaling can be used unsupervised as well as supervised. When vast scaling is applied as a supervised method, group information about the samples is used to determine group specific cvs for scaling.

The scaling methods described above use the standard deviation or an associated measure as scaling factor. The standard deviation is, within statistics, a commonly used entity to measure the data spread. In biology, however, a different measure for data spread might be useful as well, namely the biological range. The biological range is the difference between the minimal and the maximal concentration reached by a certain metabolite in a set of experiments. Range scaling [11] uses this biological range as the scaling factor (Table 1). A disadvantage of range scaling with regard to the other scaling methods tested is that only two values are used to estimate the biological range, while for the standard deviation all measurements are taken into account. This makes range scaling more sensitive to outliers. To increase the robustness of range scaling, the range could also be determined by using robust range estimators.

1.2.2.2 Scaling based on average value

Level scaling falls in the second subclass of scaling methods, which use a size measure instead of a spread measure for the scaling. Level scaling converts the changes in metabolite concentrations into changes relative to the average concentration of the metabolite by using the mean concentration as the scaling factor. The resulting values are changes in percentages compared to the mean concentration. As a more robust alternative, the median could be used. Level scaling can be used when large relative changes are of specific biological interest, for example, when stress responses are studied or when aiming to identify relatively abundant biomarkers.

1.2.3 Class III: Transformations

Transformations are nonlinear conversions of the data like, for instance, the log transformation and the power transformation (Table 1). Transformations are generally applied to correct for heteroscedasticity [7], to convert multiplicative relations into additive relations, and to make skewed distributions (more) symmetric. In biology, relations between variables are not necessarily additive but can also be multiplicative [13]. A transformation is then necessary to identify such a relation with linear techniques.

Since the log transformation and the power transformation reduce large values in the data set relatively more than the small values, the transformations have a pseudo scaling effect as differences between large and small values in the data are reduced. However, the pseudo scaling effect is not determined by the multiplication with a scaling factor as for a 'real' scaling effect, but by the effect that these transformations have on the original values. This pseudo scaling effect is therefore rarely sufficient to fully adjust for magnitude differences. Hence, it can be useful to apply a scaling method after the transformation. However, it is not clear how the transformation and a scaling method influence each other with regard to the complex metabolomics data.

A transformation that is often used is the log transformation (Table 1). A log transformation perfectly removes heteroscedasticity if the relative standard deviation is constant [7]. However, this is rarely the case in real life situations. A drawback of the log transformation is that it is unable to deal with the value zero. Furthermore, its effect on values with a large relative analytical standard deviation, usually the metabolites with a relatively low concentration, is problematic as these deviations are emphasized. These problems occur because the log transformation approaches minus infinity when the value to be transformed approaches zero.

A transformation that does not show these problems and also has positive effects on heteroscedasticity is the power transformation (Table 1) [13]. The power transformation shows a similar transformation pattern as the log transformation. Hence, the power transformation can be used to obtain results similar as after the log transformation without the near zero artifacts, although the power transformation is not able to make multiplicative effects additive.

2 Methods

2.1 Background of the data set

P. putida S12 [14] is maintained at TNO. Cultures of *P. putida* S12 were grown in batch fermentations at 30°C in a Bioflow II (New Brunswick Scientific) bioreactor as previously described by van der Werf [15]. Samples (250 ml) were taken from the bioreactor at an OD 600 of 10. Cells were immediately quenched at -45°C in methanol as described previously [16]. Prior to extracting the intracellular metabolites from the cells - by chloroform extraction at –45°C [17] - internal standards were added [18] and a sample was taken for biomass determination [19]. Subsequently, the samples were lyophilized.

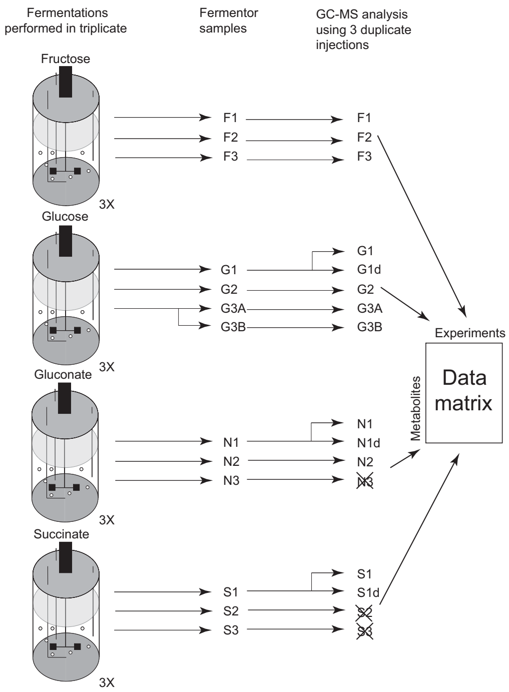


Figure 2 - Experimental design. The fermentations were performed in independent triplicates. Of the third glucose fermentation a sample was taken in duplicate and of G1, N1 and S1 the samples were analyzed in duplicate by GC-MS. The samples of N3, S2 and S3 were not taken into account in this study.

2.2 GC-MS analysis

Lyophilized metabolome samples were derivatized using a solution of ethoxyamine hydrochloride in pyridine as the oximation reagent followed by silylation with N-trimethyl-N-trimethylsilylacetamide as described by [18]. GC-MS-analysis of the derivatized samples was performed using temperature gradient from 70°C to 320°C at a rate of 10°C/min on an Agilent 6890 N GC (Palo Alto, CA, USA) and an Agilent 5973 mass selective detector. 1 µl aliquots of the derivatized samples were injected in the splitless mode on a DB5-MS capillary column. Detection was performed using MS detection in electron impact mode (70 eV).

2.3 Data preprocessing

The data from GC-MS analyses were deconvoluted using the AMDIS spectral deconvolution software package [18,20]. Zeros in the data set were replaced with small values equal to MS peak areas of 1 to allow for log transformations. The lowest peak areas in the rest of the data are in the order of 10^3. The output of the AMDIS analysis, in the form of peak identifiers and peak areas, was corrected for the recovery of internal standards and normalized with respect to biomass. The peaks resulting from a known compound were combined. The samples N3, S2 and S3 were removed from the data set, as a different sample workup protocol was followed. Furthermore, metabolites detected only once in the 13 remaining experiments were removed. This lead to a reduced data set consisting of 13 experiments and 140 variables expressed as peak areas in arbitrary units (Figure 2). This data set was used as the clean data for data pretreatment.

2.4 Data pretreatment

Data pretreatment and PCA were performed using Matlab 7 [21], the PLS Toolbox 3.0 [22], and home written m-files. Data pretreatment was applied according to the formulas in Table 1. The notation of the formulas is as follows: Matrices are presented in bold uppercase $(\mathbf{X})$, vectors in bold lowercase $(\mathbf{t})$, and scalars are given in lowercase italic (a) or uppercase italic in case of the end of a series $i = 1...I$. The data is presented in a data matrix **X** (*I x J*) with *I* rows referring to the metabolites and *J* columns referring to the different conditions. Element x_{ij} therefore holds the measurement of metabolite *i* in experiment *j*.

Vast scaling was applied unsupervised as the other data pretreatment methods were unsupervised as well.

2.5 Data analysis

PCA was applied for the analysis of the data. PCA decomposes the variation of matrix **X** into scores **T**, loadings **P**, and a residuals matrix **E**. **P** is an *I x A* matrix containing the *A* selected loadings and **T** is a *J x A* matrix containing the accompanying scores.

$$\mathbf{X} = \mathbf{P}\mathbf{T}^{T} + \mathbf{E},$$

where $\mathbf{P}^{\mathbf{T}}\mathbf{P} = \mathbf{I}$, the identity matrix.

The number of components used (*A*) in the PCA analysis was based on the scree plots and the score plots.

For ranking of the metabolites according to importance for the *A* selected PCs, the contribution *r* of all the variables to the effects observed in the *A* PCs was calculated

$$r_{Ai} = \sum_{a=1}^{A} \lambda_a^2 \cdot p_{ia}^2$$

Here, *r* is the contribution of variable *i* to *A* components, λ_a is the singular value for the *a*th PC and p_{ia} is the value for the *i*th variable in the loading vector belonging to the *a*th PC. To allow for comparison between the different data pretreatment methods, the values for r_A were sorted in descending order after which the comparisons were performed using the rank of the metabolite in the sorted list.

Table 1: Overview of the pretreatment methods used in this study. In the Unit column, the unit of the data after the data pretreatment is stated. O represents the original Unit, and (-) presents dimensionless data. The mean is estimated as: $\bar{x}_i = \frac{1}{J}\sum_{j=1}^{J} x_{ij}$ *and the standard deviation is estimated as:* $s_i = \sqrt{\frac{\sum_{j=1}^{J}(x_{ij}-\bar{x}_i)^2}{J-1}}$. $\tilde{x}$ *and* $\hat{x}$ *represent the data after different pretreatment steps.*

Class	Method	Formula	Unit	Goal	Advantages	Disadvantages
I	Centering	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$	*O*	Focus on the differences and not the similarities in the data	Remove the offset from the data	When data is heteroscedastic, the effect of this pretreatment method is not always sufficient
II	Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	(-)	Compare metabolites based on correlations	All metabolites become equally important	Inflation of the measurement errors
	Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{(x_{i\max} - x_{i\min})}$	(-)	Compare metabolites relative to the biological response range	All metabolites become equally important. Scaling is related to biology	Inflation of the measurement errors and sensitive to outliers
	Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	√*O*	Reduce the relative importance of large values, but keep data structure partially intact	Stays closer to the original measurement than autoscaling	Sensitive to large fold changes
	Vast scaling	$\tilde{x}_{ij} = \frac{(x_{ij} - \bar{x}_i)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$	(-)	Focus on the metabolites that show small fluctuations	Aims for robustness, can use prior group knowledge	Not suited for large induced variation without group structure
	Level scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$	(-)	Focus on relative response	Suited for identification of e.g. biomarkers	Inflation of the measurement errors
III	Log transformation	$\tilde{x}_{ij} = {}^{10}\log(x_{ij})$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Log *O*	Correct for heteroscedasticity, pseudo scaling. Make multiplicative models additive	Reduce heteroscedasticity, multiplicative effects become additive	Difficulties with values with large relative standard deviation and zeros
	Power transformation	$\tilde{x}_{ij} = \sqrt{(x_{ij})}$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	√*O*	Correct for heteroscedasticity, pseudo scaling	Reduce heteroscedasticity, no problems with small values	Choice for square root is arbitrary.

The measurement errors were analyzed by estimation of the standard deviation from the biological, analytical, and sampling repeats. The standard deviations were binned by calculating the average variance per 10 metabolites ordered by mean value [23].

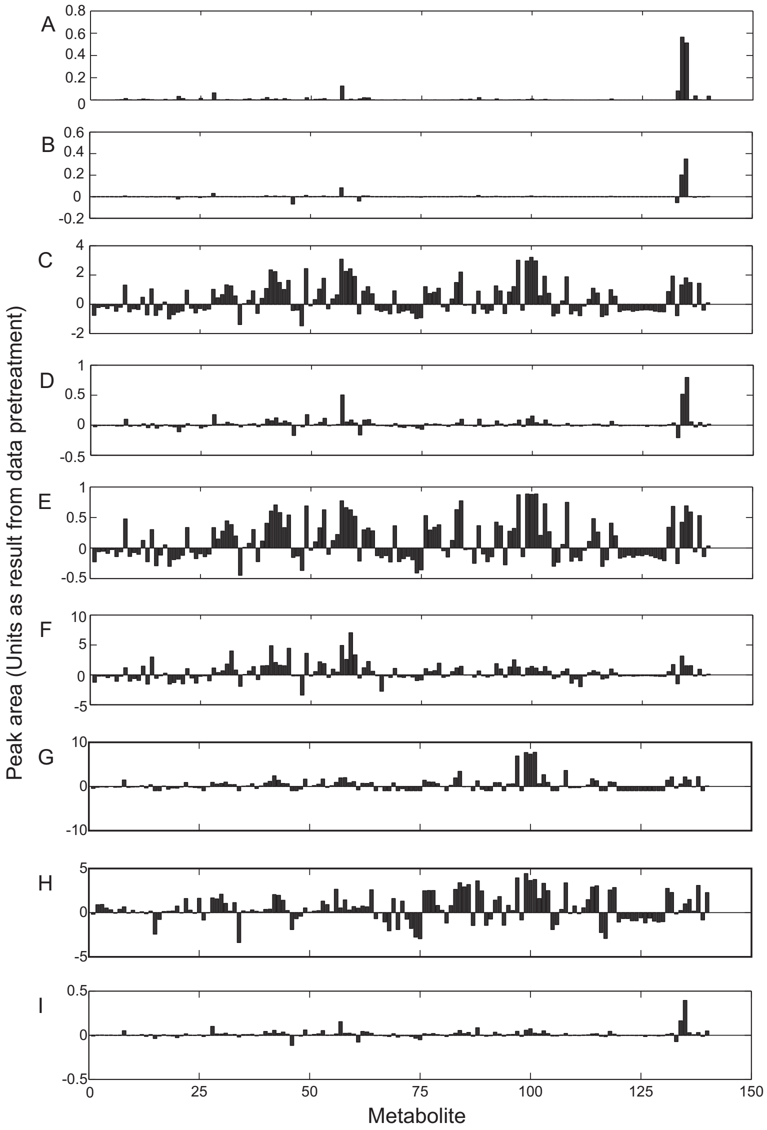


Figure 3 - Effect of data pretreatment on the original data. Original data of experiment G2 (A), and the data after centering (B), autoscaling (C), pareto scaling (D), range scaling (E), vast scaling (F), level scaling (G), log transformation (H), and power transformation (I). For units refer to Table 1.

The jackknife routine was performed according to the following setup. In round one experiments F1, G1, N1 were left out, in round two F2, G2, N1d were left out, and in round three F3, G3A, were left out. By selecting these experiments, the specific aspects of the experimental design were maintained.

3 Results and discussion

3.1 Properties of the clean data

For any data set, the total variation is the sum of the contributions of all the different sources of variation. The sources of variation in the data set used in this study were the induced biological variation, the uninduced biological variation, the sample work-up variation, and the analytical variation. The variation resulting from the sample work-up and the analytical analysis together was called technical variation. The contributions of the different sources of variation were roughly estimated from the replicate measurements by calculating the sum of squares (SS) and the mean square (MS) (Table 2). In this data set, the largest contribution to the variation originated from the induced biological variation, followed by the uninduced biological variation. The analytical variation was the smallest source of variation (Table 2).

3.2 The effect of pretreatment on the clean data

The application of different pretreatment methods on the clean data had a large effect on the resulting data used as input for data analysis, as is depicted for sample G2 in Figure 3. The different pretreatment methods resulted in different effects. For instance autoscaling (Figure 3C) showed many large peaks, while after pareto scaling (Figure 3D), only a few large peaks were present. It is evident that different results will be obtained when the in different ways pretreated data sets are used as the input for data analysis.

Source of variation	SS	MS
Analytical	0.0205	0.0102
Technical*	0.0482	0.0482
Uninduced biological	0.208	0.104
Induced biological	0.952	0.317
Total SS	1.23	

*Table 2: Estimation of the sources of variation in the data set. The SS and the MS for the different sources of variation are given, based on the experimental design presented in Figure 2. *The technical source of variation consists of the analytical error and the sample work-up error.*

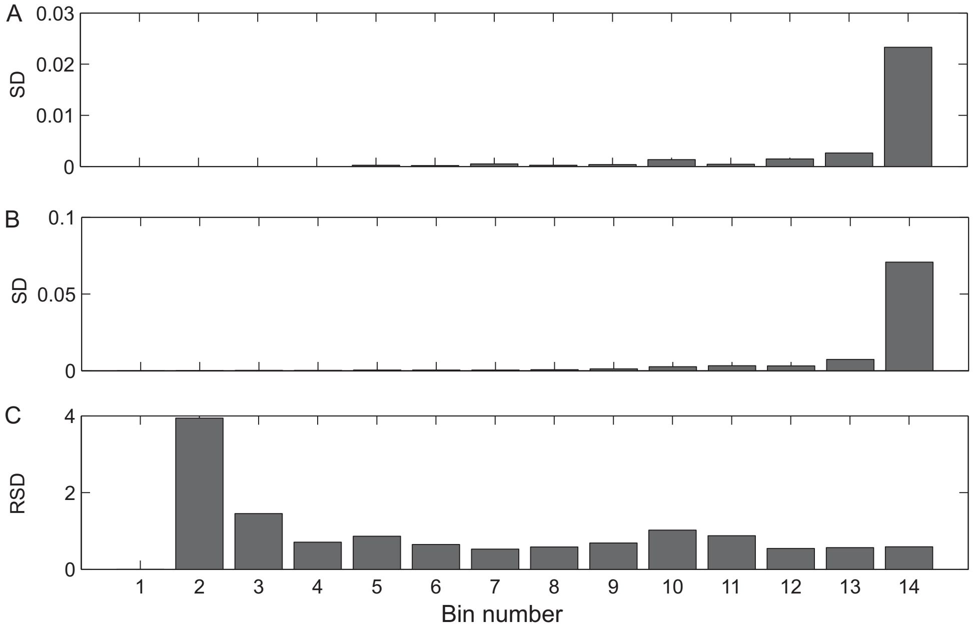


Figure 4 - Analytical and biological heteroscedasticity in the data. A: Analytical standard deviation (experiment G1), B: Biological standard deviation (all glucose experiments), and C: Relative biological standard deviation (all glucose experiments), as a function of the metabolite concentration. To obtain a clearer overview, the standard deviations were grouped together based on average mean value of the peak area (Binning, see Jansen et al. [23]). The first bin contained the metabolites whose peak area was below the detection limit.

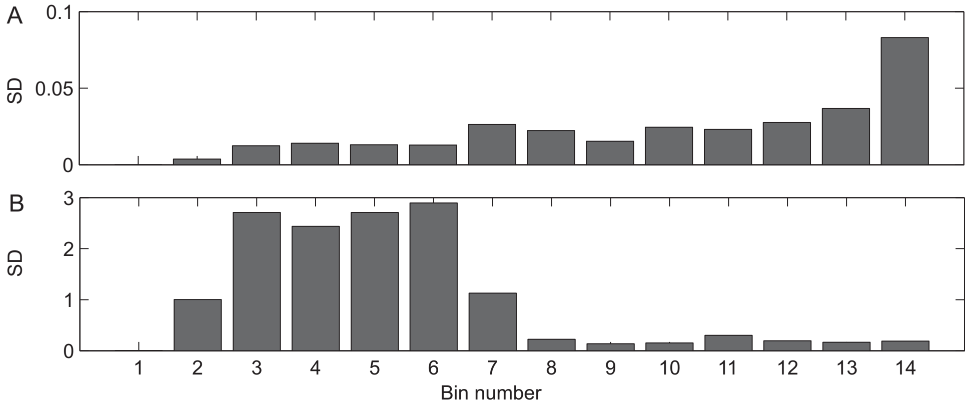


Figure 5 - Effect of data transformation on biological heteroscedasticity. A: power transformed data. B: log transformed data. The standard deviations over all glucose experiments were ordered by the mean value of the peak areas and binned per 10 metabolites. The first bin contained the metabolites whose peak area was below the detection limit.

3.2.1 Heteroscedasticity

To determine the presence or absence of heteroscedasticity in the data set, the standard deviations of the metabolites of the analytical and the biological repeats were analyzed (Figure 4). Analysis of the analytical and the uninduced biological standard deviations showed that heteroscedasticity was present both in the analytical error and in the biological uninduced variation (Figure 4A and B). In contrast, the *relative* biological standard deviation (Figure 4C), and also the relative analytical standard deviation (unpublished results), showed the opposite effect. Thus, metabolites present in high concentrations were relatively influenced less by the disturbances resulting from the different sources of uninduced variation, and were therefore more reliable.

The effect of the log and the power transformation on the data as a means to correct for heteroscedasticity is shown in Figure 5. Compared to the clean data (Figure 4B), the heteroscedasticity was reduced by the power transformation (Figure 5A), although the power transformation was not able to remove it completely. The results can possibly be improved further if a different power would be used (Box and Cox [24]). Also, the log transformation (Figure 5B) was able to remove heteroscedasticity, however only for the metabolites that are present in high concentrations. In contrast, the standard deviations of metabolites present in low concentrations were inflated after log transformation due to the large relative standard deviation of these low abundant metabolites.

Scaling approaches influence the heteroscedasticity as well, since the variation, and thus the heteroscedasticity, is converted into relative values to the scaling factor. It is likely that this aspect reduces the effect of the heteroscedasticity on the results.

3.3 The effect of data pretreatment on the data analysis results

PCA [9,25] was applied to analyze the effect on the data analysis for the in different ways pretreated data. PCA was chosen as it is an explorative tool that is able to visualize how the data pretreatment methods are able to reveal different aspects of the data in the scores and the accompanying loadings. Furthermore, it allows for identification of the most important metabolites for the biological problem by analysis of the loadings.

The score plots were judged on two aspects by visual inspection, namely the distance within the cluster of a specific carbon source and the distance between the clusters of different carbon sources. The loading plots show the contributions of the measured metabolites to the separation of the experiments in the score plots. As cellular metabolism is strongly interlinked (e.g. see [26,27]), it is expected that the concentrations of many metabolites are simultaneously affected when an organism is grown on a different carbon source. Therefore, the loadings are expected to show contributions of many different metabolites.

The data pretreatment methods used largely affected the outcome of PCA analysis (Figure 6). Three groups of data pretreatment methods could be identified in this way. After range scaling, a clear clustering of the samples was observed based on the carbon sources on which the sampled cells were grown (Figure 6A1). Furthermore, the loading plots (Figure 6A2 and 6A3) indicate that many metabolites contributed to the effects in the score plots; which is in agreement with the biological expectation. Autoscaling, level scaling, and log transformation resulted in similar PCA results as after range scaling (unpublished results).

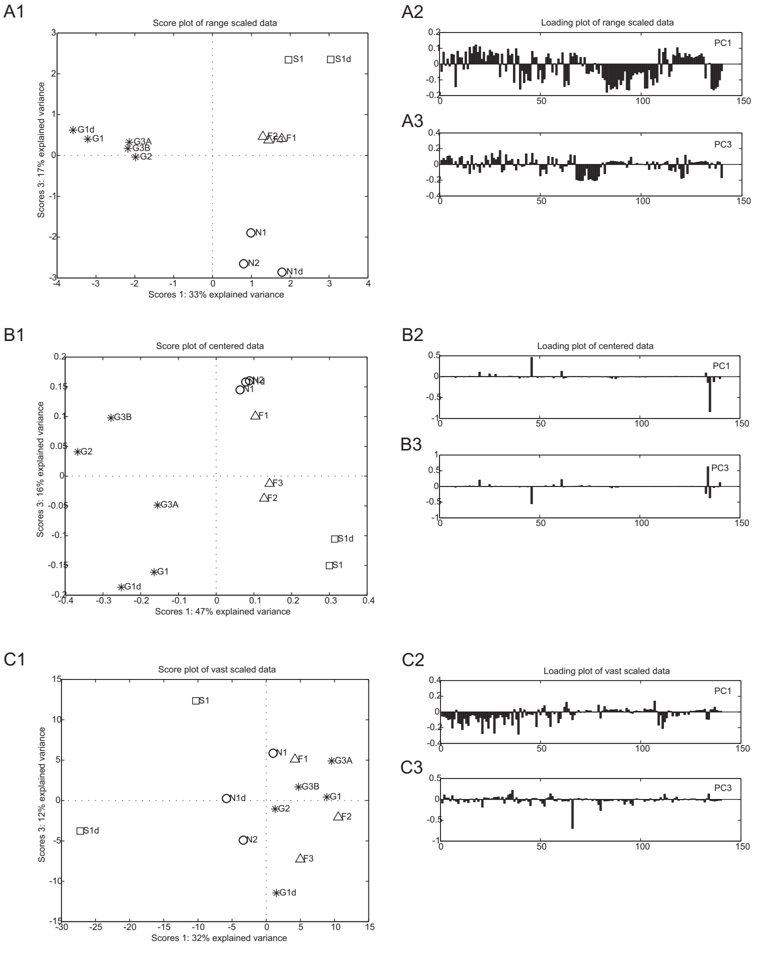
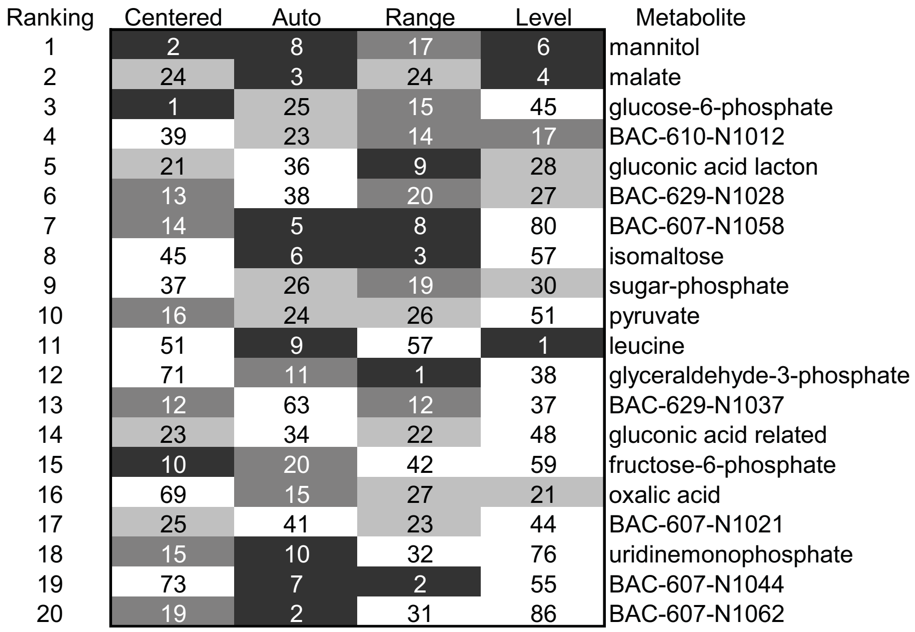


*Figure 6 - Effect of data pretreatment on the PCA results. PCA results of range scaled data (6A), centered data (6B), and vast scaled data (6C). For every pretreatment method the score plot (X1) (PC 1 vs. PC 3) and the loadings of PC 1 (X2) and PC 3 (X3) are shown. D-fructose (F, Δ), succinate (S,), D-gluconate (N,), D-glucose (G, *).*

The application of centering lead to intermediate clustering results in the score plots (Figure 6B1). The clusters were larger and less well-separated compared to the results for range scaling (Figure 6A1). The most striking results for centered data are visible in the loading plots (Figure 6B2 and 6B3). Only a few metabolites had very large contributions to the effects shown the score plot (Figure 6B1), which is in disagreement with the biological expectations. Power transformation and pareto scaling gave similar PCA results (unpublished results).

In contrast to the other pretreatment methods, vast scaling of the clean data resulted in a very poor clustering of the samples (Figure 6C1). Overlapping clusters were observed, although the loading plots (Figure 6C2 and 6C3) show contributions of many metabolites.

These results clearly demonstrate that the pretreatment method chosen dramatically influences the results of a PCA analysis. Consequently, these effects are also present in the rank of the metabolites.



Ranking	Centered	Auto	Range	Level	Metabolite
1	2	8	17	6	mannitol
2	24	3	24	4	malate
3	1	25	15	45	glucose-6-phosphate
4	39	23	14	17	BAC-610-N1012
5	21	36	9	28	gluconic acid lacton
6	13	38	20	27	BAC-629-N1028
7	14	5	8	80	BAC-607-N1058
8	45	6	3	57	isomaltose
9	37	26	19	30	sugar-phosphate
10	16	24	26	51	pyruvate
11	51	9	57	1	leucine
12	71	11	1	38	glyceraldehyde-3-phosphate
13	12	63	12	37	BAC-629-N1037
14	23	34	22	48	gluconic acid related
15	10	20	42	59	fructose-6-phosphate
16	69	15	27	21	oxalic acid
17	25	41	23	44	BAC-607-N1021
18	15	10	32	76	uridinemonophosphate
19	73	7	2	55	BAC-607-N1044
20	19	2	31	86	BAC-607-N1062

Figure 7 - Rank of the most important metabolites. The rank was based on the cumulative contributions of the loadings of the first three PCs. Top 10 metabolites are given in white characters with a black background, the top 11 to 20 is given in white characters with dark gray background, the top 21 to 30 is given in black characters with a light gray background.

3.4 Ranking of the most important metabolites

In functional genomics research, ranking of targets according to their relevance to the problem studied (for instance, strain improvement) is of great importance as it is time consuming and costly to validate the, in general, dozens or hundreds of leads that are generated in these studies[2]. As shown in Figure 6, the use of different pretreatment methods influenced the PCA analysis and the resulting loadings. For the different pretreatment methods, different metabolites were identified as the most important by studying the cumulative contributions of the loadings of the metabolites on PCs 1, 2 and 3 (Figure 7). Glucose-6-phosphate, for instance, was identified as the most important metabolite when using centering as the pretreatment method, while glyceraldehyde-3-phosphate (GAP) was identified as the most important metabolite when applying range scaling. For centering, autoscaling, and level scaling, GAP was the 71st, 11th, or 38th most important metabolite, respectively. The pretreatment of the clean data thus directly affected the ranking of the metabolites as being the most relevant.

The effect of a data pretreatment method on the rank of the metabolites is also apparent when studying the relation between the rank of the metabolites and the abundance (average peak area of a metabolite), or the fold change (standard deviation of the peak area over all experiments for a metabolite) (Figure 8). The effect of autoscaling (Figure 8B), and also range scaling (unpublished results), is in agreement with the expectation that the average concentration and the magnitude of the fold change are not a measure for the biological relevance of a metabolite. In contrast, with centering (Figure 8A), pareto scaling,

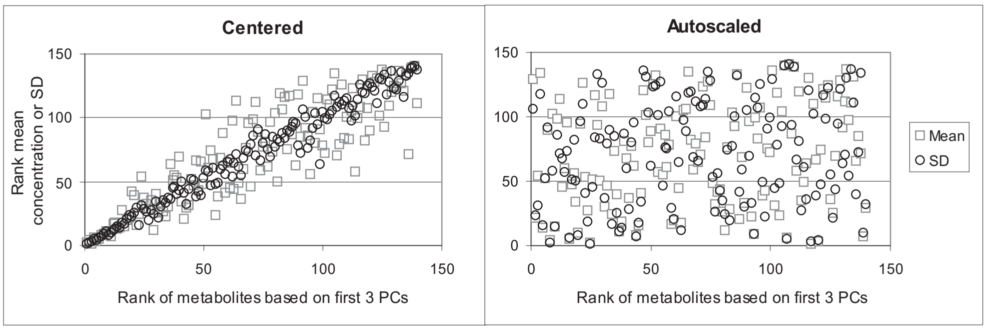


Figure 8: Relation between the abundance or the fold change of a metabolite and its rank after data pretreatment. The highest ranked metabolite after data pretreatment, based on its cumulative contributions on the loadings of the first three PCs, has position 1 on the X-axis. The metabolite that is ranked at position 1 on the Y-axis has either the highest fold change in concentration (largest standard deviation of the peak area over all the experiments in the clean data (O)); or is most abundant (largest mean concentration (□)) in the clean data.

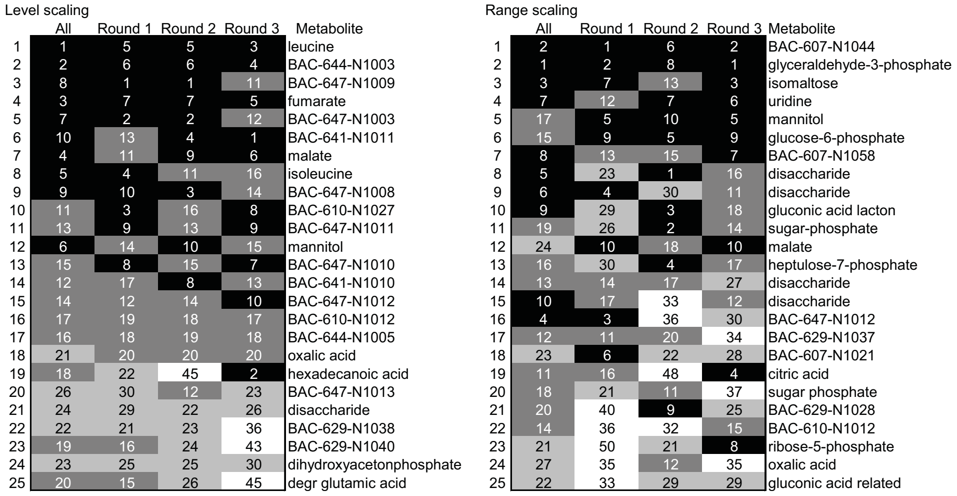
level scaling, log transformation, and power transformation (unpublished results), a clear relation between the rank of the metabolites and the abundance, or the fold change, of a metabolite was observed. This relation was less obvious for vast scaling, however still present (unpublished results).

3.5 Reliability of the rank of the metabolites

While the rank of the metabolites provides valuable information, the robustness of this rank is just as important as it determines the limits of the reliable interpretation of the rank. To test the reliability of the rank of the metabolites, a jackknife routine was applied [28].

The results for level scaling and range scaling are shown in Figure 9. The highest ranking metabolites (up to the eighth position) for both level scaled and range scaled data were relatively stable. For both methods, the fluctuations became larger for lower ranked metabolites, however, for the rank based on range scaled data the fluctuations in the rank increased faster than for the data resulting from level scaled data.

This resampling approach showed that the reliability of the rank of the most important metabolites is also dependent on the data pretreatment method. The most stable data pretreatment methods were centering, level scaling (Figure 9), log transformation, power transformation, pareto scaling, and vast scaling (results not shown). Autoscaling was less stable (results not shown), while the least stable data pretreatment method was range scaling. Two factors affect the reliability of the rank of the metabolites. The first factor relates to the reliability with which the scaling factor can be determined. For instance, level



Level scaling

	All	Round 1	Round 2	Round 3	Metabolite
1	1	5	5	3	leucine
2	2	6	6	4	BAC-644-N1003
3	8	1	1	11	BAC-647-N1009
4	3	7	7	5	fumarate
5	7	2	2	12	BAC-647-N1003
6	10	13	4	1	BAC-641-N1011
7	4	11	9	6	malate
8	5	4	11	16	isoleucine
9	9	10	3	14	BAC-647-N1008
10	11	3	16	8	BAC-610-N1027
11	13	9	13	9	BAC-647-N1011
12	6	14	10	15	mannitol
13	15	8	15	7	BAC-647-N1010
14	12	17	8	13	BAC-641-N1010
15	14	12	14	10	BAC-647-N1012
16	17	19	18	17	BAC-610-N1012
17	16	18	19	18	BAC-644-N1005
18	21	20	20	20	oxalic acid
19	18	22	45	2	hexadecanoic acid
20	26	30	12	23	BAC-647-N1013
21	24	29	22	26	disaccharide
22	22	21	23	36	BAC-629-N1038
23	19	16	24	43	BAC-629-N1040
24	23	25	25	30	dihydroxyacetonphosphate
25	20	15	26	45	degr glutamic acid

Range scaling

	All	Round 1	Round 2	Round 3	Metabolite
1	2	1	6	2	BAC-607-N1044
2	1	2	8	1	glyceraldehyde-3-phosphate
3	3	7	13	3	isomaltose
4	7	12	7	6	uridine
5	17	5	10	5	mannitol
6	15	9	5	9	glucose-6-phosphate
7	8	13	15	7	BAC-607-N1058
8	5	23	1	16	disaccharide
9	6	4	30	11	disaccharide
10	9	29	3	18	gluconic acid lacton
11	19	26	2	14	sugar-phosphate
12	24	10	18	10	malate
13	16	30	4	17	heptulose-7-phosphate
14	13	14	17	27	disaccharide
15	10	17	33	12	disaccharide
16	4	3	36	30	BAC-647-N1012
17	12	11	20	34	BAC-629-N1037
18	23	6	22	28	BAC-607-N1021
19	11	16	48	4	citric acid
20	18	21	11	37	sugar phosphate
21	20	40	9	25	BAC-629-N1028
22	14	36	32	15	BAC-610-N1012
23	21	50	21	8	ribose-5-phosphate
24	27	35	12	35	oxalic acid
25	22	33	29	29	gluconic acid related

Figure 9: Stability of the rank of the most important metabolites. The order of the metabolites is based on the average rank.

scaling uses the mean as the scaling factor. As the mean is based on all the measurements, it is quite stable. On the other hand, range scaling uses the biological range observed in the data as a scaling factor, which is based on two values only. The second factor that influences the reliability of the rank relates to those data pretreatment methods whose subsequent data analysis results show a preference for the high abundant metabolites (Figure 8). With these pretreatment methods, the stability of the rank is predetermined by this character due to the low relative standard deviation of the uninduced biological variation of the high abundant metabolites (Figure 4B).

It must be stressed that the pretreatment method that provides the most stable rank does not necessarily provides the most relevant biological answers.

4 Conclusions

This paper demonstrates that the data pretreatment method used is crucial to the outcome of the data analysis of functional genomics data. The selection of a data pretreatment method depends on three factors: (i) the biological question that has to be answered, (ii) the properties of the data set, and (iii) the data analysis method that will be used for the analysis of the functional genomics data.

Notwithstanding these boundaries, autoscaling and range scaling seem to perform better than the other methods with regard to the biological expectations. That is, range scaling and autoscaling were able to remove the dependence of the rank of the metabolites on the average concentration and the magnitude of the fold changes and showed biologically sensible results after PCA analysis. Other methods showed a strong dependence on the average concentration or magnitude of the fold change (centering, log transformation, power transformation, level scaling, pareto scaling), or lead to PCA results that were poorly interpretable in relation to the experimental setup (vast scaling).

Using a pretreatment method that is not suited for the biological question, the data, or the data analysis method, will lead to poor results with regard to, for instance, the rank of the most relevant metabolites for the biological question that is subject of study (Figure 7 and 8). This will therefore result in a wrong biological interpretation of the results.

In functional genomics data analysis, data pretreatment is often overlooked or is applied in an ad hoc way. For instance, in many software packages, such as Cluster [29] and the PLS toolbox [22], data pretreatment is integrated in the data analysis program and can be easily turned on or off. This can lead to a careless search through different pretreatment methods until the results best fit the expectations of the researcher. Therefore, we advise against method mining. With method mining, the best result translates to ‘which method fits the expectations the best’. This is poor practice, as results cannot be considered reliable when the assumptions and limitations of a data pretreatment method are not taken into account. Furthermore, it is sometimes unknown what to expect, or the starting hypothesis is incorrect.

As far as we are aware, this is the first time that the importance of selecting a proper data pretreatment method on the outcome of data analysis in relation to the identification of biologically important metabolites in metabolomics/functional genomics is clearly demonstrated.

5 Acknowledgements

The authors would like to thank Karin Overkamp and Machtelt Braaksma for the generation of the biological samples and sample work up, and Maud Koek, Bas Muilwijk, and Thomas Hankemeier for the analysis of the samples and data preprocessing. This research was funded by the Kluyver Centre for Genomics of Industrial Fermentation, which is supported by the Netherlands Genomics Initiative (NROG).

6 References

1. Reis EM, Ojopi EPB, Alberto FL, Rahal P, Tsukumo F, Mancini UM, Guimaraes GS, Thompson GMA, Camacho C, Miracca E, Carvalho AL, Machado AA, Paquola ACM, Cerutti JM, da Silva AM, Pereira GG, Valentini SR, Nagai MA, Kowalski LP, Verjovski-Almeida S, Tajara EH, Dias-Neto E, Head and Neck Annotation Consortium: **Large-scale Transcriptome Analyses Reveal New Genetic Marker Candidates of Head, Neck, and Thyroid Cancer.** *Cancer Res* 2005, **65:**1693-1699.
2. van der Werf MJ: **Towards replacing closed with open target selection strategies.** *Trends Biotechnol* 2005, **23:**11-16.
3. van der Werf MJ, Jellema RH, Hankemeier T: **Microbial Metabolomics: replacing trial-and-error by the unbiased selection and ranking of targets.** *J Ind Microbiol Biotechnol* 2005, **32:**234-252.
4. Fiehn O: **Metabolomics - the link between genotypes and phenotypes.** *Plant Mol Biol* 2002, **48:**151-171.
5. Shurubor YI, Paolucci U, Krasnikov BF, Matson WR, Kristal BS: **Analytical precision, biological variation, and mathematical normalization in high data density metabolomics.** *Metabolomics* 2005, **1:**75-85.
6. Keller HR, Massart DL, Liang YZ, Kvalheim OM: **Evolving factor analysis in the presence of heteroscedastic noise.** *Anal Chim Acta* 1992, **263:**29-36.
7. Kvalheim OM, Brakstad F, Liang Y: **Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise.** *Anal Chem* 1994, **66:**43-51.
8. Bro R, Smilde AK: **Centering and scaling in component analysis.** *J Chemom* 2003, **17:**16-33.

9. Jackson JE: *A user's guide to principal components.* John Wiley & Sons, Inc.; 1991.
10. Eriksson L, Johansson E, Kettaneh-Wold N, Wold S: **Scaling.** In *Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS).* Umetrics; 1999:213-225.
11. Smilde AK, van der Werf MJ, Bijlsma S, van der Werff-van der Vat B, Jellema RH: **Fusion of mass-spectrometry-based metabolomics data.** *Anal Chem* 2005, **77:**6729-6736.
12. Keun HC, Ebbels TMD, Antti H, Bollard ME, Beckonert O, Holmes E, Lindon JC, Nicholson JK: **Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling.** *Anal Chim Acta* 2003, **490:**265-276.
13. Sokal RR, Rohlf FJ: **Assumptions of analysis of variance.** In *Biometry.* 3rd edition. New York: W.H. Freeman and Co.; 1995:392-450.
14. Hartmans S, van der Werf MJ, de Bont JAM: **Bacterial degradation of styrene involving a novel flavin adenine dinucleotide-dependent styrene monooxygenase.** *Appl Environ Microbiol* 1990, **56:**1347-1351.
15. van der Werf MJ, Pieterse B, van Luijk N, Schuren F, van der Werff-van der Vat B, Overkamp K, Jellema RH: **Multivariate analysis of microarray data by principal component discriminant analysis: prioritizing relevant transcripts linked to the degradation of different carbohydrates in *Pseudomonas putida* S12.** *Microbiology* 2006, **152:**257-272.
16. Pieterse B, Jellema RH, van der Werf MJ: **Quenching of microbial samples for increased reliability of microarray data.** *J Microbiol Methods* 2006, **64:**207-216.
17. Ruijter GJG, Visser J: **Determination of intermediary metabolites in Aspergillus niger.** *J Microbiol Methods* 1996, **25:**295-302.
18. Koek M, Muilwijk B, van der Werf MJ, Hankemeier T: **Microbial metabolomics with gas chromatography mass spectrometry.** *Anal Chem* 2006, **78:**1272-1281.
19. Verduyn C, Postma E, Scheffers WA, van Dijken JP: **Physiology of Saccharomyces cerevisiae in anaerobic glucose-limited chemostat cultures.** *J Gen Microbiol* 1990, **136:**395-403.
20. Stein SE: **An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data.** *J Am Soc Mass Spectrom* 1999, **10:**770-781.
21. Mathworks. **Matlab 7.** 2005.
22. Eigenvector. **PLS Toolbox 3.0.** 2003.
23. Jansen JJ, Hoefsloot HCJ, Boelens HFM, van der Greef J, Smilde AK: **Analysis of longitudinal metabolomics data.** *Bioinformatics* 2004, **20:**2438-2446.
24. Box GEP, Cox DR: **An Analysis of Transformations.** *J R Statist Soc B* 1964, **26:**211-252.
25. Jolliffe IT: *Principal Component Analysis.* New York: Springer-Verlag; 2002.

26. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD: **MetaCyc: a multiorganism database of metabolic pathways and enzymes.** *Nucleic Acids Res* 2004, **32:**D438-D442.
27. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 2000, **28:**27-30.
28. Efron B, Tibshirani RJ: **The jackknife.** In *An Introduction to the Bootstrap*. New York: Chapman & Hall; 1993:141-152.
29. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.

3 Removing confounding effects from micro-array data

Robert A. van den Berg, Machtelt Braaksma, Johan A. Westerhuis, Mariët J. van der Werf and Age K. Smilde

Summary

Confounding variation is variation that obscures the induced biological variation. Removal of the confounding variation can improve the interpretation of the data. In this paper we present a strategy to remove confounding variation based on an ANOVA approach, and to assess the impact of the removal on the interpretation of the variation induced by the experimental design. Our strategy is applied to an *Aspergillus niger* micro-array data set in which the variation induced by the experimental design was obscured by confounding variation induced by the presence or absence of substrate. The confounding variation was successfully removed; however, variation induced by the experimental design was partially removed as well. This was due to correlation between the variation induced by the experimental design and the confounding variation.

1 Introduction

In micro-array experiments the response of thousands of genes to the experimental conditions is measured. The experimental conditions are selected to study a certain biological phenomenon by inducing variation in processes relevant for the biological question [1,2]. Analysis of the resulting data aims to extract the variation relevant for the biological question from the total variation present in the data set. Sometimes, however, preliminary data analysis shows that variation relevant to the biological question is not the main source of variation in the data set. This can have different causes: (i) other sources of variation, for instance unexpected biological effects, obscure the variation of interest; (ii) the induced biological variation is very small compared to uninduced biological or technical variation; or (iii) the variation of interest is not present in the data set.

Confounding variation is variation that obscures the induced biological variation. Furthermore, it originates from a certain structured source within the experimental setup. An example of a confounding effect from medical science could be an age or gender effect that influences the effectiveness of the medication. To improve the interpretation of the data, it is beneficial to remove this variation from the data set prior to data analysis.

We use an analysis of variance (ANOVA) [3,4] approach to remove variation caused by confounding factors from a genomics data set. Furthermore, we analyze the impact of the removal of this variation and discuss the consequences for providing information relevant for the biological question. Our approach is illustrated by the application on *A. niger* micro-array data in which an unwanted effect, related to substrate depletion, hampered the identification of effects related to the biological question.

2 Theory

2.1 Notation

In section 2, we will provide the theoretical background of the method. For this, the following notations will be used.

X (I x J) is the data matrix consisting of I experiments and J measured biomolecules. For simplicity, it is assumed that the variables in **X** are column mean centered.

D (I x K) is a design matrix for K design factors coded with 0 and 1.

U (I x L) is a matrix that consists of L dummy variables - also called confounders - that are used to encode the structure of the obscuring variation. **u** is a vector coded with -1 and 1 if there are two groups; and a matrix **U** coded with 0 and 1 where 1 indicates group membership if there are more groups.

F (I x [K+L]) is the concatenation of **D** and **U**.

B (*K* x *J*) and **M** (*L* x *J*) are weight matrices that describe the estimations of the contributions of each factor in respectively **D** and **U** to the *J* variables in **X**. The estimates of **B** and **M** are $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{M}}$. In the sequential approach (see below) **B** becomes $\widetilde{\mathbf{B}}$ and $\widehat{\widetilde{\mathbf{B}}}$.

R (*K*+*L* x *J*) is a weight matrix that describes the contribution of each factor in **F** to the *J* variables in **X**. The estimate of **R** is $\widehat{\mathbf{R}}$.

$\mathbf{R_D}$ (*K* x *J*) and $\mathbf{R_U}$ (*L* x *J*) are partitions of **R**, $\mathbf{R} = \begin{bmatrix} \mathbf{R_D} \\ \mathbf{R_U} \end{bmatrix}$, that correspond to partitions **D** and **U** of **F**. Their estimates are $\widehat{\mathbf{R}}_\mathbf{D}$ and $\widehat{\mathbf{R}}_\mathbf{U}$.

E (*I* x *J*), **G** (*I* x *J*), $\mathbf{G_{sim}}$ (*I* x *J*) and $\mathbf{G_{seq}}$ (*I* x *J*) are matrices containing the residuals of the model. The expected mean of these matrices is zero.

2.2 Variation in X

In the ideal situation, the variation in **X** is fully attributable to the experimental design and can be estimated as follows:

(1) $$\mathbf{X} = \mathbf{DB} + \mathbf{E}\,.$$

Sometimes, however, other sources of structural variation are present. This latter variation can be the result of co-occurring biological effects that are not directly, or only partially, related to the biological question. As a result of this structural extra variation, the confounding variation, the model has to be expanded to:

(2) $$\mathbf{X} = \mathbf{D}\widetilde{\mathbf{B}} + \mathbf{UM} + \mathbf{G}\,.$$

2.3 Removal of the confounding variation in X

An ANOVA approach is followed to estimate and remove the confounding variation. In this approach, the variation originating from **UM** is estimated and removed from **X**. In real life data, it is likely that there is correlation between **D** and **U**. As a result it is not possible to fully distinguish between variation originating from **D** or **U**. We will discuss two approaches to remove the variation originating from **U** that deal differently with the correlation between **D** and **U**. The two approaches yield the same result when **D** and **U** are uncorrelated or orthogonal.

2.4 Simultaneous estimation of the variation originating from D and U

The first option is to simultaneously estimate the variation resulting from **U** and **D** via a regression step:

(3) $$\mathbf{X} = \mathbf{D}\tilde{\mathbf{B}} + \mathbf{UM} + \mathbf{G}_{sim} = \mathbf{FR} + \mathbf{G}_{sim}$$

(4) $$\hat{\mathbf{R}} = (\mathbf{F}^{\mathbf{T}}\mathbf{F})^{-1}\mathbf{F}^{\mathbf{T}}\mathbf{X}$$

The variation originating from **D** and **U** is now divided over the weights in **R**. $\mathbf{X}^*$ is estimated by removing the variation captured by $\mathbf{UR_u}$.

(5) $$\mathbf{F}\hat{\mathbf{R}} + \mathbf{G}_{sim} = \begin{bmatrix}\mathbf{D} & \mathbf{U}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{R}}_{\mathbf{D}} \\ \hat{\mathbf{R}}_{\mathbf{U}}\end{bmatrix} + \mathbf{G}_{sim}$$

(6) $$\mathbf{X} - \mathbf{U}\hat{\mathbf{R}}_{\mathbf{U}} = \mathbf{D}\hat{\mathbf{R}}_{\mathbf{D}} + \mathbf{U}\hat{\mathbf{R}}_{\mathbf{U}} + \mathbf{G} - \mathbf{U}\hat{\mathbf{R}}_{\mathbf{U}}$$

$$\mathbf{X}^* = \mathbf{D}\hat{\mathbf{R}}_{\mathbf{D}} + \mathbf{G}_{sim}$$

2.5 Sequential estimation of the variation originating from D and U

For the sequential estimation of the variation originating from **U**, the variation originating from **D** is not taken into account in the model. The model then becomes:

(7) $$\mathbf{X} = \mathbf{UM} + (\mathbf{G}_{seq} + \mathbf{D}\tilde{\mathbf{B}})$$

$$\mathbf{X} = \mathbf{UM} + \tilde{\mathbf{G}}_{seq}$$

M can be estimated in the same way as $\hat{\mathbf{R}}$ in equation (4) by the following:

$$\hat{\mathbf{M}} = (\mathbf{U}^{\mathbf{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathbf{T}}\mathbf{X}$$

The variation from $\widehat{\mathbf{UM}}$ can then be removed to yield:

(8) $$\mathbf{X} - \widehat{\mathbf{UM}} = \tilde{\mathbf{G}}_{seq} = \mathbf{D}\tilde{\mathbf{B}} + \mathbf{G}_{seq}$$

$$\mathbf{X}^* = \mathbf{D}\tilde{\mathbf{B}} + \mathbf{G}_{seq}$$

$$\hat{\tilde{\mathbf{B}}} = (\mathbf{D}^{\mathbf{T}}\mathbf{D})^{\mathbf{-1}}\mathbf{D}^{\mathbf{T}}\mathbf{X}^*$$

$\hat{\tilde{\mathbf{B}}}$ can be estimated analogously to (4) and analyzed. The estimation of $\hat{\tilde{\mathbf{B}}}$ is not pursued here as the objective was to prepare the data for further multivariate analysis.

2.6 Comparison of the simultaneous and sequential approach

The variation induced by the confounder is estimated to subsequently remove this variation from the data matrix. By removing $\mathbf{U}\hat{\mathbf{M}}$ in the sequential approach, all the variation attributable to grouping structure of **U** is removed from **X**. This includes variation that could originate from the part of the design **D** which correlates with **U**. In the simultaneous approach, the variation resulting from correlating factors in **D** and **U** is divided over **D** and **U.** As a result, it is not possible to distinguish between the two sources of variation and it is therefore not possible to fully remove **U** from **X**. The interpretation of the variation remaining in $\mathbf{X}^*$ is therefore in the sequential approach more straightforward since only variation that is solely attributable to **D** remains. We therefore chose the sequential approach.

2.7 Design factors in D affected by the sequential removal of the confounders

To assess which factors in **D** correlate with **U**, the part of **D** that correlates with **U** can be removed from **D**:

$$\mathbf{D}^* = \mathbf{D} - \mathbf{U}(\mathbf{U}^\mathrm{T}\mathbf{U})^{-1}\mathbf{U}^\mathrm{T}\mathbf{D} \tag{9}$$

When there is a strong correlation between some factors of **D** and **U**, large parts of the variation of the factors that correlate with **U** will be removed in $\mathbf{D}^*$. $\mathbf{D}^*$ can show which sources of variation from the experimental design are affected by the removal of the confounders. Depending on how severe the factors of the experimental design are affected, the contributions of these factors to the variation in the data set $\mathbf{X}^*$ should be treated with care.

3 Results

3.1 Aspergillus niger micro array data set

The approach discussed in the previous section is applied to an *A. niger* micro-array data set originating from a study in which the proteolytic activity of *A. niger* was studied. *A. niger* is a filamentous fungus of which the genome sequence was made available recently [5]. Metabolomics [6] was applied to *A. niger* to understand the mechanisms that induce the proteolytic system in order to be able to reduce extracellular protease activity. The experimental design consisted of four environmental parameters: pH, carbon source, nitrogen source, and nitrogen concentration selected to induce variation in the extracellular protease activity. The four parameters were present on two levels. Samples were collected from controlled batch fermentations and analyzed by metabolomics. Later, however, micro-arrays became available and the transcription profile of these samples was also analyzed using the Affymetrix GeneChip platform.

The onset of the proteolytic system generally occurs during the transition from the exponential growth phase to the stationary growth phase. This phase transition usually starts when the carbon source is depleted [7]. Different samples were obtained around this transition phase. For the transcriptomics analysis, 20 samples from the samples collected for the metabolomics study were selected based on the presence of proteolytic activity. After initial analysis of the micro-array data, it became clear that a strong effect was present in these data which seemed to be related to the presence or absence of substrate. The principal component analysis (PCA) [8,9] score plot of the first two principal components of the

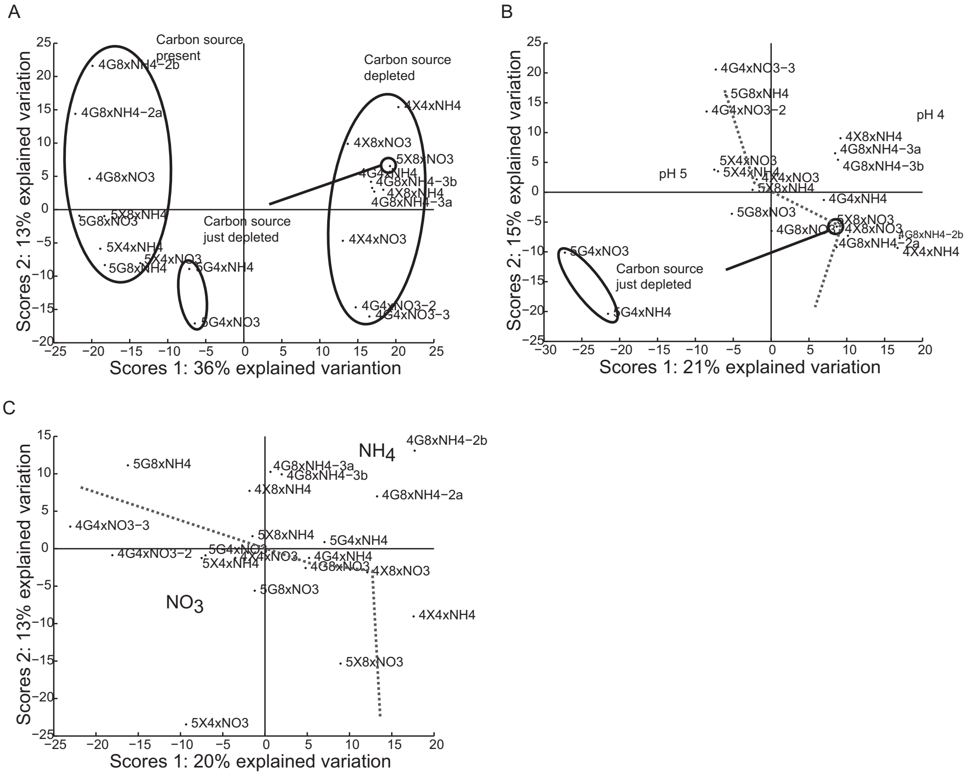


Figure 1: PCA score plots of A. niger micro-array data before and after the removal of the confounding variation. (A) The original micro-array data with a clear presence of a confounding effect. The ovals indicate different groups. (B) The confounding effect removed based on two confounding groups. The dotted line indicates the separation based on pH. (C) The confounding effect removed based on three confounding groups. The dotted line indicates separation based on nitrogen source.

range scaled [10] transcriptomics data showed a strong grouping of the experimental conditions (Figure 1A). The protease activity of these samples showed that the grouping of the experiments in the PCA score plots was not related to high or low protease activity. Closer inspection of the conditions in the fermentor at the time of sampling indicated that the grouping was based on the presence or depletion of the substrate. This effect is clearly visualized by samples obtained from the same fermentation, collected at different time points. The first samples 4G8xNH_4-2a and 4G8xNH_4-2b (a and b indicate technical duplicates for the array analysis) were taken when there was still substrate present, while 4G8xNH_4-3a and 4G8xNH_4-3b were obtained when the substrate was depleted. The samples 4G8xNH_4-2a and 4G8xNH_4-2b are grouped in a different cluster than the samples in 4G8xNH_4-3a and 4G8xNH_4-3b (Figure 1A). As the variation induced by the depletion of the substrate was therefore the likely cause that obscured the variation relevant for the induction of the proteolytic activity, we removed the variation related to the depletion of the substrate from the data set.

3.2 Removal of the confounding variation

To remove the confounding variation, a confounder matrix **U** was defined. It was anticipated based on medium composition analysis that the effect was related to either two groups: presence or depletion of substrate; or three groups: substrate present, substrate depleted for two to five hours, substrate depleted for more than 13 hours. Beforehand it was not clear whether two or three groups were sufficient to fully describe the effect. For both cases, **U** is defined as shown in Table 1.

To assess the effect of the removal of the variation induced by the different groups, a PCA analysis was performed on the resulting $\mathbf{X}^*$ (Figure 1B and 1C).

Experiment	Two effects	Three effects		
5G4xNO3	-1	1	0	0
5X4xNO3	1	0	0	1
4X8xNO3	-1	0	1	0
4G8xNO3	1	0	0	1
4X4xNO3	-1	0	1	0
5G8xNO3	1	0	0	1
5X4xNH4	1	0	0	1
4G4xNO3-2	-1	0	1	0
4G4xNO3-3	-1	0	1	0
5G4xNH4	-1	1	0	0
4G4xNH4	-1	0	1	0
5X8xNO3	-1	0	1	0
4X4xNH4	-1	0	1	0
4X8xNH4	-1	0	1	0
4G8xNH4-2a	1	0	0	1
4G8xNH4-2b	1	0	0	1
4G8xNH4-3a	-1	0	1	0
4G8xNH4-3b	-1	0	1	0
5G8xNH4	1	0	0	1
5X8xNH4	1	0	0	1

*Table 1 – Design of confounder effects (matrix **U**)*

3.2.1 Groups based on presence or depletion of substrate

When two groups were defined (Figure 1B), the grouping in the score plots was no longer dominated by the presence or depletion of the substrate. Furthermore, it seemed that there was a mild separation based on the experimental design parameter pH. This could indicate that the variation resulting from parameters of the experimental design became dominant, instead of the variation resulting from the confounders. However, samples $5G4xNO_3$ and $5G4xNH_4$ still stood out in the PCA score plots compared to the remainder of the samples (Figure 1B). These samples were obtained 2-5 hours after substrate depletion, and this could indicate that not all confounding variation was removed. The removal of the confounding variation based on two groups resulted in the removal of 52.5% of the sum of squares (SS) of the original centered data set.

Removal of the confounding variation can affect the variation originating from the experimental design. To assess the influence of this effect, the confounding effects were also removed from the original design matrix. The SS that remained relative to the SS of the original design matrix is an indication how strong a design factor is affected. For the confounding effects based on two groups, the design parameters pH and nitrogen concentration were affected the most by the removal of the confounding design effect (Table 2). Here, more than 10% of the variation of these parameters of the experimental design was removed. The experimental design parameters carbon and nitrogen source were only mildly affected, in these cases less than 1.5% of the variation was removed. Since the pH factor of the experimental design is affected by the removal of the confounder effects, it is remarkable and to a certain extent alarming that the PCA score plot indicate that a major part of the remaining variation in the data set is related to the pH parameter of the experimental design (Figure1B). This, together with the separation of samples $5G4xNO_3$ and $5G4xNH_4$ from the other samples could indicate that two groups are not sufficient to fully remove the confounding variation.

3.2.2 Three groups

Removal of the effects based on three groups seemed to fully remove the confounding effects based on absence or presence of the substrate (Figure 1C). There was no indication that grouping of experiments was related to the absence or presence of a substrate. Furthermore, there was a mild separation between the experiments based on the nitrogen source. The removal of the confounding variation removed 66.4% of the variation of the original centered data set.

Subtraction of the confounding effects from the design matrix leads to the removal of 42.2% of the variation of the pH parameter of the experimental design. Furthermore, 19.2% of the nitrogen concentration parameter, and 8.9% of the carbon source parameter of the experimental design were removed. Only the effect based on the nitrogen source in the experimental design remained relatively unaffected as only 1.5% was removed. The variation induced by this experimental design parameter became a dominant source of variation in $\mathbf{X}^*$ (Figure 1C). The importance of the nitrogen source as design parameter was also confirmed by an ANOVA analysis of protease activity (results not shown). This could indicate that three groups were better in removing the confounding variation than two groups, but may also throw away too much of the biological variation.

4 Discussion

Variation induced by confounding factors can hamper the interpretation of data sets resulting from omics experiments. In this paper, an ANOVA approach is used to remove confounding variation that was not related to extracellular protease activity, which was the main focus of this –omics study.

Removal of the confounding variation comes at a cost since the variation originating from factors in the experimental design that correlate with the confounding factors is removed as well (Table 2). This can hamper the interpretation of the relation of experimental design parameter with the biological question. In the data set studied in this paper, the pH parameter of the experimental design was affected by the removal of the confounding variation with both estimates of the confounding factors. As a consequence, large parts of the variation that could result from the pH parameter of the experimental design were removed from the original data together with the confounding variation. By removing the confounders from the original design matrix, it is possible to estimate which design factors are affected by the removal of the confounders.

The most important and also most difficult step for the removal of confounding variation is to define the confounder matrix as good as possible with regard to the biological question. The confounder matrix should be chosen based on expert knowledge of the nature of the effects that hamper the interpretation of the results. When the confounding

Design factor	Two groups (%)	Three groups (%)
pH	86.5	57.8
Carbon source	99.8	91.1
Nitrogen source	98.5	98.5
Nitrogen concentration	89.7	80.8

Table 2 – SS remaining in design factors D after removal of confounding effects.

parameter is not easily described as a discrete group structure, continuous measurements can directly be used as a parameter as well. Such an approach is called ANCOVA [4]. The choice for a certain confounder matrix can be validated by analyzing the removal of the confounding variation from the data matrix, for instance, by PCA. This, however, might not be sufficient to be confident that the correct confounding variation is removed.

In this paper, the problems of selecting the proper confounders are illustrated by the results for the removal of these confounder effects. In the two group example, the pH factor of the experimental design is affected by the removal of the confounders while the PCA analysis (Figure 1B) indicated that the pH factor might have become an important source of variation of the experimental design. The three group example strongly affected the pH factor of the experimental design, and contribution of the pH factor is not obvious anymore in the PCA score plot (Figure 1C). This example illustrates that determining the confounding effects is both very important and difficult, as the results of the removal of the confounding effects are strongly influenced.

The results of removing the confounding variation should ideally lead to the ability to extract the information relevant to the biological question under study. Depending on the biological question, and when it is unclear what the best confounder matrix is, different confounder effects can be tested empirically for better performance with regard to the biological question. For the data set analyzed in this paper, a strategy could be to validate the full model, the two group, and the three group model by the generation of PLS regression [11,12] models that model the relation between the gene expression data and protease activity, or another relevant phenotype parameter. This will most likely provide the best indication which approach gives the best results. This strategy is pursued in a follow up project.

5 References

1. Kerr MK, Churchill GA: **Experimental design for gene expression microarrays.** *Biostat* 2001, **2:** 183-201.
2. van der Werf MJ: **Towards replacing closed with open target selection strategies.** *Trends Biotechnol* 2005, **23:** 11-16.
3. Kerr MK, Martin M, Churchill GA: **Analysis of Variance for Gene Expression Microarray Data.** *J Comput Biol* 2000, **7:** 819-837.
4. Maxwell S, Delaney H: *Designing experiments and analyzing data: a model comparison perspective*. Lawrence Erlbaum Associates, Inc; 2000.
5. Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ *et al.*: **Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88.** *Nat Biotechnol* 2007, **25:** 221-231.

6. van der Werf MJ, Overkamp KM, Muilwijk B, Coulier L, Hankemeier T: **Microbial metabolomics: Toward a platform with full metabolome coverage.** *Anal Biochem* 2007, **370:** 17-25.
7. Braaksma M, Punt PJ: ***Aspergillus* as a cell factory for protein production: controlling protease activity in fungal production.** In *The Aspergilli. Genomics, Medical Aspects, Biotechnology, and Research Methods*. Edited by Goldman GH, Osmani SA. Boca Raton: CRC Press; 2008.
8. Jolliffe IT: *Principal Component Analysis*, Second Edition edn. New York: Springer-Verlag; 2002.
9. Jackson JE: *A user's guide to principal components*. John Wiley & Sons, Inc.; 1991.
10. van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ: **Centering, scaling, and transformations: improving the biological information content of metabolomics data.** *BMC Genomics* 2006, **7**.
11. Geladi P, Kowalski BR: **Partial least-squares regression: a tutorial.** *Anal Chim Acta* 1986, **185:** 1-17.
12. Boulesteix AL, Strimmer K: **Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach.** *Theoretical Biology and Medical Modelling* 2005, **2:** 23.

4 Identifying connections between a metabolic pathway and its surrounding network from metabolomics data

Robert A. van den Berg, Carina M. Rubingh, Johan A. Westerhuis, Mariët J. van der Werf and Age K. Smilde

Summary

In metabolomics it can be important to focus the data analysis to areas of specific interest within metabolism. For instance, the biological question under study can be related to a specific class of metabolites or a specific pathway. Supervised data analysis methods can bring this focus into data analysis and provide information on the behavior of the interesting metabolites in relation to the remainder of the metabolome. Here, we describe the application of consensus PCA (CPCA) and canonical correlation analysis (CCA) as a means to focus data analysis. CPCA searches for major trends in the behavior of metabolite concentrations common for the metabolites of interest and the remainder of the metabolome. CCA identifies the strongest correlations between these two subsets.

CPCA and CCA were applied to two microbial metabolomics data sets. The first data set, derived from *Pseudomonas putida*, was relatively simple and contained metabolomes obtained under four environmental conditions only. The second data set, obtained from *Escherichia coli*, was complex and contained metabolomes from 28 different environmental conditions. For the first data set, CCA and CPCA gave similar results as the variation in the two subsets was similar. In contrast, CCA and CPCA yielded different results in case of the *E. coli* data set. With CPCA the trends in the metabolites of interest – the phenylalanine biosynthesis intermediates - dominated the results. These trends were related to high and low phenylalanine productivity, and important metabolites in the CPCA analysis were associated with amino acid metabolism and regulation of the phenylalanine biosynthesis route.

With CCA neither subset dominated the data analysis. CCA described correlations between the subsets based on wild type and overproducing strain differences and different carbon sources. For the correlation based on strain differences, metabolites from the aromatic amino acid pathways were important.

Both CCA and CPCA enable one to focus the data analysis of metabolomics data to groups of metabolites that are of specific interest. Depending on the nature of the data set, they provide different, complementary, views on the relation between the metabolites of interest and the remainder of the metabolome.

1 Background

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

So far, data analysis methods that single out the behavior of groups of metabolites in relation to the behavior of other metabolites in the data set have not been applied. Consensus PCA-W (CPCA) [8], and canonical correlation analysis (CCA) [9] are methods that can perform such an analysis. CPCA searches for the largest common trends between behavior of the concentration of the metabolites of interest and the remaining metabolites. In CCA the strongest correlation between the behavior of the metabolites in the two data sets is determined. Both methods provide information on the relation between the metabolites of interest and the remaining metabolites; however the methods are based on different principles and give different views on the underlying biology, as will be explained in the next section.

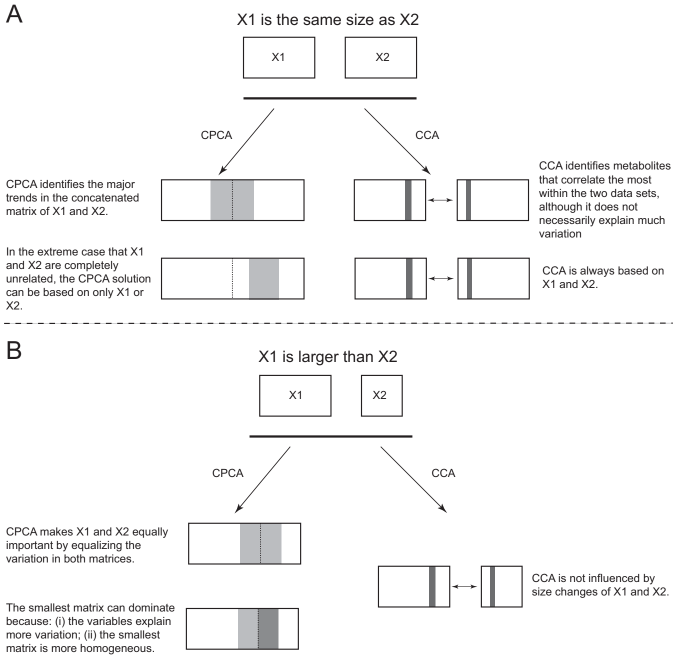
2 Theory

In the following section we will discuss different properties of CPCA and CCA. The following notations will be used: $\mathbf{X}_1$ (I x J_1) a matrix that contains the generic metabolome information, the matrix consists of I experiments and J_1 metabolites; $\mathbf{X}_2$ (I x J_2) a matrix that contains the measurements of specific interest, generally these measurements are not in $\mathbf{X}_1$; $\mathbf{X}$ (I x (J_1 + J_2)) the concatenated matrix of $\mathbf{X}_1$ and $\mathbf{X}_2$, i.e. $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2]$. The prior information in $\mathbf{X}_2$ depends on the biological question. For example, it could contain the measurements of the glycolysis intermediates, or it could contain the measurements of metabolites that belong to the same class, e.g. amino acids.

2.1 CPCA

CPCA [8] searches for common behavior in two data sets (Figure 1). As its name suggests, it is similar to a normal PCA analysis. It is, however, not straightforward to extract

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


A
X1 is the same size as X2
X1
X2
CPCA
CCA
CPCA identifies the major trends in the concatenated matrix of X1 and X2.
CCA identifies metabolites that correlate the most within the two data sets, although it does not necessarily explain much variation
In the extreme case that X1 and X2 are completely unrelated, the CPCA solution can be based on only X1 or X2.
CCA is always based on X1 and X2.
B
X1 is larger than X2
X1
X2
CPCA
CCA
CPCA makes X1 and X2 equally important by equalizing the variation in both matrices.
CCA is not influenced by size changes of X1 and X2.
The smallest matrix can dominate because: (i) the variables explain more variation; (ii) the smallest matrix is more homogeneous.

common variation from two different PCA models. The variation in the behavior of the selected metabolites and the remainder of the metabolome can be modeled with two PCA models:

(1) $$\min_{\mathbf{P}_1^{\mathbf{T}}\mathbf{P}_1=\mathbf{I}} \left\| \mathbf{X}_1 - \mathbf{T}_1\mathbf{P}_1^{\mathbf{T}} \right\|^2$$

for $\mathbf{X}_1$ and

(2) $$\min_{\mathbf{P}_2^{\mathbf{T}}\mathbf{P}_2=\mathbf{I}} \left\| \mathbf{X}_2 - \mathbf{T}_2\mathbf{P}_2^{\mathbf{T}} \right\|^2$$

for $\mathbf{X}_2$.

The symbols $\mathbf{T}_1$ (I x R_1) and $\mathbf{T}_2$ (I x R_2) represent the PCA scores and $\mathbf{P}_1$ (J_1 x R_1) and $\mathbf{P}_2$ (J_2 x R_2) represent the PCA loadings with R_1 and R_2 selected components for $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. Here, the scores $\mathbf{T}_1$ and $\mathbf{T}_2$ are different because they describe different aspects of variation of the experimental conditions. Therefore, additional steps are required to find the consensus behavior and the weights for the individual metabolites to this consensus.

CPCA models the behavior $\mathbf{X}_1$ and $\mathbf{X}_2$ as follows:

$$(3) \qquad \min_{\mathbf{P}_{\mathbf{sup}}^{\mathbf{T}}\mathbf{P}_{\mathbf{sup}}=\mathbf{I}} \left\| \mathbf{X} - \mathbf{T}_{\mathbf{sup}}\mathbf{P}_{\mathbf{sup}}^{\mathbf{T}} \right\|^2 = \min_{\mathbf{P}_{\mathbf{sup}}^{\mathbf{T}}\mathbf{P}_{\mathbf{sup}}=\mathbf{I}} \left\| [\mathbf{X}_1\ \mathbf{X}_2] - \mathbf{T}_{\mathbf{sup}} \begin{bmatrix} \mathbf{P}_{\mathbf{sup}_1} \\ \mathbf{P}_{\mathbf{sup}_2} \end{bmatrix}^T \right\|^2 .$$

The symbols $\mathbf{T}_{\mathbf{sup}}$ (I x R_{sup}) represent the CPCA score and $\mathbf{P}_{\mathbf{sup1}}$ (J_1 x R_{sup}) and $\mathbf{P}_{\mathbf{sup2}}$ (J_2 *x* R_{sup}) represent the PCA loadings with R_{sup} selected components for **X**. In this model, the variation in behavior of the experiments is captured by the common scores $\mathbf{T}_{\text{sup}}$. It is now straightforward to compare the weights of the different metabolites, as captured by the loadings $\mathbf{P}_{\text{sup1}}$ and $\mathbf{P}_{\text{sup2}}$, with each other. Furthermore, $\mathbf{T}_{\text{sup}}$ captures indeed the consensus behavior of $\mathbf{X}_1$ and $\mathbf{X}_2$.

To ensure that $\mathbf{X}_1$ and $\mathbf{X}_2$ both contribute equally to the model, they can be weighted to equal SS. This is especially important when one data matrix contains more variables than the other. In this case, when it is assumed that every metabolite has on average the same variation, it is likely that the data matrix with the most metabolites will be dominant in the data analysis. When the data matrices contain a similar amount of metabolites, the effect of block scaling is likely to be minimal (Figure 1A). As a consequence of equalizing the SS for the two data matrices, the SS per metabolite is changed. If the total SS of **X** is 100% then after block scaling $\mathbf{X}_1$ and $\mathbf{X}_2$ both contain 50% of the SS. If $\mathbf{X}_2$ is smaller than $\mathbf{X}_1$, the SS is divided over less metabolites and consequently these metabolites individually become more important. As a result of this, the behavior of the concentrations of the metabolites in the smallest block can become leading in the search for common behavior of metabolite concentrations in $\mathbf{X}_1$ and $\mathbf{X}_2$ (Figure 1B). In this paper we are interested in equal importance of the data blocksm as we want to analyze the behavior of the group of metabolites in $\mathbf{X}_2$ in relation $\mathbf{X}_1$.

There is also a second aspect that can increase the influence of the block containing the selected metabolites. The concentrations of the selected metabolites can have more homogeneous behavior than the concentrations of the metabolites in the remainder of the data set. This effect follows from the idea that the selected metabolites share a common biological background. For instance, they are chemically related or share the same regulation. This will make it easier for CPCA to identify main effects based on the selected block.

As for a normal PCA, other data pretreatment [10] steps can be taken before block scaling to emphasize different aspects of the data.

2.2 CCA

CCA searches for the largest correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$ (Figure 1). It does this by maximizing r = corr($\mathbf{X}_1\mathbf{a}$,$\mathbf{X}_2\mathbf{b}$). The vectors $\mathbf{u}$= $\mathbf{X}_1\mathbf{a}$ and $\mathbf{v}$= $\mathbf{X}_2\mathbf{b}$ are the so-called canonical variates, and describe the nature of the correlation. Besides this, the vectors **a** and **b** are the weights of the contributions of the different metabolites to the correlation detected.

Searching for the largest correlation between two data matrices can result in trivial results. First, the largest correlation could be based on the correlation between only one metabolite in each set. While the correlation is very strong, it could be only a minor effect in comparison to the total variation in both matrices. Second, when the data sets consist of more metabolites than experimental conditions, it is always possible to find perfect correlations (r = 1 or -1), and therefore the solutions will be trivial. It is possible to circumvent these effects by using a dimension reduction technique such as PCA. By reducing the data sets to their main effects, the principal components (PCs), the effects of single metabolites are limited to contributions to these components. Furthermore, the dimensionality is controlled by the number of components deemed relevant. Therefore the CCA analysis as applied in this paper, becomes the maximization of $r^* = \mathrm{corr}(\mathbf{T}_1^*\mathbf{a}_1^*, \mathbf{T}_2^*\mathbf{b}_2^*)$. Here, $\mathbf{T}_1^*$ and $\mathbf{T}_2^*$ are the selected PCs from the PCA decompositions (1) and (2). $r^* = \mathrm{corr}(\mathbf{T}_1^*\mathbf{a}_1^*, \mathbf{T}_2^*\mathbf{b}_2^*)$ can also be written in terms of the original matrices: $r^* = \mathrm{corr}(\mathbf{X}_1\mathbf{P}_1^*\mathbf{a}_1^*, \mathbf{X}_2\mathbf{P}_2^*\mathbf{b}_2^*)$. Here $\mathbf{P}_1^*$ and $\mathbf{P}_2^*$ are the loadings from the selected PCs from (1) and (2).

2.3 Validation

CPCA and CCA both provide information on the relative importance of every metabolite to the effects discovered by the analysis, namely the weights for each metabolite. These metabolite weights are the starting point for further exploration of the meaning of the results. It is therefore important that a certain degree of confidence of these metabolite weights can be obtained. For this, a validation scheme based on permutations is developed which is generic for CPCA and CCA.

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

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

3 Results

The use of CPCA and CCA was illustrated by their application on two different metabolomics data sets. The first data set consisted of metabolomes obtained from *Pseudomonas putida* S12 fermentations in which *P. putida* S12 was grown on four different carbon sources [10]. The $\mathbf{X}_2$ matrix (9 experiments, 19 metabolites) contained the concentrations of the measured nucleotides and the $\mathbf{X}_1$ matrix (9 experiments, 142 metabolites) contained the metabolome minus the nucleotides. This data set proved to be a straight forward data set with large effects induced by the selected experimental conditions. The second data set consisted of *Escherichia coli* metabolomes obtained from cells cultivated under 28 different experimental conditions aimed at inducing variation in the phenylalanine production (Table 1) [11]. $\mathbf{X}_2$ (28 experiments, 13 metabolites) contained all the measured intermediates of the phenylalanine biosynthesis pathway and $\mathbf{X}_1$ (28 experiments, 175 metabolites) contained the

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

Table 1 - Phenylalanine concentration in E. coli metabolomics samples. The phenylalanine concentrations are sorted in descending order.

remaining metabolome. This data set is a complex data set in which different effects play a role, like the environmental conditions and different growth phases in the batch process. None of concentrations of the metabolites were simultaneously in $\mathbf{X}_1$ and $\mathbf{X}_2$ to avoid trivial results.

3.1 CPCA

The CPCA analysis of the combined metabolome/nucleotide matrix from the *P. putida* S12 data set lead to a clear separation of the metabolomes resulting from growth on the four carbon sources on the first two PCs (Figure 2A). The metabolites that contribute to the first component were for $\mathbf{X}_1$ metabolites related to the carbon catabolism pathways and to central metabolism, such as, glyceraldehyde-3-phosphate, dihydroxyacetone phosphate, glucose-6-phosphate, and pyruvate (Table 2). For $\mathbf{X}_2$ most metabolites contributed significantly. It is noteworthy to see that the mono-phosphate (xMPs) and di-phosphate (xDPs) nucleotides had a positive contribution, while the tri-phosphate nucleotides (xTP) had a negative contribution. This observation suggests that the differences between growth on glucose as the sole carbon source on one hand, and succinate and fructose as sole carbon source on the other hand (Figure 2A) resulted in differences in the distribution of energy carrying molecules like the nucleotides. The variation explained for each **X** sub matrix was compared with the maximal explained variation possible for that matrix (Figure 2B). Both **X** sub matrices are very close to the maximal explained variation for the first PC. This indicated that the first PC indeed described a common direction in $\mathbf{X}_1$ and $\mathbf{X}_2$. After the first PC, the variation in $\mathbf{X}_2$ remained maximally explained while the variation $\mathbf{X}_1$ was not maximally explained. This indicated that the variation in $\mathbf{X}_2$ became more important in the analysis, except for PC 4 where $\mathbf{X}_1$ became dominant after $\mathbf{X}_2$ was almost fully explained.

For the more complex *E. coli* data set, the score plots of the first PC of the CPCA analysis (Figure 3A) showed an effect related to high and low phenylalanine productivity (Table 1) in the first PC. The most important metabolites relating to this effect were for $\mathbf{X}_1$ phenyllactate, 3,5-dihydroxypentanoate (tentatively identified), a number of unidentified metabolites, and the amino acids valine and isoleucine (Table 3). For $\mathbf{X}_2$, the most important metabolites were phenylalanine and metabolites that are regularly important [12] in the phenylalanine biosynthesis route, such as, chorismate and erythrose-4-phosphate. The enzymes that convert these metabolites are subject to end product inhibition [12].

For the second PC there was not a clear explanation for the behavior of the experimental conditions. The most important metabolites for this PC, however, suggested that the PC was related to general amino acid metabolism. The most important metabolites were for $\mathbf{X}_1$ urea, aspartate, malate, and fumarate: these metabolites are part of the citric acid cycle and the urea cycle. Also the amino acids isoleucine and valine were important for this PC as well as for the first PC. For $\mathbf{X}_2$, important metabolites were glutamate and ketoglutarate, used in amino group transfer reactions; tyrosine and tryptophan, end products of the other branch of the aromatic amino acid biosynthesis pathway; and phenylpyruvate, the precursor to phenylalanine. The third PC seemed to describe a time effect as is indicated with arrows for fermentations 4, 5, and 9 (Figure 3B). The most important metabolites for $\mathbf{X}_1$ consisted of unidentified metabolites, UDP-N-AAGDAA (UDP-*N*-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminoheptanedioate-D-alanyl-D-alanine), N-acetylglutamate, the nucleotides CMP, CDP, and UMP and tymine. For $\mathbf{X}_2$ prephenate and phosphoenolpyruvate were significant. UDP-N-AAGDAA and other metabolites in the top 20 of most important metabolites for PC 3 (UDP-N-AAGD (UDP-*N*-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminoheptanedioate) and UDP-glucose) are part of the peptidoglycan biosynthesis pathway and thus related to cell wall synthesis, which is in line with the observed time effect in Figure 3B. The changes in the cell wall synthesis could be related to the shift from exponential growth phase to the stationary growth phase.

The comparison of the explained variation per X block with the maximal explained variation for that X block showed that the CPCA analysis seemed to depend most on $\mathbf{X}_{2.}$ The explained variance of $\mathbf{X}_2$ in the solution closely followed the maximal explained variation (Figure 3C), while this is not the case for $\mathbf{X}_1$. This can be caused by two effects; first, $\mathbf{X}_2$ contains much less metabolite concentrations than $\mathbf{X}_1$, and second, $\mathbf{X}_2$ is more homogeneous than $\mathbf{X}_1$ because the selected metabolites are part of the same pathway.

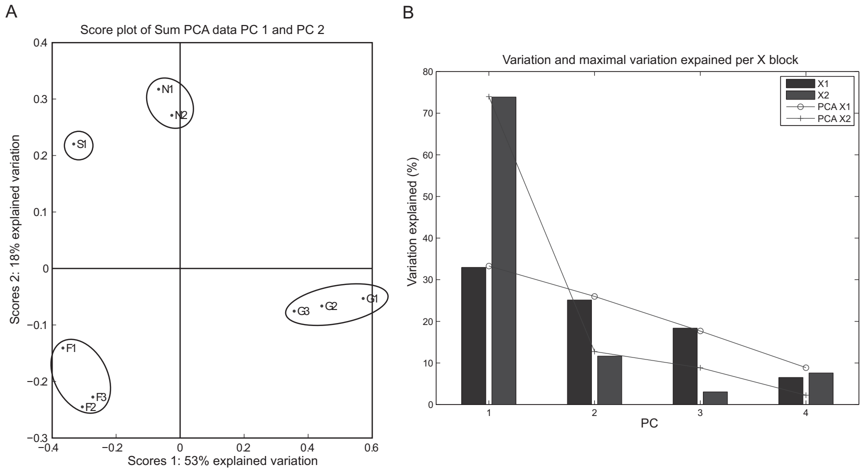


Figure 2 - CPCA results of the P. putida S12 data set.(A) The score plots of the super scores, the metabolome samples obtained from fermentations with the same carbon source are circled. N, S, G, and F refer respectively to gluconate, succinate, D-glucose, and D-fructose as sole carbon source in the fermentation. (B) The explained variation per data block (bars) and the maximal explained variance for that data block (lines).

	X1				X2			
	PC 1		PC 2		PC 1		PC 2	
	Weight	Metabolite	Weight	Metabolite	Weight	Metabolite	Weight	Metabolite
1	0.096	glyceraldehyde-3-phosphate	0.145	adenine	0.242	CMP	0.279	ADP
2	0.091	BAC-607-N1058	0.14	putrescine	0.232	UMP	0.259	CDP
3	0.091	isomaltose	0.136	BAC-607-N1038	-0.226	ATP	0.212	GDP
4	0.087	uridine*	0.136	BAC-607-N1021	0.226	AMP	-0.208	TMP
5	0.087	dihydroxyacetone phosphate	0.133	thymine	-0.224	ITP	0.165	GTP
6	0.085	uridine*	0.132	BAC-638-N1003	-0.216	UTP	0.126	UTP
7	0.084	glucose-6-phosphate	-0.128	BAC-647-N1012	0.214	UMP	-0.1	AMP
8	0.096	glyceraldehyde-3-phosphate	-0.128	BAC-647-N1013	0.242	CMP	-0.096	CMP
9	0.091	BAC-607-N1058	0.122	ketogluconate	0.232	UMP	0.074	UDP
10	0.091	isomaltose	0.116	BAC-607-N1073	-0.226	ATP	0.067	IMP

*Table 2 – Metabolite contributions to P. putida S12 CPCA model. The top 10 most important metabolites are shown. The grey areas indicate contributions which are not significant after permutation. * Uridine occurs twice because it is measured with GC-MS and LC-MS.*

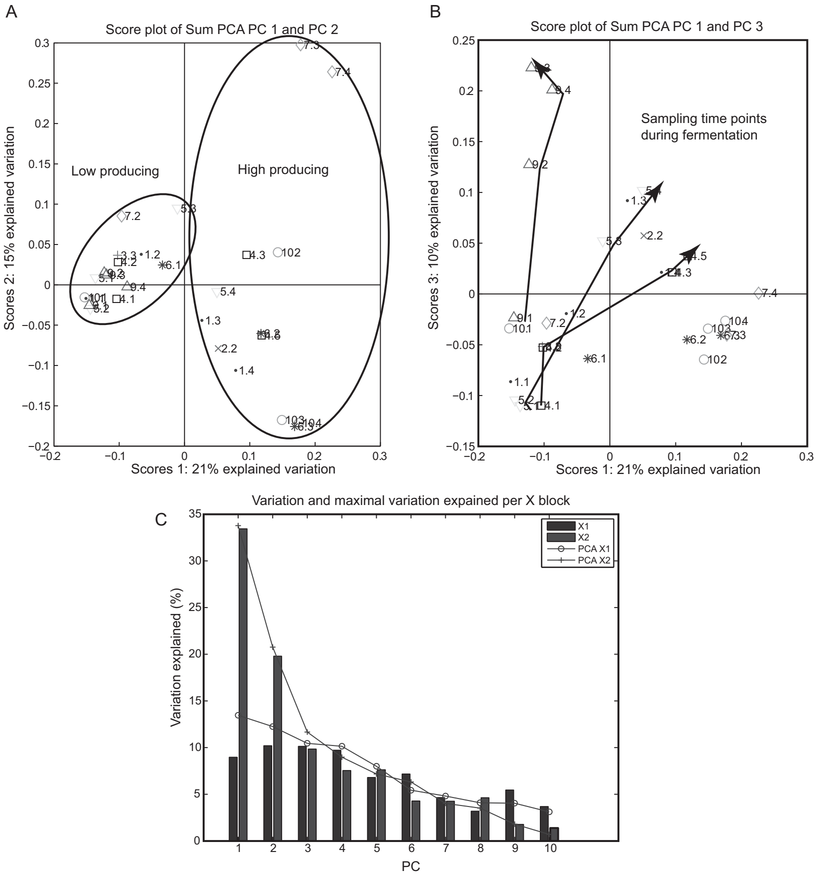


Figure 3 - CPCA results of the E. coli data set. (A) The score plots of the super scores for PC 1 and 2. The circles indicate the difference between experimental conditions that resulted in high and low phenylalanine production (Figure 4). (B) The score plots of the super scores for PC 1 and 3. The arrows indicate the order of sampling from early to late time points during the batch fermentation. (C) The explained variation per data block (bars) and the maximal explained variance for that data block (lines).

X1

	PC 1		PC 2		PC 3	
	Weight	Metabolite	Weight	Metabolite	Weight	Metabolite
1	0.109	3-phenyl-lactate or isomer	0.112	urea	0.144	spectrum not found7
2	0.096	3,5-dihydroxy-pentanoate	0.103	aspartate	-0.141	mixed spectrum5
3	-0.094	mixed spectrum3	0.103	fumarate	0.134	UDP-N-AAGDAA
4	-0.088	spectrum not complete6	0.103	malate	0.13	spectrum not found5
5	-0.08	sugar no oxim	0.099	2-hydroxy-glutarate	0.118	N-acetylglutamate
6	0.075	isoleucine	0.098	2,3-dihydroxy-3-methyl-butanoate	0.116	thymine
7	0.074	valine	0.097	pantoate	0.115	CMP
8	0.072	unknown mass 304, 319 and 406	0.096	unknown7	0.112	CDP
9	-0.07	Disaccharide4	0.094	unknown28	0.111	spectrum not complete5
10	-0.068	unknown8	0.094	organic acid with mass 261	0.107	UMP

X2

	PC 1		PC 2		PC 3	
	Weight	Metabolite	Weight	Metabolite	Weight	Metabolite
1	0.47	phenylalanine	0.452	glutamate	0.523	prephenate
2	0.435	chorismate	0.363	ketoglutarate	0.369	phosphoenol-pyruvate
3	0.314	erythrose-4-phosphate	0.323	phenylpyruvate	0.157	shikimate
4	0.304	phenylpyruvate	-0.278	tyrosine	0.138	erythrose-4-phosphate
5	0.291	tyrosine	-0.231	tryptophan	-0.112	shikimate-3-phosphate
6	0.163	shikimate-3-phosphate	-0.197	phenylalanine	-0.09	phenylpyruvate
7	0.163	glutamate	0.168	shikimate	0.08	3-dehydroquinate
8	0.135	ketoglutarate	-0.113	erythrose-4-phosphate	0.059	tryptophan
9	0.129	tryptophan	0.077	shikimate-3-phosphate	-0.056	ketoglutarate
10	0.073	3-dehydroquinate	0.057	chorismate	0.052	chorismate

Table 3 – Metabolite contributions to E. coli S12 CPCA model. The top 10 most important metabolites are shown. The grey areas indicate contributions which are not significant after permutation.

3.2 CCA

CCA searches for the largest correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$. For the *P. putida* S12 data set of both $\mathbf{X}_1$ and $\mathbf{X}_2$ the dimensions were reduced by PCA after range scaling. For $\mathbf{X}_1$ four and for $\mathbf{X}_2$ three PCs were used. The correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$ was very large - all the experiments are on the diagonal line - and the significant association is 0.999 (Figure 4). This value for the association was significant after validation by permutation of the experimental conditions and repetition of the data analysis. The metabolites responsible for this large correlation were for $\mathbf{X}_1$ metabolites related to catabolic pathways, such as, glyceraldehyde-3-phosphate, dihydroxyacetone-phosphate, and glucose-6-phosphate (Table 4). This was similar to the CPCA results. The responsible metabolites for $\mathbf{X}_2$ were the xMPs and the xTPs. Unlike the CPCA results, the xDPs were less important. For both data sets, the variation modeled by the correlation between the two matrices was close to the maximal explained variation for those matrices. This indicated that the behavior of the metabolite concentrations in $\mathbf{X}_1$ and $\mathbf{X}_2$ correlates very well and that the correlation is a major effect in the behavior of these concentrations.

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

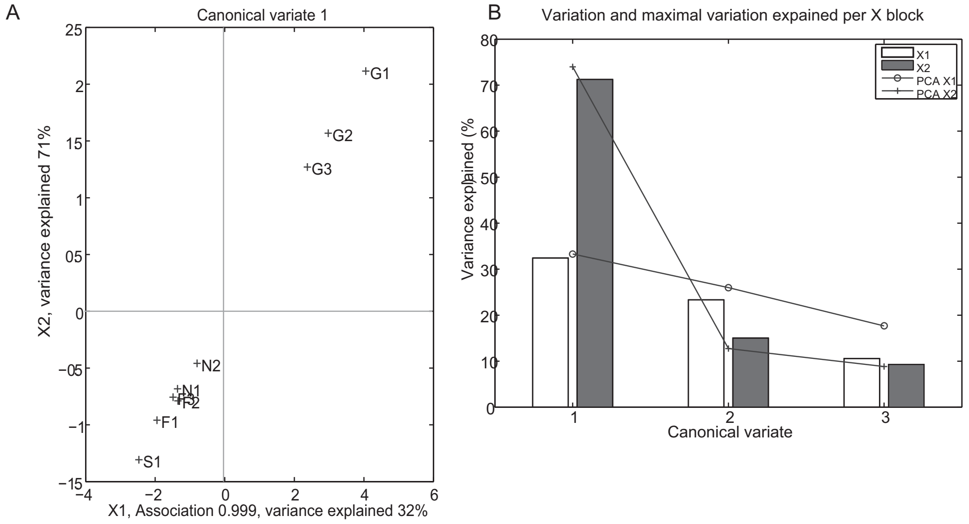


Figure 4 - CCA results for the P. putida S12 data set. (A) The nature of the correlation of canonical variate 1. (B) The explained variation per data block (bars) and the maximal explained variation for that data block (lines).

	X1		X2	
	Variate 1		Variate 1	
	Weight	Metabolite	Weight	Metabolite
1	0.175	glyceraldehyde-3-phosphate	0.351	CMP
2	0.164	dihydroxyacetone phosphate	-0.307	GTP
3	0.16	BAC-607-N1058	-0.284	UTP
4	0.159	isomaltose	0.282	TMP
5	0.153	uridine	0.282	AMP
6	0.151	glucose-6-phosphate	0.265	AMP
7	0.151	sugar phosphate	0.243	UMP
8	0.149	gluconic acid lacton	0.242	UMP
9	0.147	BAC-607-N1102	-0.242	ITP
10	0.146	BAC-629-N1028	0.242	GMP

Table 4 - Metabolite contributions to P. putida S12 CCA model. The top 10 most important metabolites are shown. The grey areas indicate contributions which are not significant after permutation.

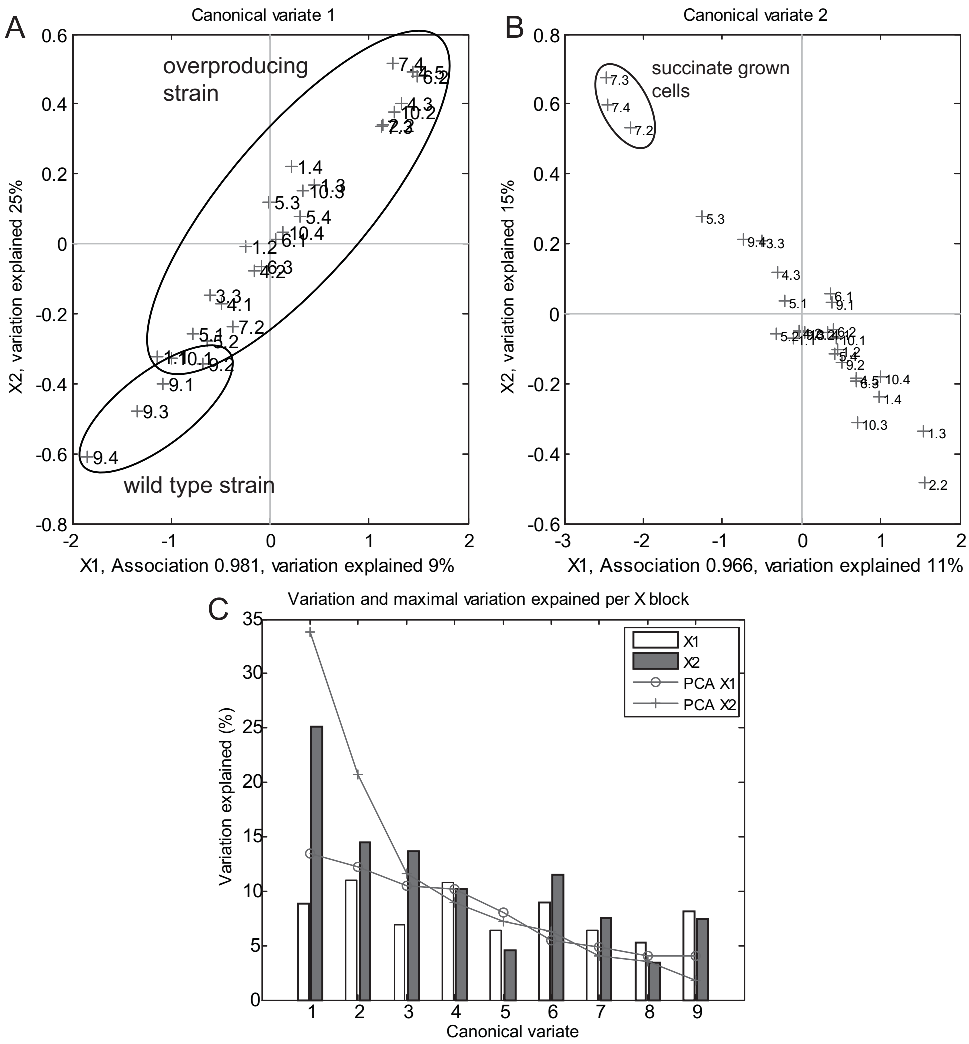


Figure 5 - CCA results for the E. coli data set. (A) The nature of the correlation of canonical variate 1. The ovals indicate grouping of the metabolomes resulting from fermentations with wild type and the overproducing strain. (B) The nature of the correlation of canonical variate 2. The metabolomes resulting from succinate grown cells are indicated with a circle. (C) The explained variation per data block (bars) and the maximal explained variation for that data block (lines).

	X1				X2			
	Variate 1		Variate 2		Variate 1		Variate 2	
	Weight	Metabolite	Weight	Metabolite	Weight	Metabolite	Weight	Metabolite
1	-0.182	unknown 8	-0.188	urea	0.65	phenylalanine	0.622	shikimate
2	-0.171	spectrum not found5	-0.178	pantoate	0.391	3-dehydroquinate	0.404	ketoglutarate
3	0.158	unknown 7	0.17	isoleucine	-0.353	tryptophan	-0.369	phenylalanine
4	0.154	unknown 32	-0.168	spectrum not specific	-0.277	tyrosine	-0.305	phosphoenolpyruvate
5	0.153	unknown mass 304, 319 and 406	-0.163	fumarate	0.274	erythrose-4-phosphate	0.286	phenylpyruvate
6	0.152	ribulose (?)	-0.159	adenosine	-0.229	prephenate	0.22	glutamate
7	0.149	FMN	-0.149	sugar phosphate 4	0.193	shikimate-3-phosphate	-0.212	shikimate-3-phosphate
8	0.148	spectrum not found3	-0.147	malate	0.171	phenylpyruvate	0.143	chorismate
9	-0.147	N-acetyl-aspartate + β-phenylpyruvate	-0.145	acetyl-amino acid	0.123	chorismate	-0.101	prephenate
10	0.145	3-phenyllactate (?)	-0.14	aspartate	0.1	glutamate	0.089	tryptophan

Table 5 - Metabolite contributions to E. coli CCA model. The top 10 most important metabolites are shown. The grey areas indicate contributions which are not significant after permutation. (?) indicates a metabolite whose identification is not certain.

4 Discussion

CPCA and CCA are valuable methods to emphasize specific areas of the metabolic network in data analysis of metabolomics data. They make it possible to focus on groups of metabolites be it functionally or chemically related metabolites as for the *P. putida* S12 data, or a metabolic pathway as for the *E. coli* data; both methods result in biologically meaningful results.

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

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

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

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

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

5 Methods

5.1 Data

The first data set consisted of *P. putida* S12 [13] metabolomes. Cultures of *P. putida* S12 were grown as previously described [14]. In short, samples were grown in triplicate on four carbon sources: D-fructose (sample F1, F2 and F3), D-glucose (sample G1, G2 and G3), gluconate (sample N1 and N2) and succinate (sample S1). Samples were analyzed by GC-MS [15] and LC-MS [16]. The GC-MS and LC-MS data set were fused together by concatenating the measurement tables [11]. The final data set was manually cleaned up, removing spurious and double entries and consisted of 9 experiments and 161 metabolites. The second data set consisted of *E. coli* metabolomics (*E. coli* NST 74, a phenylalanine overproducing strain, and *E. coli* W3110, the wild-type strain). The *E. coli* strains were grown under different experimental conditions as described elsewhere [11]. Samples were analyzed by GCMS [15] and LCMS [16] and fused together [11]. The final data set was manually cleaned up, removing spurious and double entries and consisted of 28 experiments and 188 metabolites.

5.2 Data analysis

CPCA [8] and CCA [9] were implemented in Matlab 7.3.0 [17]. In the data analysis, the data was range scaled [10]. The significance of the data analysis results were validated as described in the text.

6 Acknowledgements

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

7 References

1. Jolliffe IT: *Principal Component Analysis.* New York: Springer-Verlag; 2002.
2. Jackson JE: *A user's guide to principal components.* John Wiley & Sons, Inc.; 1991.
3. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
4. Kaufman L, Rousseeuw PJ: *Finding groups in data: an introduction to cluster analysis.* Wiley-Interscience; 1990.
5. Geladi P, Kowalski BR: **Partial least-squares regression: a tutorial.** *Anal Chim Acta* 1986, **185:**1-17.
6. Höskuldsson A: **PLS regression methods.** *J Chemom* 1988, **2:**211-228.
7. Hoogerbrugge R, Willig SJ, Kistemaker PG: **Discriminant Analysis by Double Stage Principal Component Analysis.** *Anal Chem* 1983, **55:**1710-1712.

8. Smilde AK, Westerhuis JA, de Jong S: **A framework for sequential multiblock component methods.** *J Chemom* 2003, **17:**323-337.
9. Krzanowski WJ: *Principles of Multivariate Analysis, a User's Perspective.* New York: Oxford University Press Inc.; 1988.
10. van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ: **Centering, scaling, and transformations: improving the biological information content of metabolomics data.** *BMC Genomics* 2006, **7**.
11. Smilde AK, van der Werf MJ, Bijlsma S, van der Werff-van der Vat B, Jellema RH: **Fusion of mass-spectrometry-based metabolomics data.** *Anal Chem* 2005, **77:**6729-6736.
12. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham JL, Paley S, Paulsen IT, Peralta-Gil M, Karp PD: **EcoCyc: a comprehensive database resource for Escherichia coli.** *Nucleic Acids Res* 2005, **33:**D334-D337.
13. Hartmans S, van der Werf MJ, de Bont JAM: **Bacterial degradation of styrene involving a novel flavin adenine dinucleotide-dependent styrene monooxygenase.** *Appl Environ Microbiol* 1990, **56:**1347-1351.
14. van der Werf MJ, Pieterse B, van Luijk N, Schuren F, van der Werff-van der Vat B, Overkamp K, Jellema RH: **Multivariate analysis of microarray data by principal component discriminant analysis: prioritizing relevant transcripts linked to the degradation of different carbohydrates in *Pseudomonas putida* S12.** *Microbiology* 2006, **152:**257-272.
15. Koek M, Muilwijk B, van der Werf MJ, Hankemeier T: **Microbial metabolomics with gas chromatography mass spectrometry.** *Anal Chem* 2006, **78:**1272-1281.
16. Coulier L, Bas R, Jespersen S, Verheij E, vanderWerf MJ, Hankemeier T: **Simultaneous Quantitative Analysis of Metabolites Using Ion-Pair Liquid Chromatography-Electrospray Ionization Mass Spectrometry.** *Anal Chem* 2006, **78:**6573-6582.
17. The Mathworks Inc.. **Matlab 7.3 (R2006b).** 2006.

5 Genetic algorithm based two-mode clustering of metabolomics data

Jos A. Hageman, Robert A. van den Berg, Johan A. Westerhuis, Mariët J. van der Werf, and Age K. Smilde

Summary

Metabolomics and other omics tools are generally characterized by large data sets with many variables obtained under different environmental conditions. Clustering methods, and more specifically two-mode clustering methods, are excellent tools for analyzing this type of data. Two-mode clustering methods allow for analysis of the behavior of subsets of metabolites under different experimental conditions. In addition, the results are easily visualized. In this paper we introduce a two-mode clustering method based on a genetic algorithm that uses a criterion that searches for homogeneous clusters. Furthermore we introduce a cluster stability criterion to validate the clusters and we provide an extended knee plot to select the optimal number of clusters in both experimental and metabolite modes.

The genetic algorithm-based two-mode clustering gave biologically relevant results when it was applied to two real life metabolomics data sets. It was, for instance, able to identify a catabolic pathway for growth on several of the carbon sources.

This chapter is published J.A. Hageman, R.A. van den Berg, J.A. Westerhuis, M.J. van der Werf, and A.K. Smilde. Metabolomics, 2008, 4 (2): 141-149.

1 Introduction

Functional genomics approaches have been applied in many different areas for the unraveling of complex biological questions. A functional genomics approach aims to obtain a complete overview of a certain biological response, for instance, gene expression levels or metabolite concentrations, in relation to the experimental conditions of interest. Obtaining a complete overview of the biological response enables the identification of interesting effects that would not be noticed if a subset of the genes or metabolites is analyzed.

Within functional genomics, metabolomics focuses on the analysis of the metabolome, the complete set of small organic molecules in, or outside, a cell. The metabolome is the most direct reflection of the phenotype of the organism under study, as regulatory effects, like post-transcriptional processing, or post-translational modification, do not hamper its interpretation [1]. In a metabolomics experiment, metabolome samples of an organism are generated under conditions that result in (large) variations of the metabolome composition.

The resulting variations are often analyzed with latent variable techniques or clustering methods. Latent variable techniques, such as PCA [2], PCDA [3], reduce the dimensions of the data to make interpretation easier. Clustering methods, on the other hand, order the data in groups that are similar according to a particular similarity measure, such as the Euclidean distance, or the correlation coefficient [4,5]. The popularity of clustering methods results from their visualization and clear interpretation.

Clustering methods can be divided in two groups. The first group clusters the data set in either experiment or metabolite clusters; this is called one mode clustering. Here, the experiments or the metabolites are clustered based on the similarity of the behavior of all metabolite concentrations under an experimental condition or on the similarity of behavior of the concentration of a metabolite under all experimental conditions, respectively. The second group simultaneously creates experiment and metabolite clusters, which is called two-mode clustering or biclustering [6,7]. Here the metabolites and experiments are clustered simultaneously to obtain groups of experiments and metabolites that behave as similar as possible. It is possible to apply a one-mode clustering method (e.g. hierarchical clustering, or k-means clustering) first to the metabolite mode and subsequently to the experiment mode, or vice versa. However, this will not result in identical results as by using two-mode clustering, as the clusters are not optimized for homogeneity in both the experimental and the metabolite mode. Therefore, two-mode clusters obtained by one-mode clustering methods are sub-optimal and the interpretation of these results will be hampered.

Two-mode clustering algorithms aim to find the best partitioning of the data in clusters. We define the best partitioning as the cluster assignment which results in the minimal difference between the model of the data and the original data. Different two-mode clustering algorithms exist, of which some algorithms are based on global optimization approaches, such as Simulated Annealing (SA) and Tabu Search (TS) [6,8]. The main advantage of global optimization methods is that they are able to find the global solution and not a locally optimal solution; something that is likely to happen with local optimization methods like steepest descent.

In this paper we introduce two-mode clustering of metabolomics data based on a Genetic Algorithm (GA). As GA's work on a group of solutions it can take large steps in the solution space and it is less likely to get stuck in local optima compared to SA and TS. The GA approach used in this paper is based on a cluster homogeneity criterion and not on distances between clusters. This means that clusters are based on metabolites that behave as similar as possible for a group of experimental conditions. Furthermore, quite some attention is paid to assess the cluster stability using a leave one out resampling of the two-mode clustering results. The selection of the number of clusters in both experimental and metabolite modes is performed using a generalized knee plot. Most two-mode clustering methods are specifically designed for gene expression data, but we apply our new two-mode clustering approach to metabolomics data which improves their interpretation considerably. Two different metabolomics data sets with different complexity are analyzed to show the generality and usefulness of the new method.

2 Methods and Materials

2.1 Data

The first data set *(Pseudomonas putida* S12) is maintained at TNO (Zeist, the Netherlands). Cultures of *P. putida* S12 [9] were grown in batch fermentations at 30°C in a Bioflow II (New Brunswick Scientific) bioreactor as previously described [10]. In short, samples were grown in triplicate on four carbon sources: D-fructose (sample F1, F2 and F3), D-glucose (sample G1, G2 and G3), gluconate (sample N1 and N2) and succinate (sample S1). Samples were analyzed by GC-MS and LC-MS. A detailed description is given elsewhere [11-13]. The GC-MS and LC-MS data set were fused together by concatenating the measurement tables [14]. The final data set was manually cleaned up, removing spurious and double entries and consisted of 9 experiments and 162 metabolites.

The second data set (*Escherichia coli* NST 74, a phenylalanine overproducing strain, and *E. coli* W3110, the wild-type strain) were grown at 30°C in a bioreactor containing 2 liters of a medium with 30 g/l glucose as carbon source. A constant pH (pH 6.5) and oxygen tension (30 %) was maintained. Samples were taken from the bioreactor after 16, 24, 40, 48 hours, and immediately quenched. Variations in this standard fermentation protocol were introduced by changing one of the default conditions, resulting in a screening experiment. Samples were analyzed by GC-MS and LC-MS and fused together. A detailed description of this data set is given elsewhere [14]. The final data set was manually cleaned up, removing spurious and double entries and consisted of 28 experiments and 188 metabolites.

2.2 Genetic Algorithms

GAs are a special class of global optimizers based on the theory of evolution. A GA minimizes a function F(x), where x represents a parameter vector, by searching the parameter space of x for the optimal solution. In the case of two-mode clustering, GAs will search for the optimal partitioning of objects and variables by minimizing the residuals. The residuals are the difference between the model of the data and the original data matrix. Several steps in a GA are identical for all GAs and will be explained shortly in the following.

1. Initialization: GAs operate on a group of solutions, called a population. At the start of the GA, all solutions, also called strings or chromosomes, are set to random values.
2. Evaluation: All strings in the population are evaluated by an evaluation function (see Section 2.3.1).
3. Stop: A stop criterion is checked.
4. Selection: A percentage of the best strings in a population is selected to form the next generation.
5. Recombination: To form the new population, new solutions are created by combining two selected existing solutions (parents) to yield two different ones (children). This is called crossover.
6. Mutation: Parts of a string in the new population are selected randomly and modified. To prevent the search from random behavior, the probability of mutation is usually chosen to be quite low.

Several parameters, such as the rate of crossover and mutation, regulate the performance of the GA. Each specific optimization problem has its own specific set of parameters for which the GA performs at its optimum. This so-called meta-optimization of the GA parameters can be tedious and can be considered a disadvantage of GAs in general. For more information regarding GAs, we refer to [15].

2.3 Two-mode clustering

2.3.1 The model

The goal of two-mode clustering is to simultaneously find the optimal partitioning between objects and variables of data matrix X, as depicted in Figure 1. For two-mode clustering, data matrix X is approximated by

(1) $\mathbf{X} = \mathbf{U}\mathbf{Y}\mathbf{V}^{\mathbf{T}} + \mathbf{E}$

Where

X (*M* x *N*): data matrix of *M* rows and *N* columns.

U (*M* x *P*): membership matrix for *M* rows (metabolites) of matrix **X** allowing for *P* row clusters. This matrix contains on each row (*P*-1) zeros and a single 1. The location of this 1 indicates the cluster membership.

Y (*P* x *Q*): matrix containing the clusters averages for *P* row and *Q* column clusters.

V (*N* x *Q*): membership matrix for *N* columns (experiments) of matrix **X** allowing *Q* column clusters. The structure of this matrix is similar to that of matrix **U**.

E (*M* x *N*): matrix containing the difference between each measurement and the average of the cluster it belongs to.

A schematic representation of this decomposition is given in Figure 2.

Pretreatment of the data is an important aspect of data analysis that can dramatically influence the results of data analysis [11]. In this paper, range scaling was applied to accentuate the biological information content of the metabolomics data set by converting the concentrations to values relative to the biological range of a metabolite. The biological range is defined as the difference between the minimum and maximum concentration measured for a metabolite in the data set. In this way, high or low metabolite concentrations

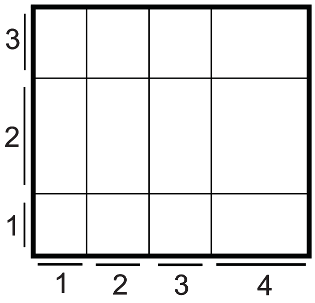


Figure 1 - Schematic representation of two-mode partitioning

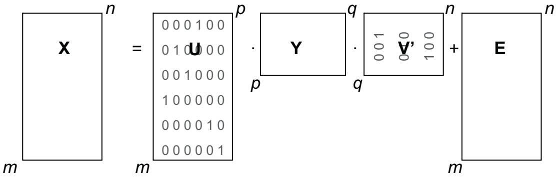


Figure 2 - Schematic representation of the decomposition of matrix X. See text for details.

and the way in which the concentrations of metabolites are affected by different environmental conditions are seen within the context of the natural variation of the concentration (dynamic range) of those metabolites.

2.3.2 Evaluation function

For the evaluation function, the partitioning information on the string is used to construct membership matrices **U** and **V**. Matrix **Y** is obtained in two steps. In the first step, the sums of all metabolites in a cluster are obtained:

$$\text{(2a)} \qquad \tilde{\mathbf{Y}} = \mathbf{U}^{\mathbf{T}}\mathbf{X}\mathbf{V}$$

In the second step, all elements are divided by the number of members in that cluster to obtain cluster averages in Y.

$$\text{(2b)} \qquad y_{p,q} = \frac{\tilde{y}_{p,q}}{u_p \cdot v_q}$$

Here $y_{p,q}$ is the average value of a two-mode cluster (*p*,*q*), u_p and v_q indicate the number of metabolites and experiments respectively for two-mode cluster (*p*,*q*)

The residual matrix **E** is then given by:

$$\text{(3)} \qquad \mathbf{E} = \mathbf{X} - \mathbf{U}\mathbf{Y}\mathbf{V}^{\mathbf{T}}$$

Matrix $\mathbf{UYV}^{\mathrm{T}}$ is the approximation of **X** and contains for each metabolite a value equal to its cluster average. For an optimal two-mode clustering result, the GA minimizes the sum of squares (SS) of the elements of **E**. The smaller the values in **E**, the tighter the corresponding clusters are.

2.3.3 Software

The two-mode genetic algorithm clustering method was programmed in Matlab 7.1 [16] using the Genetic Algorithm and Direct Search (GADS) [17]. A special integer type coding scheme was written for use with this toolbox. This scheme encodes the cluster number for each *M* metabolites and *N* experiments, so each string in the GA population has length *M*+*N*. The cluster number is an integer between 1 and the maximum number of clusters. The mutation operator replaces, with a certain probability, a value from the string with a random number between 1 and the maximum number of clusters. The settings used for the GA are listed in Table 1.

All GA runs were executed in five-fold with different random seeds to exclude any (un)lucky starting positions. The results from the five runs should be similar, and the best solution is chosen. The evaluation function was optimized for speed using the profile

Description of GA-parameters	Value
Data Type	Integer
Population Size	200
Mutation rate	0.005
Number of Generations	4000
Crossover Rate	0.8

Table 1 – Settings of the genetic algorithm

function of Matlab, resulting in run-times of five minutes for five replicate runs for the *P. putida* S12 data set and run-times of ten minutes for the *E. coli* data set. Since two-mode k-means is a local optimizer and is known to get easily stuck in local optima, the two-mode k-means was restarted 50 times for each solution and the best solution out a possible 50 was kept. All calculations were performed on an AMD Athlon XP 2400+ 2.00 GHz 512 MB RAM PC running Windows XP. The GA two-mode clustering routines applied in this paper are available at http://www.bdagroup.nl.

2.4 Number of clusters

Partitioning clustering algorithms require a predefined number of clusters. There are a number of methods for finding the most suited number of clusters in the data, such as, the Bayesian Information Criterion (BIC) [18] the GAP statistic [19] and the knee or 'L' method[20].

We chose the knee method which finds the knee or 'L' in a plot of the number-of-clusters versus the SS of the residuals. The assumption of this method is that an additional cluster gives a sharp decrease in the SS of the residuals as long as the optimal number of clusters is not reached. When more than the optimal number of clusters is chosen, the decrease in SS of the residuals is less sharp and more or less equal for each additional cluster.

The knee method can be generalized to two-mode clustering. In this case, the curve of the number-of-clusters versus the sum of squared residuals plot is a contour plot. In this plot there is a combination of cluster numbers for the experiments and metabolites for which an additional cluster no longer sharply decreases the SS of the residuals.

2.5 Validation

The two-mode clustering method was validated by leaving one experiment out (LOO) of the data set, clustering this data set again and comparing the obtained results with the clustering of the full data set. In this way, the dependence of the clustering on one single experiment can be assessed. A stable clustering will less likely be influenced by leaving one experiment out. For the *P. putida* S12 data set, at least one experiment per group remained in the data set to maintain the structure of the experimental design. All LOO-data sets were pretreated and clustered. When comparing the content of a cluster obtained with the LOO procedure, it was made sure that it was compared with the correct cluster obtained with the

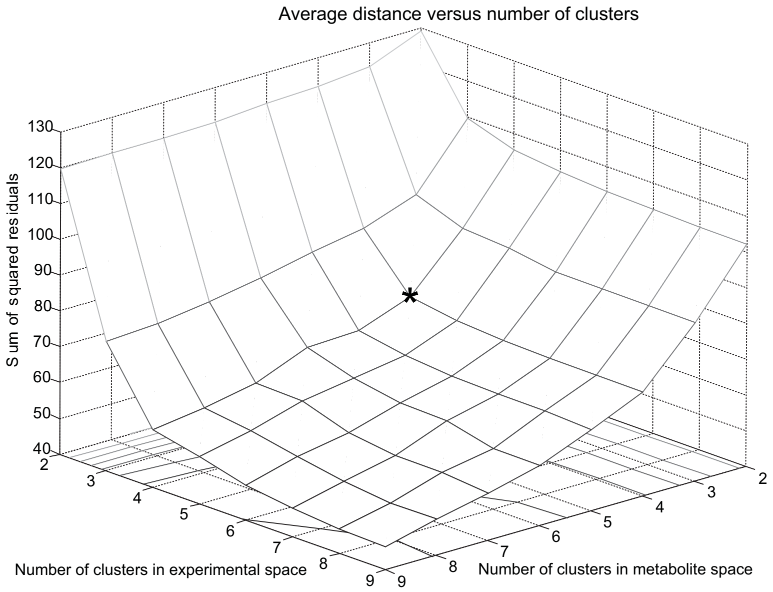


Figure 3 - The number-of-clusters versus the sum of squared residuals plot for the metabolome data set of P. putida S12. Point '4 experimental' and '4 metabolite' clusters is the combination of clusters where an increase in the number of clusters no longer sharply decreases the sum of squared residuals.

complete data by first establishing which clusters have to most overlap and linking them together. The LOO validating scheme only validates the effect of the experiments on the metabolite clustering. If desired, it is possible to validate the effect of the metabolites on the experiment clustering in a similar way.

In order to analyze the spread within clusters, the cluster variances are used as a diagnostic tool:

$$s_k^2 = \frac{\sum_{n_k=1}^{N_k} (x_{n_k} - y_k)^2}{N_k - 1} \tag{5}$$

Here, x_{n_k} indicates the cluster element n of cluster k for a total of N_k elements and y_k is the mean of the cluster k. The variances of the different clusters can be compared; a relatively low variance indicates small and compact clusters. In contrast, a relatively high variance indicates large and/or heterogeneous clusters and this could be a sign of, for instance, outliers.

The cluster variances are a natural diagnostic of the cluster quality as they are directly linked to the evaluation function. The variances of each two-mode cluster can be combined to give the pooled variance:

$$s_{pooled}^2 = \frac{\sum_{k=1}^{K}(N_k - 1)s_k^2}{\sum_{k=1}^{K}(N_k - 1)} \tag{6}$$

The evaluation function (Eq. 3) and the pooled variance are identical up to a scaling factor as is proven in Appendix A.

3 Results

3.1 Estimation of the number of clusters

3.1.1 *P. putida* data

The generalized knee method is used to obtain an estimate of the number of clusters in the partitioning. The rate of decrease for the residuals became smaller after four experimental clusters and four/five metabolite clusters (Figure 3). Obtaining four experiment clusters may seem trivial, however, it is possible that some of the experiments are rather similar and end up in the same cluster. For the metabolite clusters, both the four and five cluster solutions were analyzed and the five cluster choice was found to be more meaningful.

When comparing the results from the two-mode clustering with two single k-means clustering on the metabolites and the experiments (results not shown), the sum of squared residuals was 7.8% lower when applying two-mode clustering. The data set was also subjected to a classical non-GA based two-mode k-means method with the same evaluation function as the GA two-mode algorithm [21,22]. Figure 4 shows the comparison of the resulting sum-of-squares. For a larger number of clusters, GA tends to give better results than two-mode k-means. In the cases that two-mode k-means has a lower sum-of-squares it is usually only lower by a small amount, indicating that both algorithms have reached the same global minimum but with different precisions.

3.1.2 *E. coli* data

A similar analysis was performed for the *E. coli* data showing seven experimental clusters and six metabolite clusters was optimal. The performance of GA against k-means was again tested (see Figure 4) and showed that relatively quickly the GA outperforms the two-mode k-means solution.

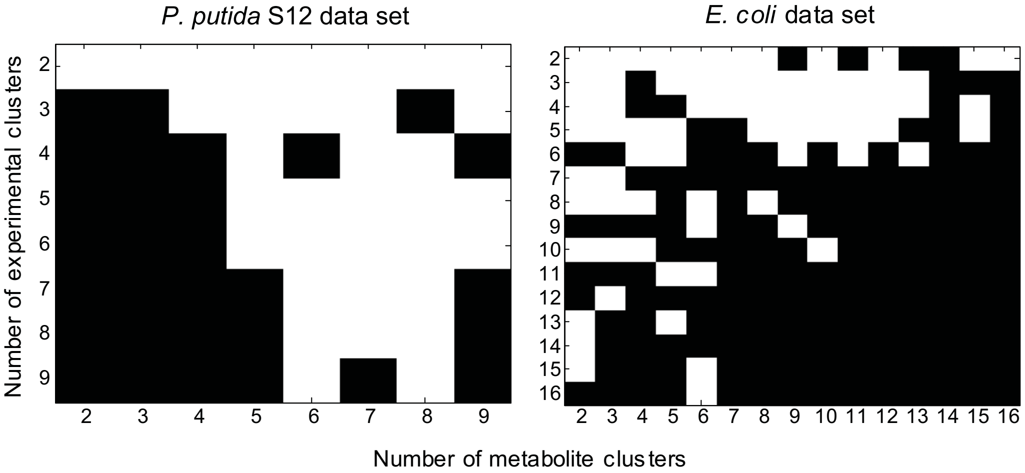


Figure 4 - Comparison of GA two-mode clustering and two-mode k-means clustering results. P. putida (left) and E. coli (right). The black area shows when GA two-mode clustering gave better results in terms of the evaluation criterion for a certain metabolite/experiment cluster combination. The white area shows when two-mode k-means gave the best results.

Figure 5 (page 75) - Two-mode clustering results for the metabolome data set of P. putida S12 grown on four different carbon sources. The roman numerals I to V, and F, G, N and S, are used to refer to the corresponding clusters throughout the text. The black/white bar indicates the number of cluster swaps a certain metabolite has made during the loo validation. Some metabolites were analyzed by GC-MS and LC-MS but their identity is not known, or were only identified as part of a class of metabolites, e.g. disaccharides. These metabolites were given a number behind the metabolite name to be able to distinguish between them during validation. For some metabolites there is uncertainty about the identification. These metabolites were given a question mark.

Figure 8 (page 76) - Two-mode clustering results for the metabolome data set of E. coli. The numerals I–VI, and characters A–G are used to refer to the corresponding clusters throughout the text. The black/white bar indicates the number of cluster swaps a certain metabolite has made during the loo validation. Some metabolites were analyzed by GC-MS and LC-MS but their identity is not known, or were only identified as part of a class of metabolites, e.g. disaccharides. These metabolites were given a number behind the metabolite name to be able to distinguish between them during validation. For some metabolites there is uncertainty about the identification. These metabolites were given a question mark.

Figure 5
see p. 74

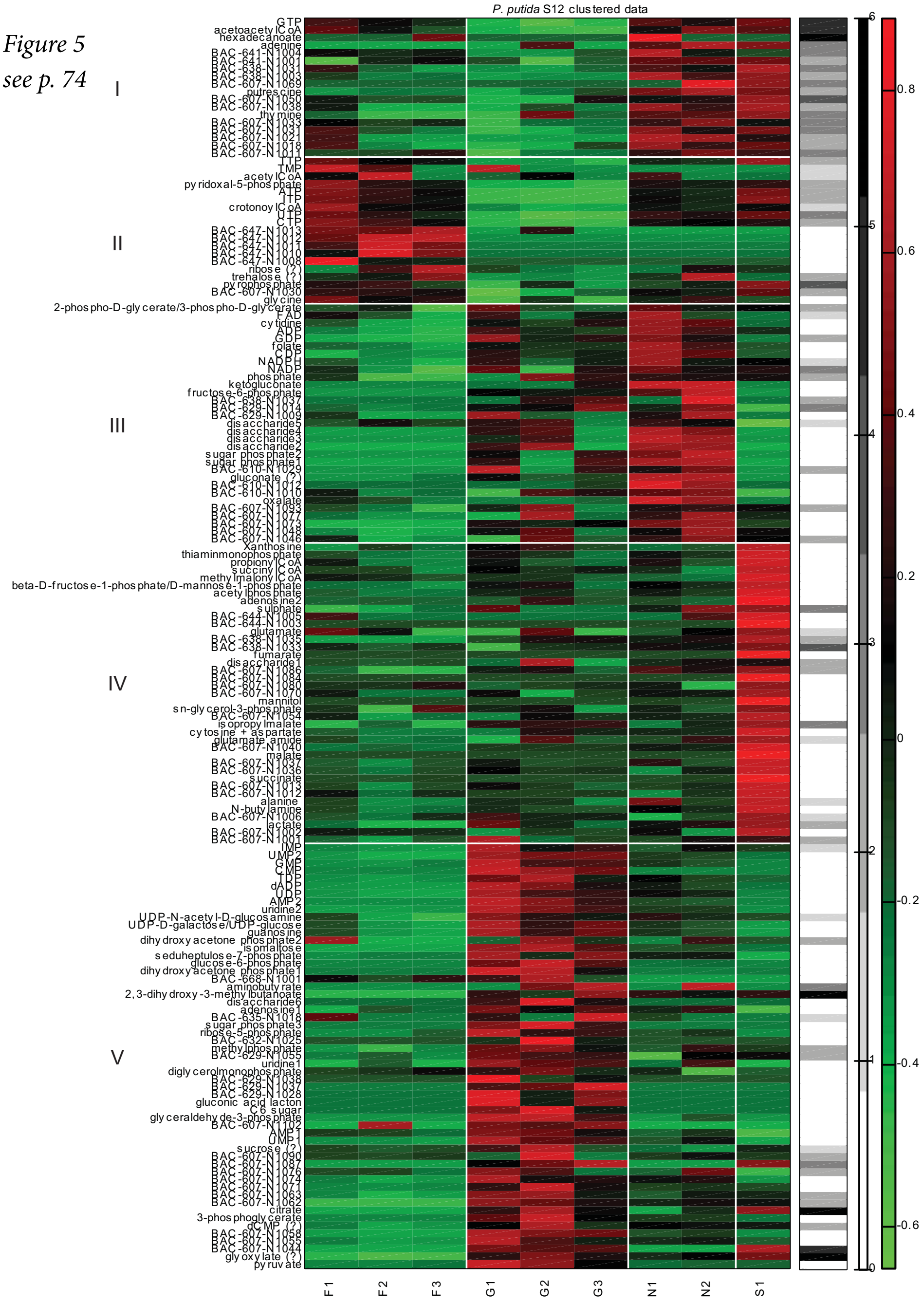
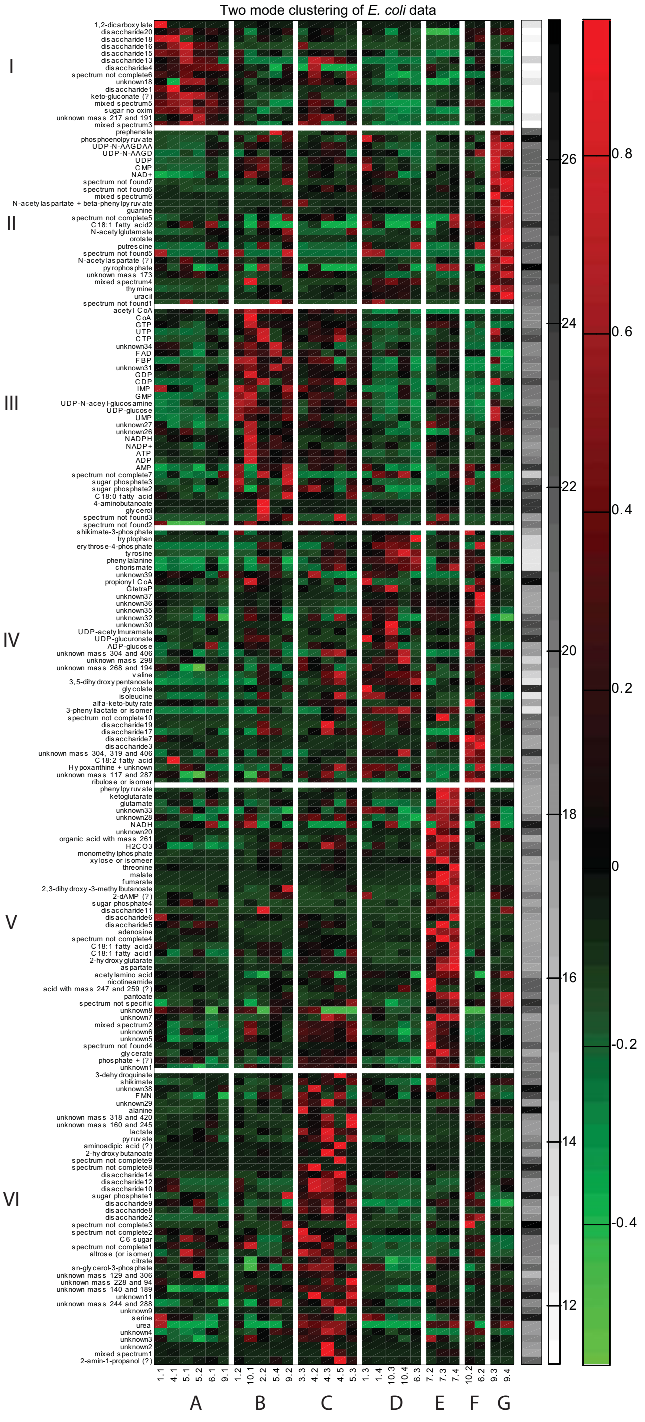


Figure 8
see p. 74



3.2 Two-mode clustering

3.2.1 *P. putida* data

The two-mode cluster result is presented in Figure 5. The different patterns of the metabolites under the different growth conditions are clearly visible. For example, the behavior of the metabolites in D-fructose and D-glucose grown cells was most different for the metabolites in clusters II and V. The stability of the clustered metabolites was tested with a leave-one-out validation strategy (see Section 1.4). The gray scale shows how often metabolites switch to another cluster during LOO validation. Only a few metabolites switch often, so the results are stable.

The visualization of the two-mode clustering result allows for the instant detection of outliers, as the color of an outlying variable is different from the consensus color of a cluster. In cluster FV, for instance, BAC-607-N1102 in experiment F2 is bright red, while most of the cluster is green, just as the results for F1 and F3 (Figure 5). This indicates that BAC-607-N1102 is a deviating point in the result of F2.

It is important to know whether the estimated cluster average is a suitable estimate of a cluster. By calculating the variance of a cluster, a measure for the homogeneity of the cluster is obtained. The variances for the two-mode clustering are presented in Figure 6. Most of the variances are comparable. FIII is the most homogeneous clusters found, while SII and GIII

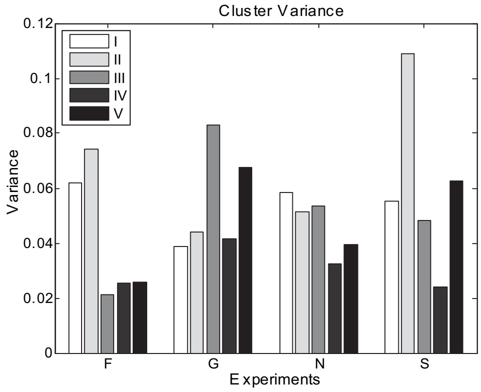


Figure 6 - Variances of the clusters in Figure 5

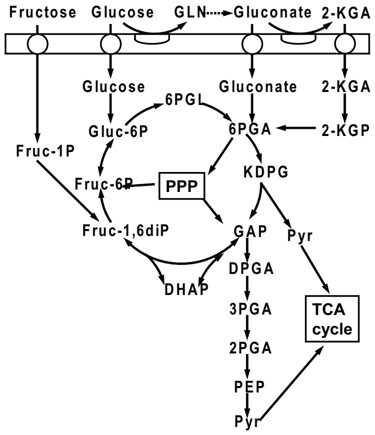


Figure 7 - Degradation of fructose, glucose and gluconate by the cyclic Entner-Doudoroff pathway in Pseudomonas. Taken with permission from SGM Microbiology [10].

contain the most variance. Analysis of the cluster variance can thus be applied as a quick assessment of the capability of the cluster to summarize the containing data.

When the resulting clusters are studied in more detail, several clusters contain interesting information. For instance, cluster V contains dihydroxyacetonephosphate1 (DHAP), pyruvate, glucose-6-phosphate (G6P), 3-phosphoglycerate (3PGA), glyceraldehyde-3-phosphate (GAP) and gluconic-acid-lactone (GLN). These metabolites are catabolic intermediates of the degradation pathway of D-fructose, gluconate and D-glucose (Figure 7).

On the other hand, fructose-6-phosphate (F6P) is member of cluster III, even though it is also an intermediate of the catabolic pathway of D-fructose, gluconate, and D-glucose. F6P connects the degradation pathways of D-fructose, gluconate, and D-glucose with the pentose phosphate pathway (PPP) (Figure 7). It is possible that the switch between the PPP and the degradation pathway explains why F6P was assigned a different cluster. The lack of 6-phosphofructokinase in *Pseudomonas* [23] probably contributes to this behavior as well. This result shows that two-mode clustering can find clusters that are informative from a biological point of view.

3.2.2 *E. coli* data

The two-mode clustering results of the *E. coli* data is shown in Figure 8. This data set is more complicated than the *P. putida* data because more perturbations were performed and longitudinal measurements were analyzed. The leave-one-out results are again shown as a gray scale bar (see Figure 8). The complexity of the data set is reflected in these results since the clustering is less stable compared to the *P. putida* data. Yet, biological meaningful results were obtained with respect to both the clustering of the metabolites and the samples. For instance, most nucleotides cluster together (cluster III) and the ketoglutarate/glutamate and malate/fumarate/aspartate pairs that are converted into each other by one enzymatic reaction, cluster together. On the other hand, with the clustering of the samples it was observed that the samples taken at the earlier time points cluster together, but also the samples of the wild-type strain, and samples collected from fermentations using succinate as the carbon source, cluster together.

4 Concluding remarks

Genetic algorithm based two-mode clustering is a valuable tool for the identification of biologically meaningful clusters in metabolomics data. Furthermore, it visualizes which subset of metabolites responds to which experimental condition. The results are validated by the use of a leave-one-out validation scheme that allows for the identification of metabolites that have an unstable clustering. A second validation measure is the analysis of the cluster variance. This gives insight in the homogeneity of the clusters and thus how well the clusters fit the data. Application of the newly developed approach to metabolomics data results in the identification of biologically relevant clusters.

The algorithm compares favorably to other approaches (e.g. two-mode k-means and single one-mode clustering). Hence, the genetic algorithm based two-mode clustering, together with an extensive validation of the results, is a valuable addition to the omics data analysis toolbox, as it provides a detailed overview of the data.

5 Acknowledgements

The authors would like to thank Richard Bas and Leon Coulier for analyzing the samples by LC-MS. Joost van Rosmalen is kindly thanked for sharing his implementation of the two-mode k-means algorithm. This research was funded by the Kluyver Centre for Genomics of Industrial Fermentation, which is supported by the Netherlands Genomics Initiative (NROG) and the Netherlands Bioinformatics Consortium (NBIC) and this work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI)."

6 References

1. Fiehn O: **Metabolomics - the link between genotypes and phenotypes.** *Plant Mol Biol* 2002, **48:**151-171.
2. Jolliffe IT: *Principal Component Analysis.* New York: Springer-Verlag; 2002.
3. Hoogerbrugge R, Willig SJ, Kistemaker PG: **Discriminant Analysis by Double Stage Principal Component Analysis.** *Anal Chem* 1983, **55:**1710-1712.
4. Vandeginste BGM, Massart DL, Buydens LMC, Jong Sd, Lewi PJ, Smeyers-Verbeke J: *Handbook of chemometrics.* Amsterdam: Elsevier; 1998.
5. Kaufman L, Rousseeuw PJ: *Finding groups in data: an introduction to cluster analysis.* Wiley-Interscience; 1990.
6. Van Mechelen I, Bock H-H, De Boeck P: **Two-mode clustering methods: a structured overview.** *Stat Methods Med Res* 2004, **13:**363-394.
7. Madeira SC, Oliveira AL: **Bicluster Algorithms for Biological Data Analysis: A Survey.** *IEEE Trans Comput Biol Bioinform* 2004, **1:**24-45.
8. Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E: **A systematic comparison and evaluation of biclustering methods for gene expression data.** *Bioinformatics* 2006, **22:**1122-1129.
9. Hartmans S, van der Werf MJ, de Bont JAM: **Bacterial degradation of styrene involving a novel flavin adenine dinucleotide-dependent styrene monooxygenase.** *Appl Environ Microbiol* 1990, **56:**1347-1351.

10. van der Werf MJ, Pieterse B, van Luijk N, Schuren F, van der Werff-van der Vat B, Overkamp K, Jellema RH: **Multivariate analysis of microarray data by principal component discriminant analysis: prioritizing relevant transcripts linked to the degradation of different carbohydrates in *Pseudomonas putida* S12.** *Microbiology* 2006, **152:**257-272.
11. van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ: **Centering, scaling, and transformations: improving the biological information content of metabolomics data.** *BMC Genomics* 2006, **7**.
12. Koek M, Muilwijk B, van der Werf MJ, Hankemeier T: **Microbial metabolomics with gas chromatography mass spectrometry.** *Anal Chem* 2006, **78:**1272-1281.
13. Coulier L, Bas R, Jespersen S, Verheij E, vanderWerf MJ, Hankemeier T: **Simultaneous Quantitative Analysis of Metabolites Using Ion-Pair Liquid Chromatography-Electrospray Ionization Mass Spectrometry.** *Anal Chem* 2006, **78:**6573-6582.
14. Smilde AK, van der Werf MJ, Bijlsma S, van der Werff-van der Vat B, Jellema RH: **Fusion of mass-spectrometry-based metabolomics data.** *Anal Chem* 2005, **77:**6729-6736.
15. Wehrens R, Buydens LMC: **Evolutionary optimisation: a tutorial.** *Trends Anal Chem* 1998, **17:**193-203.
16. The Mathworks Inc.. **Matlab 7.3 (R2006b).** 2006.
17. The Mathworks Inc.. **Genetic Algorithm Direct Search Toolbox 2.0.** 2005.
18. Raftery AE: **Choosing Models for Cross-Classifications.** *Am Sociol Rev* 1986, **51:**145-146.
19. Tibshirani R, Walther G, Hastie T: **Estimating the number of clusters in a data set via the gap statistic.** *J R Statist Soc B* 2001, **63:**411-423.
20. Salvador S, Chan P: **Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms.** In *Proceedings of the 16th IEEE International Conference on Tools with Arificial Intelligence (ICTAI 2004)*. 2004:576-584.
21. Vichi M: **Double k-means Clustering for Simultaneous Classification of Objects and Variables.** In *Advances in Classification and Data Analysis*. Edited by Edited by Borra S, Rocci R, Vichi M, Schader M. Heidelberg: Springer; 2001:43-52.
22. Baier D, Gaul W, Schader M: **Two-Mode Overlapping Clustering With Applications to Simultaneous Benefit Segmentation and Market Structuring.** In *Classification and knowledge organization*. Edited by Edited by Klar R, Opitz O. Heidelberg: Springer; 1997.
23. Lessie TG, Phibbs PVJ: **Alternative pathways of carbohydrate utilization in Pseudomonads.** *Annu Rev Microbiol* 1984, **38:**359-387.

7 Appendix

Appendix A – Proof that the evaluation function and the pooled variance are identical up to a scaling factor.

X = Data matrix

Variance: $s_k^2 = \dfrac{\sum_{n_k=1}^{N_k}(x_{n_k} - \bar{y}_k)^2}{N_k - 1}$

Pooled Variance: $s_{pooled}^2 = \dfrac{\sum_{k=1}^{K}(N_k - 1)s_k^2}{\sum_{k=1}^{K}(N_k - 1)}$

Evaluation function **E**=**X**-**UYV**$^{\mathrm{T}}$ can also be written as: $SS_{res} = \sum_{k=1}^{K}\sum_{n_k=1}^{N_K}(x_{n_k} - \bar{y}_k)^2$

$$s_{pooled}^2 = \frac{\sum_{k=1}^{K}(N_k - 1)\left(\dfrac{\sum_{n_k=1}^{N_k}(x_{n_k} - \bar{y}_k)^2}{(N_k - 1)}\right)}{\sum_{k=1}^{K}(N_k - 1)} = \frac{\sum_{k=1}^{K}\sum_{n_k=1}^{N_K}(x_{n_k} - \bar{y}_k)^2}{\sum_{k=1}^{K}(N_k - 1)}$$

$$s_{pooled}^2 = \frac{SS_{res}}{\sum_{k=1}^{K}N_k - \sum_{k=1}^{K}1} = \frac{SS_{res}}{(P \cdot Q) - K}$$

6 Discovery of functional modules in metabolomics data: regulation of cellular metabolite concentrations

Robert A. van den Berg, Age K. Smilde, Jos A. Hageman, Uwe Thissen, Johan A. Westerhuis, and Mariët J. van der Werf

Summary

In metabolism, functional modules can be defined as groups of metabolites that have a related function. Functional modules can be determined on different levels within the cellular organization. In response to normal, not stressful conditions, global regulatory effects will control the major physiological processes. These global regulatory effects are characterized by metabolites whose concentrations show a similar behavior in response to varying environmental conditions. Changes in environmental conditions that perturb specific areas in the metabolism will provoke local regulatory effects. Metabolites whose concentration responds similar to such local perturbations will be part of the same local functional module. In this paper we identified both local and global functional modules based on two real-life microbial metabolomics data sets using a top-down systems biology approach. Furthermore we discuss the nature of homeostasis, as is reflected by the regulation of metabolite concentrations.

Local functional modules were identified in microbial metabolomics data sets originating from *Escherichia coli* and *Pseudomonas putida* S12 by a two-mode clustering approach. Their identification proved strongly dependent on the variation in environmental conditions under which the metabolome data were obtained. For instance, a local functional module containing citric acid cycle and redox-related metabolites was identified when *E. coli* was grown on succinate instead of D-glucose. The global functional modules were discovered by a correlation network analysis. Here, modules related to amino acid biosynthesis and the central metabolism were found. Comparison of the metabolite composition of local and global functional modules revealed that metabolites which are member of the same global functional module are not necessarily member of the same local functional module, and vice versa.

Regulation of metabolite concentrations was found to occur on different hierarchical levels. Whether these different hierarchical regulation levels could be identified in the metabolomics data set depended strongly on the environmental conditions – and thus the experimental design behind the data sets - and how these conditions perturb the metabolism. By the application of two different data analysis methods both local and global functional modules could be identified.

1 Background

Functional modules are components of a system that have a specific function. For example, a computer consists of a mother board, hard disk, monitor, and video card. These different components all have a specific functionality and are therefore considered as functional modules. Functional modules depend on the level of detail in which a system is studied. The video card of the computer, for instance, consists of different functional modules as well, such as, the connector to the mother board, cooler, memory, and processor. In cell biology, functional modules in the broadest sense relate to specific metabolic processes like carbon metabolism, stress response, or redox/energy balance. These broad definitions include all layers of cellular organization; genes, proteins, and metabolites [1]. Functional modules can be subdivided down to the level of a single metabolic pathway, or a part of a pathway, or a single operon. Generally, functional modules are considered on distinct biochemical layers within the cellular organization, for instance, gene regulons [2], protein interaction networks [3], or the energy/redox metabolite pools [4]. As functional modules relate to biological function, their behavior and regulation can provide information on the physiological state of an organism and thus reflect its response to environmental conditions.

To grow and to maintain themselves, cells attempt to regulate their important processes within certain boundaries. Even after (extreme) changes in environmental conditions, the cells try to maintain these boundaries. This regulatory phenomenon is called homeostasis.

For metabolomics data, homeostasis refers to cellular metabolite concentrations. The concentrations of metabolites in a cell are influenced by different factors: (i) enzyme activity, which is regulated via several mechanisms, such as enzyme concentration and allosteric regulation; (ii) the concentrations of metabolites connected in the metabolic network as these determine the thermodynamic push or pull in a certain direction for the biochemical reactions; (iii) the possibilities of the organism to control the influx and efflux of metabolites, for instance, it can be very difficult to control the influx of organic acids, as organic acids in undissociated form can travel freely over the cell membrane [5,6]. This factor is also related to (ii).

Cellular regulation of homeostasis is often hierarchically organized; global regulatory effects coordinate central metabolism in an organism cultivated under normal, not stressful conditions (base-level homeostasis), while increasingly smaller-scale sub-processes are controlled to respond to specific changes in environmental conditions (Figure 1). In other words, global regulation mechanisms are dominant when the organism is cultivated under

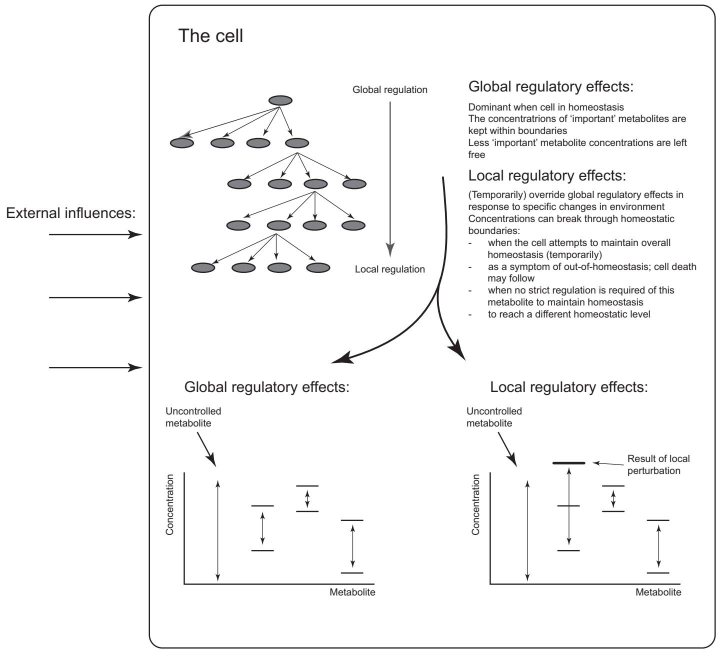

The cell
External influences:
Global regulation
Local regulation
Global regulatory effects:
Dominant when cell in homeostasis
The concentratrions of 'important' metabolites are kept within boundaries
Less 'important' metabolite concentrations are left free
Local regulatory effects:
(Temporarily) override global regulatory effects in response to specific changes in environment
Concentrations can break through homeostatic boundaries:
- when the cell attempts to maintain overall homeostasis (temporarily)
- as a symptom of out-of-homeostasis; cell death may follow
- when no strict regulation is required of this metabolite to maintain homeostasis
- to reach a different homeostatic level
Global regulatory effects:
Uncontrolled metabolite
Concentration
Metabolite
Local regulatory effects:
Uncontrolled metabolite
Result of local perturbation
Concentration
Metabolite

Figure 1 - Schematic overview of the regulation of homeostasis in the cell. Global regulatory effects control 'important' processes under normal, not stressful conditions. Local regulatory effects become visible due to specific changes in the environmental conditions.

conditions that generally are not stressful.

In contrast, local regulatory mechanisms become dominant when the processes they control become perturbed beyond the boundaries of this base-level homeostasis. Such a perturbation could result in a new homeostatic state, or in the inability of the cells to grow and maintain themselves and thus in loss of homeostasis. In the example of the organic acids, the production of these compounds by fermentative micro-organisms results in the accumulation of toxic concentrations of these organic acids [5,6] disrupting the membrane potential and thus cellular homeostasis. Pieterse and co-workers [6] discovered that *Lactobacillus plantarum* redirects its metabolism towards other fermentation end products in response to this disruption. This can be considered a local response to the specific perturbation of the cell.

The layout of cellular regulation implies that at the extremes two types of functional modules can be discerned: (i) functional modules on the level of global regulation, and (ii) functional modules on the level of local organization. Identifying functional modules between these two extremes in metabolomics data sets will require different strategies. Local regulatory effects will manifest themselves in response to specific perturbations of that area in the metabolic network that they control, while the global regulatory effects will be reflected by metabolites whose concentration behaves similar under a wide range of environmental conditions.

Until recently, the study of functional modules in experimental metabolite data was limited by the poor detection of large numbers of metabolites. Therefore, the study of functional modules in metabolism was limited to *in silico* analysis based on the topology of the metabolic network [7-9]. These are bottom up systems biology approaches in which the study of the functional modules is based on the modeling of existing knowledge about the metabolism [10]. These models model flux distributions under steady state assumptions, and they are often not able to fully capture concentration-based regulation mechanisms [11]. Here we describe a top down systems biology approach to identify functional modules. Instead of analyzing flux distributions, we analyze metabolite concentrations. This is now possible due to the advancements in the field of metabolomics [12,13] that allows the measurement of the (relative) concentrations of several hundreds of metabolites [13]. In this paper we identify both global and local functional modules in two different microbial metabolomics data sets.

2 Results

For the purpose of this study, two different metabolomics data sets were used. The first data set was a metabolomics data set in which *P. putida* S12 was grown on four different carbon sources as the sole carbon source [14]. The second data set consisted of metabolomes of *E. coli* grown under 28 distinct conditions with regard to strain, environmental conditions, and time point of harvesting [15].

2.1 Homeostasis of metabolites

Homeostasis refers to the desire of a cell to maintain important processes within certain boundaries in order to grow and maintain itself. In order to study whether homeostasis could be identified in a microbial metabolome data set, we determined the behavior of the relative concentration of the average metabolite in the *E. coli* data set (Figure 2). To this end, the concentrations of all the individual metabolites were converted into relative concentrations with respect to the maximal concentration for that metabolite in on of the samples. Subsequently, the relative concentrations of every metabolite were sorted (ranked) from low to high. Next, the concentration profile of an average concentration was established by averaging the relative concentrations of all the metabolites per rank.

There was a large difference between the average relative metabolite concentration and the maximal relative concentration achieved for a metabolite in the *E. coli* data set (Figure

2). The average relative concentration, estimated by the median as a robust estimator, was 13.8% of the total relative range (100%). This observation is an indication that some of the 28 distinct environmental conditions studied in this experiment perturbed the homeostasis of the average metabolite. Furthermore, this perturbation seemed to have induced local regulatory phenomena. A more linear behavior would have suggested that global regulation is dominant, as the metabolite concentrations are left free or are maintained within certain boundaries (Figure 1).

Studying the metabolite concentration profiles of individual metabolites could provide further insight in the nature of homeostasis and of (specific) regulatory processes controlling its concentration. Therefore, the ordered relative concentration profiles for all metabolites in the *E. coli* data set were plotted and visually inspected. En large, we could discriminate six different concentration profiles, and possibly regulation patterns. Figure 3A

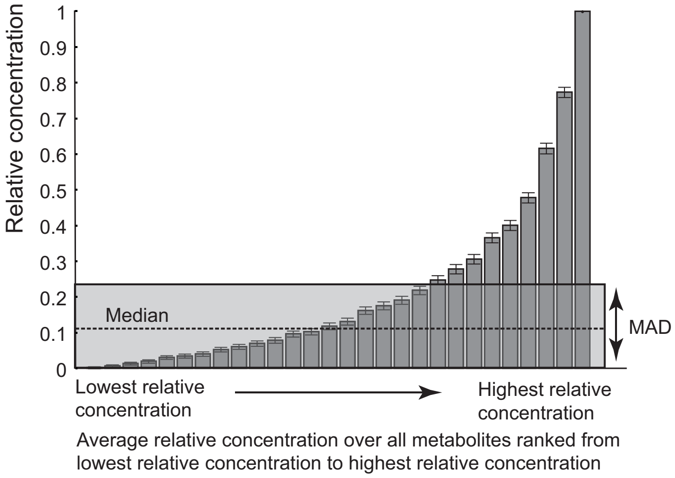


Figure 2 - Relative concentration profile of an average metabolite. The metabolite concentrations of the E. coli data set were per metabolite sorted from low concentration to high concentration. Furthermore, the concentrations were per metabolite normalized relative to the highest concentration reached for that metabolite. The sorted and normalized metabolite concentrations were averaged per sort position using the mean as a robust estimator. This means that over all metabolites the lowest relative concentration was averaged. Next the second lowest concentration was averaged. This was continued until the highest relative concentration was averaged. This resulted in the relative concentration profile of an average metabolite in the E. coli data set sorted in ascending order. The shaded area is the median +/- the median absolute deviation (MAD). The shaded area is a robust indication for the dispersion around the median relative concentration in the 28 experiments in the E. coli data set. The error bars represent the standard error (n=188).

shows the pattern of a metabolite that matched the average pattern (Figure 2); the average concentration was low compared to the maximally attained concentration. This could indicate that the last four experimental conditions somehow strongly perturbed the concentration of phenylpyruvate. The profiles of the metabolites in Figure 3B and 3C did not show extreme relative concentrations, that is, the concentrations were evenly distributed between the maximal and the minimal concentration. This could point to at least two possible situations: (i) the metabolites were controlled within the 'normal' boundaries of homeostasis; (ii) the concentrations of these metabolites were not regularly controlled. The profile in Figure 3C described a smaller relative concentration range than Figure 3B, 30-100% for Figure 3C and 5-100% for Figure 3B respectively. This could be an indication that the metabolite in profile B was not under strong regulation, while FMN, the metabolite in Figure 3C was regulated at a base level. Figure 3D is an example of a metabolite, aspartate, showing extreme differences between a base level homeostasis and deviations from it. In *E. coli*, the metabolites fumarate and malate can be interconverted from aspartate via simple enzymatic conversions by generally highly active enzymes. We therefore expected that fumarate and malate would show a similar concentration profile, which was indeed the case (Supplemental data I). For all three metabolites the extreme concentrations were observed when the cells were cultivated on succinate. Figure 3E displays the pattern of a metabolite whose concentration is below the detection limit under half of the environmental conditions and is detectable in the other half.

There are three possible explanations for this behavior: (i) either the flux through the enzyme that converts this metabolite is very high; (ii) the metabolite was not produced under those experimental conditions; or (iii) a concentration profile similar to the other profiles is present, although not detectable with the used equipment. The last profile (Figure 3F) demonstrates the pattern of a metabolite that suggests that there can be different regulation levels or homeostatic states at which the concentration of a metabolite can be. There is a low concentration state, a mid range concentration state, and a very high concentration state. These metabolite profiles indicate that different local and global regulatory mechanisms are indeed present in the *E. coli* data set.

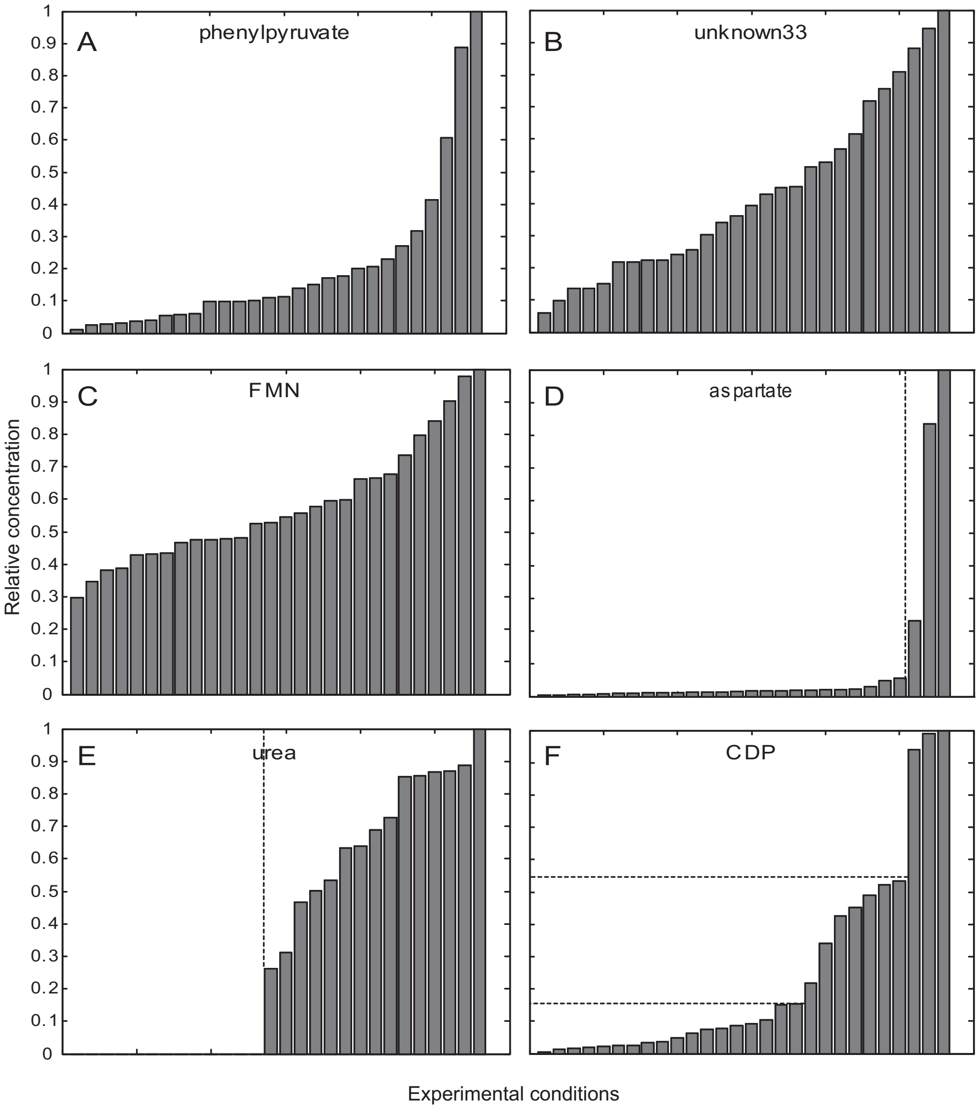


Figure 3 - Profiles of the distribution of the relative metabolite concentration of selected metabolites. All profiles are relative to the highest concentration in the data set of that metabolite. The concentrations are sorted in ascending order. The dashed lines in 3D, E and F indicate possible local perturbations.

2.2 Identification of functional modules

2.2.1 Global functional modules

Global functional modules are based on the similarity in behavior of a group of metabolites under the full range of experiments in the data set (Figure 4). Therefore, as a measure of similarity, the pair wise correlation between metabolite concentrations was calculated [16]. Correlation analysis as a means to identify regulatory effects in metabolomics is also discussed in a recent paper by Müller-Linow and co-workers [17]. Calculating correlations on a small number of measurements will likely result in unreliable results due to the amount of false positives that can be expected. Since the *P. putida* data set consisted of metabolomes obtained from three biological replicates of four different environmental conditions, we refrained from correlation analysis with the *P. putida* S12 data set.

The correlation networks of the *E. coli* data set were studied for different cut-off values for the correlation coefficient (Figure 5, additional results not shown). Three global functional modules were present at all cut-off values and grew in size when a lower correlation coefficient was selected as cut-off (Figure 5). These global functional modules

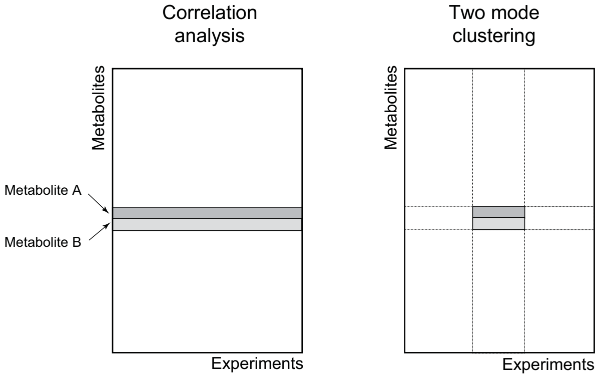


Figure 4 - Principle behind correlation analysis and two-mode clustering for the identification of global and local functional modules. In a correlation analysis, the behavior of the concentrations of metabolite A and B is compared over all environmental conditions. The highest correlation coefficient is obtained when the concentrations of the compared metabolites behaves the same under all experimental conditions. In contrast, two-mode clustering clusters metabolites that behave most similar within subsets of the environmental conditions [23]. The two-mode clustering method simultaneously partitions the normalized [25] metabolite concentration data by searching for those partitions of metabolite concentrations and experimental conditions that result in the most homogeneous partitions.

were based on metabolites related to (i) Nucleotides, energy and cell wall intermediates, (ii) amino acids, the pentose phosphate pathway, nucleotide bases, and (iii) a module that is strongly interconnected but consisted initially of unknowns. At rho = 0.725 (Figure 5B) citrate, fumarate, malate, glycerate, 3-dehydroquinate, 2-hydroxyglutarate, and FMN have joined this module as well.

The large global functional modules (more than six members) in Figure 5B contain metabolites that are close to each other in the metabolic network as well as metabolites that are further way. For instance, the amino acid/PPP global functional module contains many intermediates from the aromatic amino acid biosynthesis pathway, e.g. erythrose-4P, chorismate, phenylpyruvate, tyrosine, tryptophan, and phenylalanine. However, metabolites that are further away in the metabolic network, e.g. the amino acids valine and isoleucine; the nucleotide bases thymine, guanine, and uracil; are part of this module as well. This global functional module could be the result of regulation of amino acid concentrations and regulation of the distribution of PPP intermediates towards (i) C5-sugars (building blocks for nucleotides) and (ii) the aromatic amino acid biosynthesis pathway (erythrose-4P).

Another example of a global functional module, with closely related as well as more distinct metabolites, is the nucleotides/energy/cell wall global functional module. In this module, many of the nucleotides are present, but also acetyl CoA, fructose-1,6-bisphosphate (FBP), and phosphoenolpyruvate. Moreover, notable missing nucleotides are GTP and ATP, that both cluster with CoA (Figure B, red module labeled “Energy transfer”); and AMP that clusters with a disaccharide.

The finding that the identified global functional modules did not comply to the distance in the metabolic network is in line with the work of Notebaart and co-workers [18], who showed for flux analysis that closeness in the metabolic network is not a good indicator for flux predictions; and with the work of Steuer and co-workers [19].

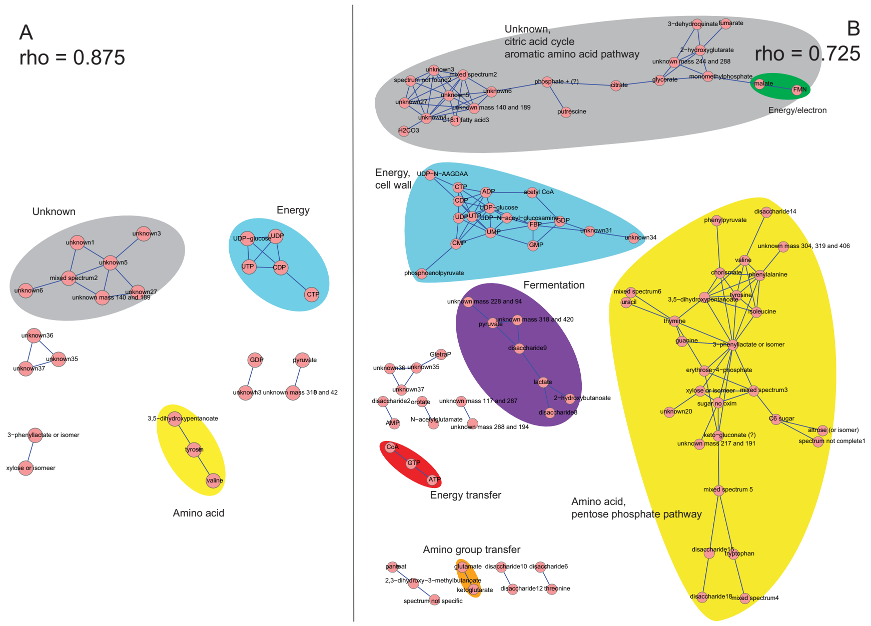


Figure 5 - Networks of significantly correlating metabolites in the E. coli metabolome data.(A) Modules at a correlation coefficient of 0.875, (B) Modules at a correlation coefficient of 0.725. Modules with the same color indicate modules that were small at rho = 0.875 and were extended further at rho = 0.725. If necessary, the description of the module was extended. The metabolites on top of the green oval were not coupled to the grey module at the previous correlation coefficient cut-off value (Results not shown).

Figure 6 (page 93) - Local functional modules in P. putida S12 data set using two-mode clustering. The characters F, G, N, and S refer to respectively D-fructose, D-glucose, gluconate, and succinate, the carbon source on which P. putida S12 was grown. The colors refer to the metabolite concentrations relative to their biological range. The BAC codes refer to unidentified metabolites, and the (?) refers to metabolites whose identification is uncertain.

Figure 7 (page 94) - Local functional modules in E. coli data set using two-mode clustering. The numbers below the x axis refer to the experimental conditions. The colors refer to the metabolite concentrations relative to their biological range. The (?) refers to metabolites whose identification is uncertain.

Figure 6
see p. 92

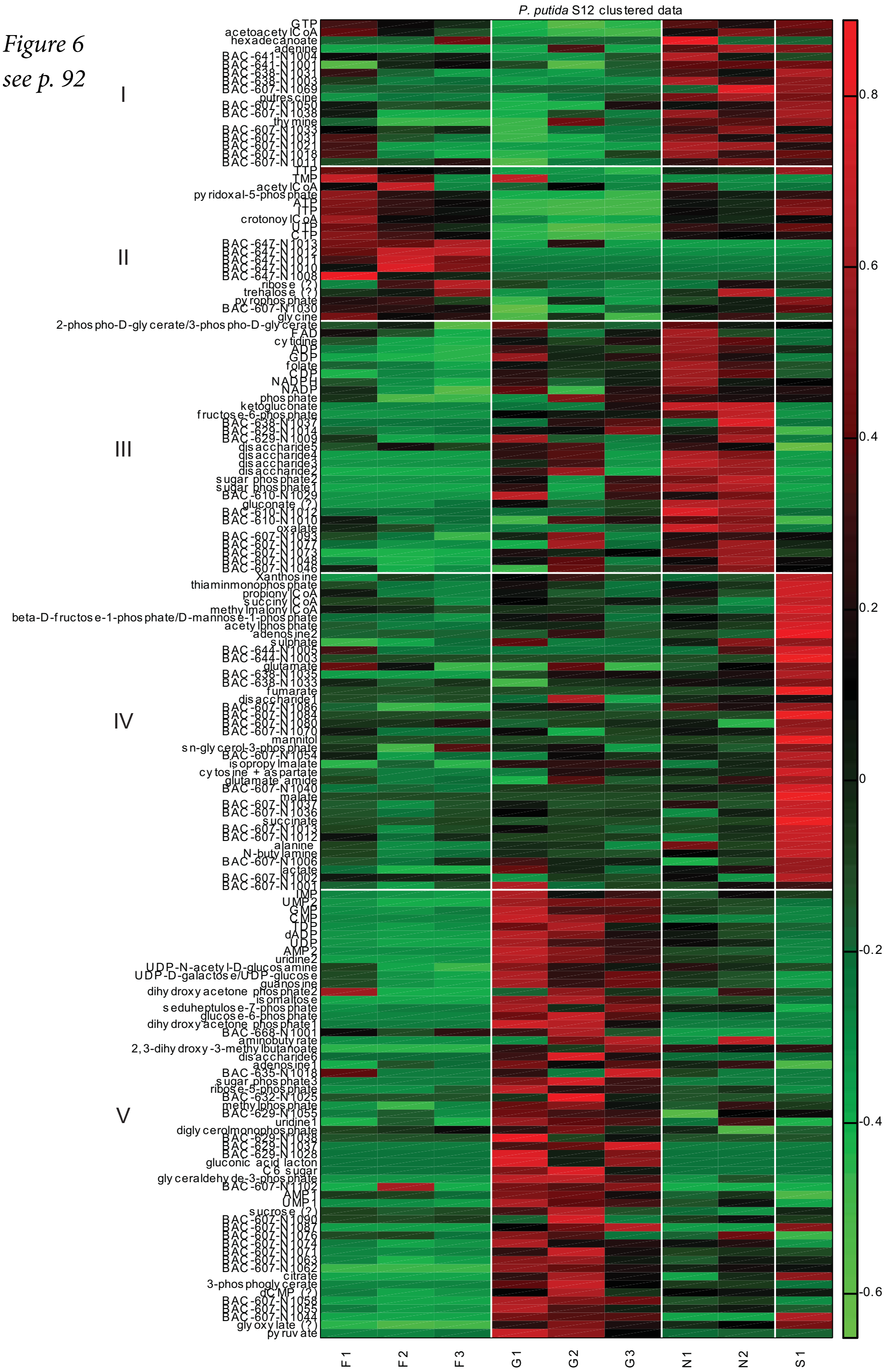
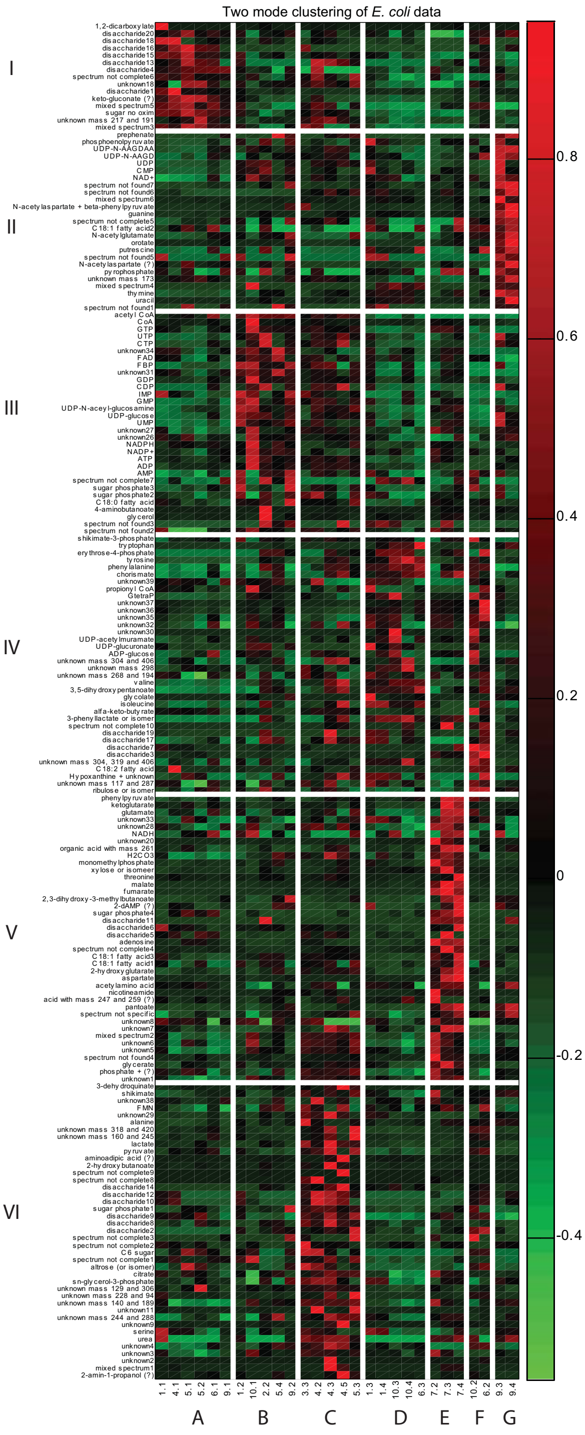


Figure 7
see p. 92



Global functional module	Metabolites involved (total metabolites in module)
Nucleotides, energy, cell wall	10 (17)
Unknown, aromatic amino acids, citric acid cycle	3 (21)
Energy transfer	3 (3)
Amino acids, PPP	2 (29)
Amino group transfer	1 (2)
Fermentation	1 (7)

Table 1 - Global functional module membership for metabolites involved in the core reactions of E. coli. From the global functional modules identified at rho = 0.725, 20 metabolites were involved in the core reactions of E. coli.

The global functional modules in Figure 5B were compared with the metabolic core of *E. coli* as identified by flux-balance analysis [20]. According to Almaas and co-workers [20], the fluxes of the metabolic core reactions were highly correlated under the 30 000 tested simulation environments. In 44 of the 90 reactions of the metabolic core, 20 metabolites (of a total of 96) from Figure 5B were involved in one or more reactions. ATP was involved in 22 of these 44 reactions, whereas ADP was involved 11 times. The measured concentration profiles of the metabolites involved in the core reactions of *E. coli* are not as strongly correlated as the fluxes [20], since the metabolites are divided over six global functional modules (Table 1), instead of one functional module. However, it is remarkable that from the 20 metabolites involved in the core reactions, 10 metabolites were member of the nucleotides/energy/cell wall global functional module (Table 1).

2.2.2 Local functional modules

Local regulatory mechanisms become active when certain processes in the central metabolisms are (temporarily) perturbed in order to (i) maintain homeostasis by reaching another local homeostatic level; (ii) to be able to return to homeostasis; (iii) or as a sign that the cell is out of homeostasis (Figure 1). Local functional modules are groups of metabolites that behave similar in the conditions that overrule the global regulatory effects, or that behave similarly in that subset of the experimental conditions that only perturb specific processes of the metabolism. In order to identify local functional modules in real life metabolomics data sets, the data analysis tool two-mode clustering was applied. Two-mode clustering searches for groups of metabolites that behave the same for subsets of experimental conditions (Figure 4) [21,22].

In the *P. putida* S12 data, five local functional modules were identified after two-mode clustering (Figure 6). For instance, a part of the catabolic pathway for D-fructose, gluconate and D-glucose was recovered in local functional module V [23]. The clustering of these metabolites seemed primarily determined by the behavior of the metabolites under growth on D-glucose and D-fructose. Here, the highest and lowest relative metabolite concentrations were detected. As the two-mode clustering method searches for clusters that are as homogenous as possible, clustering extreme values in data with this type of concentration profiles (Figure 2) correctly is more important than clustering average values correctly, since wrongly clustering extreme values will generally have a larger impact on cluster homogeneity than wrongly clustering average values. The behavior of these metabolites under growth on D-glucose and D-fructose was especially interesting as the preferred carbon sources for *P. putida* S12 are organic acids [24]. These results suggested that this functional module was specifically perturbed by these sub-optimal carbon sources, D-glucose and D-fructose, and that this functional module was locally regulated in order to maintain homeostasis.

For the *E. coli* data set six local functional modules were identified [23] (Figure 7). The number of local functional modules in the *E. coli* data set could not be determined with certainty as the large number of experimental conditions made it difficult to establish an optimal number of clusters, and thus local functional modules in this data set [23]. The clustering lead to functional modules which were biologically relevant, for instance, ketoglutarate, glutamate, malate, fumarate, aspartate, and NADH, all citric acid cycle and redox related metabolites, were part of the same local functional module (Figure 7, V). The local functional module seemed to be perturbed the most when *E. coli* was grown on succinate instead of D-glucose (Figure 7, E). This provided an explanation for the joint behavior of these metabolites as growth on succinate as sole carbon source requires gluconeogenesis, which results in a lower energy yield compared to growth on D-glucose. The lactate and pyruvate metabolite pair were part of another functional module (Figure 7, VI) that had extreme concentrations when *E. coli* was cultivated under oxygen limited conditions (Figure 7, C). As lactate is a major fermentation end product of *E. coli* produced from pyruvate under fermentative conditions, this is another example of a biologically meaningful local functional module in this data set. These results demonstrate that the behavior of these metabolites closely related to, or part of the citric acid cycle behave differently after specific perturbations.

2.2.3 Global versus local functional modules

Identifying functional modules on either the local or the global level means exploring different views on metabolism (Figure 1). Comparing these regulation levels will therefore reveal more information about the relation between them. The composition of the global functional modules (rho = 0.725) of *E. coli* (Figure 5) was compared with the identified local functional modules (Figure) and the comparison is presented in Table 2. Although the global functional modules analyzed often contain metabolites which are for a majority member of one local functional module, it is clear that there is no direct correspondence between the composition of the global and the local functional modules. The metabolites that are part of the same global functional modules are frequently divided over several local functional modules. Examples of this are the global functional modules “Unknown/citric acid cycle/aromatic amino acid pathway” and “Amino acid/PPP”, which contain metabolites that are member of respectively four and five local functional modules. In addition, the metabolites in local functional modules are spread over different global functional modules. For instance, metabolites in local functional module VI are divided over five different global functional modules, among which are the global functional modules “Unknown/citric acid cycle/aromatic amino acid pathway”, “Amino acid/PPP”, and “Fermentation”. Local functional module VI was perturbed the most under low oxygen conditions. The results presented in Table 2 illustrate the differences between global and local functional modules. It shows that metabolites that are part of the same global functional module can be part of different local functional modules due to different responses to local perturbations and vice versa.

3 Discussion

In this paper we have identified functional modules in two real life microbial metabolomics data sets. Using two different data analysis tools, we were able to identify functional modules based on global (Figure 5) and local (Figure 7,8) regulatory effects in two data sets of a very different nature.

These results provide insight in the mechanisms of homeostasis, where local perturbations of metabolism by specific environmental conditions lead to local responses in the metabolic network in order to maintain homeostasis. This is demonstrated by the average metabolite profile in the *E. coli* data set (Figure 2).

Moreover, differences in local behavior of the concentrations of metabolites that were member of the same global functional module were identified. While the metabolites were part of the same global functional module, they belonged to different local functional modules in response to specific changes in the environmental conditions (Figure 6 and 7).

The nature of the two data sets had a strong influence on the ability to identify local and global functional modules. The *P. putida* S12 experimental design consisted of only four different experimental conditions, while the *E. coli* experimental design had a broad range of different experimental conditions with regard to environmental conditions, strains, and time points at which the samples were taken. The number of different experimental conditions and the impact of the changes in environmental conditions (e.g. low oxygen versus normal oxygen levels, or D-glucose versus succinate as carbon source) made the *E. coli* data set better suited for the identification of local functional modules, as many different local perturbations were present. In contrast, the *P. putida* S12 data set is more suited for the discovery of global functional modules as the global regulatory mechanisms are perturbed only by the different carbon sources. However the small number of samples prohibited us to perform this analysis reliably.

Generally, we anticipate that the nature and type of functional modules that can be identified with the methods presented in this paper depend highly on the experimental design. This means that the functional modules discovered vary between data sets and are not static as parts of a jigsaw puzzle. For instance, different environmental conditions, such as carbon source, pH, nitrogen source, will induce different local perturbations. Furthermore, experimental designs based on different and strong perturbations will favor the discovery of local functional modules, while experimental designs with no or only mildly varying environmental conditions will favor the identification of global functional

Global functional module	Number of metabolites of global functional module found in local functional module
Unknown, citric acid cycle, aromatic amino acid pathway	Module II (1), Module III (2), Module V (12), Module VI (6), Total (21)
Nucleotides, energy, cell wall	Module II (4), Module III (13), Total (17)
Energy transfer	Module III (3), Total (3)
Amino group transfer	Module V (2), Total (2)
Amino acid, PPP	Module I (7), Module II (5), Module IV (10), Module V (3), Module VI (4), Total (29)
Fermentation	Module VI (7), Total (7)

Table 2 – Comparison module membership for global and local functional modules. The cluster composition of the global functional modules at rho = 0.725 was compared with the local functional modules. Between brackets, the number of metabolites from the global functional module present in the local functional module is given. Also the total number of metabolites in the global functional module is given.

modules. Therefore, to study global or local metabolic regulation, the experimental setup should be designed in such a way that these regulatory effects are present in the data.

Besides the influence of the experimental design on the discovery of global and local functional modules, the settings of the data analysis methods influence the results as well, as discussed in Chapter 5 for two-mode clustering [23]. Varying the cut-off value for the correlation coefficient will provide different views on the global networks from loosely related to strongly related functional modules. There is no straightforward answer to which settings are the right or best ones. The different settings will provide different views on the data; selecting more functional modules for two-mode clustering will allow a more refined view on local effects, limited by the resolution or information present in the data. Biological interpretation, combined with rigorous statistical validation will therefore provide the most relevant views.

To our knowledge, this is the first time that metabolic regulation on the level of global and local functional modules has been identified in real life microbial metabolomics data sets. This work could therefore provide a basis for future work on metabolic homeostasis and regulation of metabolite concentrations in micro organisms.

4 Methods

4.1 Data set

The *P. putida* S12 data set was generated by cultivating *P. putida* S12 in independent triplicate controlled fermentations on D-fructose, D-glucose, gluconate, and succinate [25]. Metabolome samples were taken, quenched, extracted, and analyzed by GC-MS [26] and LC-MS [27]. The resulting data was normalized and manually curated as described previously [25], metabolites that were not detected in more than 80% of the experiments were removed from the data set. The data set consisted of 162 metabolite measurements.

The *E. coli* data set was generated by cultivating *E. coli* NST74, a phenylalanine overproducing strain, in controlled batch fermentations in which one environmental parameter compared to a reference condition was varied [15]. One fermentation was performed with the wild type strain W3110, instead of the NST74 strain. Metabolome samples were taken at different time points during the fermentations, quenched, extracted, and analyzed by GC-MS and LC-MS. The resulting data was normalized and manually curated as described previously[23], metabolites that were not detected in more than 80% of the experiments were removed from the data set. The data set consisted of 188 metabolite measurements.

4.2 Global functional modules

The pairwise correlation between each metabolite pair was determined based on Spearman rank correlation, as this correlation measure is robust with regard to outliers and non-linear behavior [28]. Metabolite pairs with an absolute correlation coefficient (ρ) larger than 0.7 were considered relevant for interpretation (p-value for every correlation with $|\rho| >= 0.7$ and 28 observations (*E. coli* data set) is equal or smaller than $5.1 \cdot 10^{-5}$). The absolute correlation takes into account positive as well as negative correlations between metabolite pairs. Correlation coefficients lower than 0.7 were not further analyzed, as the correlation between two metabolites then explains less than roughly 49% ($\rho^2 = 0.49$) of the variation of the metabolite pair. Additional testing by permutation tests (1000 runs) confirmed their significance. The significant correlations were visualized for different cut-off values (Figure 5, additional results not shown) with the software program Cytoscape [29].

4.3 Local functional modules

For the discovery of the local functional modules a genetic algorithm based two-mode clustering method was applied [23]. The two-mode clustering method searched for clusters of experimental conditions and metabolites that were as homogeneous as possible. Before the two-mode clustering was applied the data was range scaled [25] in order to search for functional modules based on the behavior of metabolites relative to their biological range. The optimal number of clusters in the metabolite and experiment mode was based on the generalized knee method and the biological interpretation of the clustering results [23]. To decrease the chance of identifying a local minimum, five different start solutions were chosen and the best solution was used in further analysis.

4.4 Computations

The computations were performed on personal computers running Windows XP, Matlab [30], the statistics toolbox, the genetic algorithm and direct search toolbox, both from Mathworks, and home-made scripts.

5 Authors' contributions

RAB interpreted the results, contributed to conceptual discussion, and wrote the manuscript. AKS, JAW and MJW provided feedback on the interpretation of the results and contributed to conceptual discussion on functional modules. JAH developed the methodology to identify local functional modules and performed the calculations on local functional modules. UT developed the methodology for the identification of global functional modules and performed the calculations.

6 Acknowledgements

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

7 References

1. Oltvai ZN, Barabasi AL: **Systems Biology: Life's Complexity Pyramid.** *Science* 2002, **298:**763-764.
2. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nat Genet* 2003, **34:**166-176.
3. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP et al.: **Evidence for dynamically organized modularity in the yeast protein-protein interaction network.** *Nature* 2004, **430:**88-93.
4. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5:**101-113.
5. Axe DD, Bailey JE: **Transport of lactate and acetate through the energized cytoplasmic membrane of *Escherichia coli*.** *Biotechnol Bioeng* 1995, **47:**8-19.
6. Pieterse B, Leer RJ, Schuren FHJ, van der Werf MJ: **Unravelling the multiple effects of lactic acid stress on Lactobacillus plantarum by transcription profiling.** *Microbiology* 2005, **151:**3881-3894.
7. Papin JA, Reed JL, Palsson BO: **Hierarchical thinking in network biology: the unbiased modularization of biochemical networks.** *Trends Biochem Sci* 2004, **29:**641-647.
8. Guimera R, Nunes Amaral LA: **Functional cartography of complex metabolic networks.** *Nature* 2005, **433:**895-900.
9. Ravasz E, Somera A, Mongru D, Oltvai Z, Barabasi A: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297:**1551-1555.
10. Bruggeman FJ, Westerhoff HV: **The nature of systems biology.** *Trends Microbiol* 2007, **15:**45-50.
11. Covert MW, Schilling CH, Palsson BO: **Regulation of Gene Expression in Flux Balance Models of Metabolism.** *J Theor Biol* 2001, **213:**73-88.
12. Fiehn O: **Metabolomics - the link between genotypes and phenotypes.** *Plant Mol Biol* 2002, **48:**151-171.
13. van der Werf MJ, Overkamp KM, Muilwijk B, Coulier L, Hankemeier T: **Microbial metabolomics: Toward a platform with full metabolome coverage.** *Anal Biochem* 2007, **370:**17-25.
14. van der Werf MJ, Pieterse B, van Luijk N, Schuren F, van der Werff-van der Vat B, Overkamp K, Jellema RH: **Multivariate analysis of microarray data by principal**

component discriminant analysis: prioritizing relevant transcripts linked to the degradation of different carbohydrates in *Pseudomonas putida* S12. *Microbiology* 2006, **152:**257-272.

15. Smilde AK, van der Werf MJ, Bijlsma S, van der Werff-van der Vat B, Jellema RH: **Fusion of mass-spectrometry-based metabolomics data.** *Anal Chem* 2005, **77:**6729-6736.
16. Steuer R: **On the analysis and interpretation of correlations in metabolomic data.** *Brief Bioinform* 2006, **7:**151-158.
17. Müller-Linow M, Weckwerth W, Hutt MT: **Consistency analysis of metabolic correlation networks.** *BMC Systems Biology* 2007, **1:**44.
18. Notebaart RA, Teusink B, Siezen RJ, Papp B: **Co-Regulation of Metabolic Genes Is Better Explained by Flux Coupling Than by Network Distance.** *PLoS Comput Biol* 2008, **4:**e26.
19. Steuer R, Kurths J, Fiehn O, Weckwerth W: **Observing and interpreting correlations in metabolomic networks.** *Bioinformatics* 2003, **19:**1019-1026.
20. Almaas E, Oltvai ZN, Barabasi AL: **The Activity Reaction Core and Plasticity of Metabolic Networks.** *PLoS Comput Biol* 2005, **1:**e68.
21. Madeira SC, Oliveira AL: **Bicluster Algorithms for Biological Data Analysis: A Survey.** *IEEE Trans Comput Biol Bioinform* 2004, **1:**24-45.
22. Van Mechelen I, Bock H-H, De Boeck P: **Two-mode clustering methods: a structured overview.** *Stat Methods Med Res* 2004, **13:**363-394.
23. Hageman JA, van den Berg RA, Westerhuis JA, van der Werf MJ, Smilde AK: **Genetic algorithm based two mode clustering.** *Metabolomics* 2008, **4:**141-149.
24. Lessie TG, Phibbs PVJ: **Alternative pathways of carbohydrate utilization in Pseudomonads.** *Annu Rev Microbiol* 1984, **38:**359-387.
25. van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ: **Centering, scaling, and transformations: improving the biological information content of metabolomics data.** *BMC Genomics* 2006, **7**.
26. Koek M, Muilwijk B, van der Werf MJ, Hankemeier T: **Microbial metabolomics with gas chromatography mass spectrometry.** *Anal Chem* 2006, **78:**1272-1281.
27. Coulier L, Bas R, Jespersen S, Verheij E, vanderWerf MJ, Hankemeier T: **Simultaneous Quantitative Analysis of Metabolites Using Ion-Pair Liquid Chromatography-Electrospray Ionization Mass Spectrometry.** *Anal Chem* 2006, **78:**6573-6582.
28. Zar JH: *Biostatistical analysis.* Prentice Hall; 1996.
29. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.** *Genome Res* 2003, **13:**2498-2504.
30. The Mathworks Inc.. **Matlab 7.3 (R2006b).** 2006.

7 Summary and outlook

Robert A. van den Berg

1 Important factors in a top down systems biology study

In top down systems biology, the answer to a certain biological question is sought in the systems wide response of a biological system to the chosen experimental conditions. The response of the biological system is measured with –omics tools, such as, transcriptomics or metabolomics, and advanced data analysis tools are applied to extract biologically relevant information from the measurements. The success of a top down systems biology approach is highly dependent on the information richness of the data obtained from the –omics measurements. Therefore, it is essential for a successful top down systems biology study to balance the three key factors: (i) biological question, (ii) experimental design, and (iii) data analysis. In Chapter 1, we discuss these three key factors, their interdependence, and their significance for successful top down systems biology. In Chapters 2 to 6, different aspects of the relation between a biological question and a data analysis strategy, such as, data pretreatment and selection of the most suited data analysis method, are further explored.

2 Data pretreatment

2.1 Translation of a biological question into the expected behavior of relevant biomolecules

Extracting relevant biological information from large data sets is a major challenge in functional genomics research. Different aspects of the data hamper their biological interpretation. For instance, 5000-fold differences in concentration for different metabolites are present in a metabolomics data set, while these differences are not proportional to the biological relevance of these metabolites. However, data analysis methods are not able to make this distinction. Data pretreatment methods can correct for aspects that hinder the biological interpretation of metabolomics data sets by emphasizing the biological information in the data set and thus improving their biological interpretability.

In Chapter 2, different data pretreatment methods i.e. centering, autoscaling, pareto scaling, range scaling, vast scaling, log transformation, and power transformation, were tested on a real-life metabolomics data set. They were found to greatly affect the outcome of the data analysis and thus the ranking of the, from a biological point of view, most important metabolites. Furthermore, the stability of the ranking, the influence of technical errors on data analysis, and the preference of data analysis methods for selecting highly abundant metabolites were affected by the data pretreatment method used prior to data analysis.

We found that different pretreatment methods emphasize different aspects of the data and each pretreatment method has its own merits and drawbacks. The choice for a pretreatment method depends on the biological question to be answered, the properties of the data set and the data analysis method selected. For the explorative analysis of the validation data set used in this study, autoscaling and range scaling performed better than the other pretreatment methods. That is, range scaling and autoscaling were able to remove the dependence of the ranking of the metabolites on the average concentration and the magnitude of the fold changes and showed biologically sensible results after PCA (principal component analysis). In conclusion, selecting a proper data pretreatment method is an essential step in the analysis of metabolomics data and greatly affects the metabolites that are identified to be the most important.

2.2 Removal of confounding variation from micro-array data

Confounding variation is variation that obscures the induced biological variation. Removal of the confounding variation can improve the interpretation of the data. In Chapter 3, we present a strategy to remove confounding variation based on an ANOVA approach, and to assess the impact of the removal on the interpretation of the variation induced by the experimental design. Our strategy is applied to an *Aspergillus niger* micro-array data set in which the variation induced by the experimental design was obscured by variation induced by the presence or absence of substrate. It was possible to remove the confounding variation; however, variation induced by the experimental design was partially removed as well. This was due to correlation between the variation induced by the experimental design and the confounding variation.

3 Conversion of a biological question into a data analysis question

3.1 The relation of a class of metabolites and its surrounding metabolic network

In metabolomics research it can be important to focus the data analysis to areas of specific interest within metabolism. For instance, the biological question under study can be related to a specific class of metabolites or a specific pathway. Supervised data analysis methods can bring this focus into data analysis and provide information on the behavior of the interesting metabolites in relation to the remainder of the metabolome. In Chapter 4, we describe the application of consensus PCA (CPCA) and canonical correlation analysis (CCA) as a means to focus data analysis. CPCA searches for major trends in the behavior of metabolite concentrations common for the metabolites of interest and the remainder of the metabolome. CCA identifies the strongest correlations between these two subsets.

CPCA and CCA were applied to two microbial metabolomics data sets. The first data set, derived from *Pseudomonas putida*, was relatively simple and contained metabolomes obtained under four environmental conditions only. The second data set, obtained from *Escherichia coli,* was complex and contained metabolomes from 28 different environmental conditions. For the first data set, CCA and CPCA gave similar results as the variation in the two subsets was similar. In contrast, CCA and CPCA yielded different results in case of the *E. coli* data set. With CPCA the trends in the metabolites of interest – the phenylalanine biosynthesis intermediates - dominated the results. These trends were related to high and low phenylalanine productivity, and important metabolites were associated with amino acid metabolism and regulation of the phenylalanine biosynthesis route.

With CCA neither subset dominated the data analysis. CCA described correlations between the subsets based on wild type and overproducing strain differences and different carbon sources. For the strain differences, metabolites from the aromatic amino acid pathways were important.

Both CCA and CPCA enable to focus the data analysis of metabolomics data to groups of metabolites that are of specific interest. Depending on the nature of the data set, they provide different, complementary, views on the relation between the metabolites of interest and the remainder of the metabolome.

3.2 Analysis of the behavior of subsets of metabolites under different environmental conditions

Metabolomics and other omics tools are generally characterized by large data sets with many variables and obtained under different environmental conditions. Clustering methods and more specifically two-mode clustering methods are excellent tools for analyzing this type of data. Two-mode clustering methods allow for analysis of the behavior of subsets of metabolites under different experimental conditions. In addition, the results are easily visualized. In Chapter 5 we introduce a two-mode clustering method based on a genetic algorithm that uses a criterion that searches for homogeneous clusters. Furthermore we introduce a cluster stability criterion to validate the clusters and we provide an extended knee plot to select the optimal number of clusters in both experimental and metabolite modes.

The genetic algorithm-based two-mode clustering gave biological relevant results when it was applied to two real life metabolomics data sets. It was, for instance, able to identify a catabolic pathway for growth on several of the carbon sources.

3.3 Discovery of functional modules in metabolomics data: regulation of cellular metabolite concentrations

In metabolism, functional modules can be defined as groups of metabolites that have a related function. Functional modules can be determined on different levels within the cellular organization. In response to normal, not stressful conditions, global regulatory effects will control the major physiological processes. These global regulatory effects are characterized by metabolites whose concentrations show a similar behavior in response to different environmental conditions. Changes in environmental conditions that perturb specific areas in the metabolism will provoke local regulatory effects. Metabolites whose concentration responds similar in response to such local perturbations will be part of the same local functional module. In Chapter 6, we identified both local and global functional modules based on two real-life microbial metabolomics data sets. Furthermore we discuss the nature of homeostasis, as is reflected by the regulation of metabolite concentrations.

Local functional modules were identified in two microbial metabolomics data sets originating from *Escherichia coli* and *Pseudomonas putida* S12 by a two-mode clustering approach. Their identification proved strongly dependent on the variation in environmental conditions under which the metabolome data were obtained. For instance, a local functional module containing citric acid cycle and redox-related metabolites was identified when *E. coli* was grown on succinate instead of D-glucose. The global functional modules were discovered by a correlation network analysis. Here, modules related to amino acid biosynthesis and the central metabolism were found. Comparison of the metabolite composition of local and global functional modules revealed that metabolites which are member of the same global functional module are not necessarily member of the same local functional module, and vice versa.

Regulation of metabolite concentrations was found to occur on different hierarchical levels. Whether these different hierarchical regulation levels could be identified in the metabolomics data set depended strongly on the environmental conditions – and thus the experimental design of the data sets - and how the selected conditions perturb the metabolism. By the application of two different data analysis methods both local and global functional modules could be identified.

4 Outlook further research

The results presented in this thesis offer several leads for further research both biologically as well as data analysis oriented. The previous chapters of this thesis were inspired by the translation of a biological question into a data analysis strategy. Although the research was illustrated by real life data sets, the final step, that is, validation of the findings in the laboratory was not made. Therefore, following up leads found in the previous chapters would be the proof of the pudding for the proposed approaches. Nevertheless, there remain more generic topics for research as well.

4.1 Functional modules

The search for functional modules in metabolomics data in Chapter 6 was a step towards a better understanding of the regulation of metabolite concentrations. The next step would be to set up a top down systems biology study in which the properties of global as well as local regulatory mechanisms are further explored. Ideally, such study would be developed for a well studied and relatively simple organism, for instance *Escherichia coli* or *Bacillus subtilis*. The study should be designed around (parts of) the central metabolism and around more condition dependent modules, e.g. amino acid biosynthesis. In this way, the regulatory differences with regard to global and local regulation between constitutive and inducible areas of the metabolism can be analyzed. While it is difficult, if not impossible, to establish a clear distinction [1] between local and global functional modules; studying the differences between, for instance, the response of the central metabolism and more condition dependent modules to certain perturbations in the experimental design can teach us more about this hierarchical distinction.

The intriguing global functional module found in Chapter 6 (Chapter 6, Figure 5) which consists for a large part of unidentified metabolites illustrates an important problem within metabolomics: the identification of unknown metabolites. In mass spectrometry, unknown compounds often remain unidentified due to the absence of reference compounds. It would therefore be useful for biologists and analytical chemists to try to relate the masses and hence possible chemical structures of the unidentified compounds to the behavior displayed by these unidentified compounds under the measured experimental conditions. The behavior of these unidentified compounds could reduce the number of possible chemical structures and provide additional information regarding their possible identity.

4.2 Integration of information from other sources

The application of the data analysis methods discussed in Chapter 4 can be extended to combine biologically relevant information obtained on different levels, such as, transcriptomics, proteomics and metabolomics, in the cellular organization in a broad top down systems biology fashion. Depending on the biological question, it is possible to emphasize the synergy, or to search for distinctive behavior between the different biochemical layers.

4.2.1 Potential data sources

The data that can be combined and utilized in a data analysis strategy does not have to be restricted to -omics measurements; other data sources can also be integrated. For instance, process parameters like nutrient consumption rates, pH control data, or biomass formation rates can be a valuable data source for the data analysis. They can help explain the behavior of the measured biomolecules. Also "soft" data like technicians observations, i.e. foam production or color changes, could be utilized to help extract information relevant to the biological question from the data. In addition, the data can be qualitative, e.g. color of the culture, and does not have to be limited to quantitative data.

For two data blocks, we briefly discussed weighting options, such as, every variable equally important, or each data block equally important (for *P. putida* metabolome and nucleotides; for *E. coli* metabolome and phenylalanine pathway) (Chapter 4). These weighting schemes can be extended to different weighting schemes for different data blocks depending on the nature of the data blocks, e.g. the reliability of the measurements in the different data blocks. Also subjective criteria, such as, the confidence researchers have in the different data blocks, could be applied as a weighting factor.

The examples of data sources discussed above are all directly linked to the conducted experiments and therefore share the experimental mode of the data. However, the data analysis could also benefit from information unrelated to the specific experiments, but related to, for example, literature knowledge regarding regulation mechanisms of metabolites/proteins/genes. An example of linking this type of information to gene expression is the use of DNA sequence information for the prediction of gene expression [2]. Specifically for metabolomics, linking the knowledge obtained from genome wide bottom up systems biology analyses (e.g. [3,4]) to metabolomics analysis could be very useful, for instance, in an experiment in which metabolomics samples are taken in short time intervals analogously to traditional flux analyses.

4.2.2 Challenges for data integration

It is not straightforward to utilize and balance the different sources of probably heterogeneous data mentioned in the previous section in a data analysis method. There are different strategies to utilize the additional data and these strategies depend on various factors: (i) the type and nature of data offered; (ii) the type and nature of the data with which the offered data should be combined; (iii) the biological question, et cetera. The examples provided below will therefore be limited to general ideas. For instance, the additional data could provide better estimates of starting positions for methods like two-mode clustering (Chapter 5). In this way, the influence of the added information is fairly mild as the starting position of the algorithm does not give guarantees regarding the end solution. An example of this could be to use transcription factor binding site information as a starting point for the two-mode clustering of gene expression data. In another analysis, the additional data could be fused with the measured data and be fully part of the data analysis (e.g. merging process data with metabolomics data in a CPCA setting), or the information could be used to compare behavior between different measurements (e.g. metabolomics data with simulation results from metabolic models in a CCA setting).

4.3 The performance of multivariate data analysis methods in systems biology

Combining different data sources can result in new biological knowledge. However, especially in the case of linking different types of –omics data sets, the reliable estimation of model parameters (i.e. PCA loadings) becomes more difficult due to addition of many extra biomolecules. Due to the additional biomolecules, it becomes increasingly more difficult to determine the specific contribution of one biomolecule to the data analysis model (collinearity). Since it is still not clear how severe this problem is, and how many experiments are required to be able to make reliable estimates of model parameters (see also Chapter 1), studying the performance of multivariate data analysis tools in –omics data analysis is still an important topic. An analysis of this problem could benefit from the different databases containing microarray expression data from many different experimental conditions and many different organisms. These micro-array databases could form a basis for studies of the dependence of the performance of data analysis methods on the number of samples. The micro-arrays selected for such a study should be of decent quality, e.g. with regard to reproducibility, and there should be some coherence in the experimental conditions of which samples were taken for the micro-array measurements.

5 Conclusion

This thesis discussed essential aspects of top down systems biology and focused on the relation between biological question and data analysis strategy. This thesis clearly demonstrates the interdependence of a biological question and a data analysis strategy. Depending on the articulation of the biological question, different choices within the data analysis strategy were made. For instance in Chapter 4 focusing on major trends or strongest correlation of a group of metabolites between its surrounding network determined the choice between CPCA and CCA; and in Chapter 6 different data analysis strategies were developed for the identification of global and local functional modules.

The interdependence of the different factors of the top down systems biology research triangle (Chapter 1) underlines the multidisciplinary nature of top down systems biology. To be able to make the best choices for a particular top down systems biology study, the collaboration of experts in biology, data analysis, and – depending on the studied biological level (transcriptome, proteome, or metabolome) – analytical chemistry is required. These multidisciplinary areas are, in my opinion, the most exiting areas as there are possibilities to generate synergy between the constituent fields. It seems that especially data integration offers these possibilities, and therefore I hope to be able to contribute to this area in further studies.

6 References

1. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402:**C47-C52.

2. Beer MA, Tavazoie S: **Predicting Gene Expression from Sequence.** *Cell* 2004, **117:**185-198.

3. Reed JL, Vo TD, Schilling CH, Palsson BO: **An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR).** *Genome Biol* 2003, **4:**R54.

4. Teusink B, van Enckevort FHJ, Francke C, Wiersma A, Wegkamp A, Smid EJ, Siezen RJ: **In Silico Reconstruction of the Metabolic Pathways of *Lactobacillus plantarum*: Comparing Predictions of Nutrient Requirements with Those from Growth Experiments.** *Appl Environ Microbiol* 2005, **71:**7253-7262.

Samenvatting

R. A. van den Berg

1 Belangrijke factoren bij een top down systeembiologie studie

In top down systeembiologie wordt het antwoord op een bepaalde biologische vraag gezocht in de reactie van een biologisch systeem op de gekozen experimentele condities. De reactie van het biologische systeem wordt gemeten met –omics methodes, zoals transcriptomics en metabolomics. Geavanceerde data-analysemethodes worden gebruikt om biologisch relevante informatie uit de meetgegevens te extraheren. Of een top down systeembiologieonderzoek succesvol is hangt sterk af van de informatierijkdom van de gegevens die met de –omics methodes zijn verkregen. Voor een succesvol systeembiologieonderzoek is het daarom essentieel dat er een balans wordt gevonden tussen drie sleutelfactoren: (i) de biologische vraag, (ii) de studieopzet, (iii) de data-analyse. In hoofdstuk 1 bespreken we deze drie sleutelfactoren, hun onderlinge afhankelijkheid, en hun belang voor succesvol systeembiologieonderzoek. In hoofdstukken 2 tot en met 6 worden verschillende aspecten van de relatie tussen de biologische vraag en de data-analysestrategie, zoals datavoorbewerking en de keuze voor de meest geschikte data-analysemethode verder onderzocht.

2 Datavoorbewerking

2.1 Het vertalen van een biologische vraag in het verwachtte gedrag van relevante biomoleculen

Het extraheren van biologisch relevante informatie uit grote datasets is binnen functional genomics onderzoek een van de grote uitdagingen. Verschillende aspecten van de data verstoren de biologische interpretatie van deze data. Het is bijvoorbeeld mogelijk dat de concentraties van verschillende metabolieten in een metabolomics dataset een factor 5000 verschillen, terwijl deze verschillen niet in verhouding staan tot de biologische relevantie van deze metabolieten. Data-analysemethodes kunnen dit onderscheid echter niet altijd maken. Datavoorbewerkingsmethodes zijn in staat om te corrigeren voor deze factoren die de biologische interpretatie van metabolomics datasets bemoeilijken. Er wordt gecorrigeerd door de biologische informatie in de dataset te benadrukken en hierdoor de biologische interpretatie te verbeteren.

In hoofdstuk 2 worden de prestaties van verschillende datavoorbewerkingsmethodes, zoals centreren, autoschalen, paretoschalen, rangeschalen, vastschalen, logtransformeren, en machtverheffen, met elkaar vergeleken door ze toe te passen op een metabolomics dataset. De datavoorbewerkingsmethodes hadden een zeer grote invloed op de data-analyseresultaten en dus de rangorde van de, vanuit een biologisch oogpunt, belangrijkste metabolieten. Bovendien had de keuze voor een bepaalde datavoorbewerkingsmethode invloed op (i) de stabiliteit van rangorde; (ii) de invloed van technische fouten op de data-analyse; en (iii) de voorkeur van data-analysemethodes om metabolieten te selecteren die in hoge concentraties voorkomen.

Uit ons onderzoek bleek dat verschillende datavoorbewerkingsmethodes andere aspecten van de data benadrukken en dat elke datavoorbewerkingsmethode zo zijn eigen voor- en nadelen heeft. De keuze voor een bepaalde data-analysemethode hangt af van de biologische vraag, de eigenschappen van de dataset, en de gekozen data-analysemethode. Voor de verkennende analyse van de dataset die in deze studie is gebruikt, presteerden autoschalen en rangeschalen beter dan de andere datavoorbewerkingsmethodes. Dat wil zeggen, rangeschalen en autoschalen waren in staat om de rangorde van de metabolieten onafhankelijkheid te laten zijn van de gemiddelde concentratie en van de grootte van de gemiddelde spreiding van deze metabolieten. Dit leidde tot biologisch zinvolle resultaten na PCA (principale componenten analyse). Wij concluderen dat het kiezen van een geschikte datavoorbewerkingsmethode een essentiële stap is in de analyse van metabolomics data en dat deze keuze een grote invloed heeft op identificatie van de rangorde van de meest belangrijke metabolieten.

2.2 Het verwijderen van storende variatie van micro-array data

De geïnduceerde biologische variatie in een micro-array dataset kan worden overschaduwd door storende variatie. Het verwijderen van deze storende variatie kan de interpretatie van de micro-array data verbeteren. In hoofdstuk 3 presenteren wij een strategie gebaseerd op ANOVA om deze storende variatie te verwijderen, en om de impact van deze verwijdering op de interpretatie van de door het experimenteel ontwerp geïnduceerde variatie te analyseren. Deze strategie is toegepast op een *Aspergillus niger* micro-array dataset. In deze dataset werd de variatie geïnduceerd door het experimentele ontwerp overschaduwd door variatie geïnduceerd door de aan- of afwezigheid van substraat. Het was mogelijk om de storende variatie te verwijderen. Echter, een deel van de geïnduceerde variatie werd hierdoor ook verwijderd. Dit effect werd veroorzaakt doordat een deel van de geïnduceerde variatie correleerde met de storende variatie.

3 Het vertalen van een biologische vraag in een data-analysevraag

3.1 De relatie van een klasse van metabolieten en het omringende metabole netwerk

Het kan binnen metabolomics onderzoek van belang zijn om de data-analyse te richten op specifieke belangrijke gebieden binnen het metabolisme. De biologische vraag kan bijvoorbeeld betrekking hebben op een specifieke klasse van metabolieten of een specifieke metabole route. Zogenaamde supervised data-analysemethodes kunnen de data-analyse hierop richten en zo informatie geven over het gedrag van de belangrijke metabolieten in relatie tot het omringende metabole netwerk. In hoofstuk 4 bediscussieren we de toepassing van consensus PCA (CPCA) en canonische correlatie analyse (CCA) om data-analyse te richten. CPCA zoek voor hoofdtrends in het gedrag van de metabolietconcentraties gemeenschappelijk zowel voor de belangrijke metabolieten als voor de rest van het metabolisme. CCA zoekt naar de sterkste correlaties tussen deze twee subgroepen.

CPCA en CCA zijn toegepast op twee microbiologische metabolomics datasets. De eerste dataset, verkregen met experimenten met *Pseudomonas putida*, was relatief simpel en bevatte metabolomes die verkregen waren onder slechts vier condities. De tweede dataset, verkregen met experimenten met *Escherichia coli*, was complex en bevatte metabolomes van 28 verschillende condities. Bij de eerste dataset gaven CCA en CPCA vergelijkbare resultaten omdat de variate in de twee subgroepen sterk vergelijkbaar was. Daarentegen leidde CCA en CPCA tot verschillende resultaten voor de *E. coli* dataset. Met CPCA domineerden de trends in de belangrijke metabolieten – de fenylalanine biosyntheseroute – de resultaten. Deze trends hadden betrekking op hoge en lage fenylalanineproductie. De voor deze trends belangrijke metabolieten waren onderdeel van het aminozuurmetabolisme en betrokken bij de regulatie van de fenylalanine biosyntheseroute.

Geen van de subgroepen domineerden de CCA. CCA beschreef de correlaties tussen de subgroepen die gebaseerd waren op verschillen tussen de wildtype en de overproducerende stammen, en verschillende koolstofbronnen. Belangrijk voor de verschillen tussen de stammen waren metabolieten van de aromatische aminozuurbiosyntheseroutes .

Zowel CCA als CPCA kunnen de data-analyse richten op groepen metabolieten met een specifiek belang. Afhankelijk van de eigenschappen van de dataset, kunnen deze methodes verschillende complementaire visies geven op de relatie van de belangrijke metabolieten en de rest van het metabole netwerk.

Analyse van het gedrag van subgroepen van metabolieten bij subgroepen van experimentele omstandigheden

Metabolomics en andere –omics tools resulteren in de regel in grote datasets met zeer veel variabelen welke gemeten zijn onder verschillende omstandigheden. Clustermethodes en in het bijzonder two-mode clustermethodes zijn uitstekende methodes om dit type data te analyseren. Zij kunnen het gedrag van subgroepen van metabolieten onder verschillende experimentele condities analyseren. Bovendien kunnen de resultaten gemakkelijk worden gevisualiseerd. In hoofdstuk 5 introduceren we een two-mode clustermethode gebaseerd op een genetisch algoritme dat zoekt naar homogene clusters. Bovendien presenteren we een clusterstabiliteitscriterium om de gevonden clusters te valideren, en een uitbreiding van de knieplot welke help bij de identificatie van het optimale aantal clusters.

De op genetische algoritmes gebaseerde two-mode clustermethode gaf biologisch relevante resultaten nadat het op twee metabolomics datasets was toegepast. De methode kon bijvoorbeeld een katabole metabole route identificeren betrokken bij de groei op een aantal van de gebruikte koostofbronnen.

3.3 Identificatie van functionele modules in metabolomics data: regulatie van cellulaire metabolietconcentraties

Binnen het metabolisme kunnen functionele modules worden gedefinieerd als groepen van metabolieten die een gerelateerde functie hebben. Functionele modules kunnen worden gevonden op verschillende niveaus binnen de cellulaire organisatie. Globale regulatiemechanismes zullen de belangrijkste fysiologische processen controleren als reactie op normale, niet stressvolle condities. Deze globale regulatiemechanismes worden gekarakteriseerd doordat de concentraties van bepaalde metabolieten hetzelfde gedrag vertonen als reactie op verschillende condities. Veranderingen in de condities die specifieke gebieden binnen het metabolisme verstoren zullen lokale regulatiemechanismes activeren. Metabolieten zullen onderdeel zijn van hetzelfde lokale functionele module wanneer de concentraties vergelijkbaar reageren in reactie op een verandering binnen een specifiek gebied van het metabolisme. In hoofdstuk 6 hebben we zowel lokale als globale functionele modules geïdentificeerd in twee microbiële metabolomics datasets. We bespreken hiernaast homeostase zoals dat wordt gereflecteerd in de regulatie van metabolietconcentraties.

Lokale functionele modules werden geïdentificeerd door middel van een two-mode clusteringmethode in twee microbiële metabolomics datasets afkomstig van *E. coli* en *P. putida* S12. De identificatie van lokale functionele modules bleek sterk afhankelijk van de variatie in de experimentele condities waaronder de metabolomes werden verkregen. Zo werd bijvoorbeeld een lokale functionele module gevonden die metabolieten uit de citroenzuurcyclus en redoxbalans gerelateerde metabolieten bevatte onder condities waar *E. coli* op succinaat werd gekweekt in plaats van D-glucose. De globale functionele modules werden ontdekt door middel van een correlatienetwerkanalyse. Hier werden modules gerelateerd aan aminozuurbiosynthese en het centraal metabolisme gevonden. Een vergelijking van de compositie van lokale en globale functionele modules liet zien dat metabolieten die lid zijn van hetzelfde globale module niet noodzakelijk lid zijn van hetzelfde lokale functionele module, en vice versa.

De regulatie van metabolietconcentraties vind plaats op verschillende hiërarchische niveaus. De identificatie van deze regulatieniveaus hing sterk af van de experimentele condities, en dus het experimentele ontwerp waarmee deze datasets zijn gegenereerd, en hoe deze condities het metabolisme verstoren. Door twee verschillende data-analysemethodes toe te passen konden zowel lokale als globale functionele modules worden gevonden.

4 Conclusie

Dit proefschrift bediscussieert essentiële aspecten van top down systeembiologie en richt zich op de relatie tussen de biologische vraag en de data-analysestrategie. De hoofdstukken van dit proefschrif illustreren de onderlinge afhankelijkheid van een biologische vraag en een data-analysestrategie. Afhankelijk van het benadrukken van verschillende aspecten van de biologische vraag werden andere keuzes binnen de data-analysestrategie gemaakt. In hoofdstuk 4 bijvoorbeeld bepaalde de focus op hoofdtrends of sterkste correlaties tussen een groep metabolieten en het omringende metabole netwerk de keuze tussen CPCA en CCA. In hoofdstuk 6 werden verschillende data-analysestrategieën ontwikkeld voor het identificeren van globale of lokale functionele modules.

De onderlinge afhankelijkheid van de verschillende factoren van de top down systeembiologiedriehoek (hoofstuk 1) onderstreept het multidisciplinaire karakter van top down systeembiologie. Om de beste keuzes te maken voor een bepaald top down systeembiologieonderzoek, is de samenwerking van experts binnen biologie, data-analyse, - en voor metabolomics en proteomics - analytische chemie vereist. Ik denk dat deze multidisciplinaire gebieden de interessantste onderzoeksgebieden zijn, omdat hier mogelijkheden zijn om synergie te genereren tussen deze onderzoeksvelden.

Nawoord

Met de afronding van dit proefschrift is er een einde gekomen aan mijn academische opleidingstraject die in 1997 aan de Landbouw Universiteit Wageningen begon. In de eerste fase van dit traject ben ik opgeleid tot labmicrobioloog, en in de tweede fase lijk ik het lab achter me gelaten te hebben om me verder te specialiseren in een veld waarvan ik in een kroeg het nut voor moleculair biologen nog weleens heb betwijfeld. Bij deze koerswijziging hebben Age, Mariët en Johan me met veel enthousiasme en veel discussies begeleid.

Mariët, bedankt voor de begeleiding bij zowel mijn afstudeervak als mijn promotieonderzoek. Ik ben erg blij voor je begeleiding met je openheid en je kritische en positieve blik. Ook waardeer ik de tijd die je zonder morren vrijmaakte en de snelheid waarmee je mijn stukken nakeek. Age, ook jij bedankt voor je kritische blik en je enthousiasme. Ik vind het heel prettig om met iemand te hebben samengewerkt die zo gemotiveerd is om continue te blijven leren. Ook ben ik blij met je bijdrage en deelname aan de studiereis naar Canada. Johan, bedankt voor de spoedcursus multivariate data-analyse die je me hebt gegeven en het geduld waarmee je mijn vragen hebt beantwoord. Ook bedankt voor je belangstelling en je toegankelijkheid. Het kluyver centre for genomics of industrial fermentations wil ik bedanken voor de financiering van mijn project.

Machtelt en Suzanne, mijn paranimfen. Ik vond jullie hele bijzondere collega's met een originele kijk op dingen. Machtelt, ik vond het heel gezellig met je op de kamer, op de fiets, en met etentjes en borrels in de stad. Suzanne, het was mede dankzij jou heel gezellig in Amsterdam. Ik vond het erg leuk met je bij te kletsen en zeker ook om de Canadareis met jou en Jos te organiseren.

Jos, jou wil ik bedanken voor de leuke samenwerking bij inmiddels twee en binnenkort drie artikelen. Het was leuk om met jou op te trekken, of dit nu op de kamer, cursus of congres was.

Erik en Henk-Jan, ik vond het erg gezellig bij jullie op de kamer, de discussies en verhalen waren interessant en vaak heel grappig.

Bart, jij hebt mijn keuze voor dit aioproject "gekatalyseerd" door mij nog eens op te bellen en over deze aioplaats te praten. Hierdoor zat ik onverwacht met Age en Mariët te praten over dit project en de invulling hiervan. Ook was het carpolen, borrelen en waren de etentjes heel gezellig en interessant.

Roelie, bedankt voor de gezelligheid op de kamer en in de auto.

Rolf, bedankt voor de gezelligheid en de lol op kamer. Ik hoop dat jij ook zo snel mogelijk je boekje af krijgt. Succes met de laatste loodjes.

Hennie, bedankt voor de leuke gesprekken, je belangstelling en je hartelijkheid.

Douwe, bedankt voor je vriendschap, en inmiddels dan ook eindelijk een samenwerking.

Ook wil ik mijn overige kamergenoten bedanken, waarmee ik wat minder intensief, maar niet minder leuk mee op de kamer heb gezeten, Olja, Ewoud, Helen, Amandine, en Rob. En collega's met wie ik heb samengewerkt of waarmee ik leuke gesprekken mee heb gevoerd. Van TNO: Martien, Norbert, Jan, Rob Leer, Karin, Nicole, Mieke, Sabina, Sabine, Carina, Bianca, Marian, Cora, Renger, Uwe, Bas, Leon, Eddy, Maud. En van Amsterdam: Huub, Daniël, Maikel, Hans, Susanna, Jeroen, Janko, en Tunahan. De collega's die ik niet met naam heb genoemd wil ik bedanken voor de leuke sfeer in Zeist en Amsterdam. Ook wil ik de facilitaire diensten van de UvA en van TNO bedanken voor de ondersteuning.

Rens, ik vond het superleuk om jou te begeleiden bij je stage. Ik heb heel veel van je geleerd, en je enthousiasme was erg motiverend.

Mijn vrienden, familie en schoonfamilie wil ik bedanken voor de belangstelling voor zowel de enthousiaste verhalen als mijn frustraties.

Pap, Ramona, en Miranda; bedankt voor jullie belangstelling en liefde.

Annelieke, bedankt voor je liefde, enthousiasme, steun, en geduld. Ik kijk uit naar de stappen die we samen nog gaan zetten.

List of publications

J. A. Hageman, R. A. van den Berg, J. A. Westerhuis, M. J. van der Werf, and A. K. Smilde. 2008. Genetic algorithm based two-mode clustering of metabolomics data. Metabolomics 4: 141-149.

J. A. Hageman, R. A. van den Berg, J. A. Westerhuis, H. C. J. Hoefsloot, and A. K. Smilde. 2006. Bagged K-Means Clustering of Metabolome Data. Critical Reviews in Analytical Chemistry 36: 211-220.

R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf. 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7: 142.

R. A. van den Berg, A. K. Smilde, J. A. Hageman, U. Thissen, J. A. Westerhuis, and M. J. van der Werf. Discovery of functional modules in metabolomics data: regulation of cellular metabolite concentrations. Submitted.

R. A. van den Berg, C. M. Rubingh, J. A. Westerhuis, M. J. van der Werf and A. K. Smilde. Identifying connections between a metabolic pathway and its surrounding network from metabolomics data. In preparation.

R.A. van den Berg, A.K. Smilde, J.A. Westerhuis, M.J. van der Werf. Key factors for successful top down systems biology in biotechnology. In preparation.

Curriculum vitae

Robert van den Berg werd op 15 December 1978 te Woerden geboren. Hij heeft VWO gedaan op het Kalsbeek College in Woerden. In 2003 voltooide hij zijn studie Bioprocestechnologie aan de Wageningen Universiteit met afstudeervakken in de richtingen Microbiologie en Moleculaire genetica van industriële micro-organismen. In dit jaar werd ook begonnen met het promotieonderzoek wat is beschreven in dit proefschrift. Het onderzoek werd uitgevoerd voor het Kluyver centre for genomics of industrial fermentations bij TNO Kwaliteit van Leven te Zeist en bij de Universiteit van Amsterdam. Het onderzoek werd begeleid door prof. Dr. A. K. Smilde, Dr. Ir. M. J. van der Werf en Dr. J. A. Westerhuis. Vanaf 2008 is hij werkzaam als post-doc bij de Research Group of Quantitative Psychology and Individual Differences aan de Katholieke Universiteit Leuven, te Leuven.