

# Quantitative multivariate NMR spectroscopy in Food Science and Nutrition

PhD thesis by  
Hanne Winning, 2009

Supervisors:

Assistant Professor Nanna Viereck

Professor Søren Balling Engelsen

FACULTY OF LIFE SCIENCES  
UNIVERSITY OF COPENHAGEN  
DEPARTMENT OF FOOD SCIENCE  
ROLIGHEDVEJ 30  
1958 FREDERIKSBERG C  
DENMARK



Cover illustration by Hanne Winning  
Title: Quantitative multivariate NMR spectroscopy in Food Science and Nutrition  
2009 © Hanne Winning  
ISBN: 978-87-7611-279-0  
Printed by SL grafik, Frederiksberg C, Denmark

## Preface

This PhD thesis is written to fulfil the requirements for obtaining a PhD degree at the Faculty of Life Sciences (LIFE) at the University of Copenhagen (KU). The presented work has been conducted in the research group Quality and Technology (Q&T) at the Department of Food Science, LIFE, KU under the supervision of Assistant Professor Nanna Viereck and Professor Søren Balling Engelsen.

I am utterly grateful to both my supervisors. They have helped and supported me in completing this thesis, what I at times thought impossible. Thank you Nanna for always being an optimistic and warm person, your support has been indispensable. And thank you Søren, for giving me the opportunity to make this PhD thesis and for always believing in me.

The presented publications involved several co-authors to whom I am grateful for their help as well as constructive and inspiring critique. Especially Flemming Hofmann Larsen is greatly acknowledged for help with NMR spectroscopy in general.

The Ministry of Science, Technology, and Innovation is greatly acknowledged for financially supporting the project “Exploring the asynchronous protein metabolism of wheat”.

I would also like to thank all my colleagues from Q&T for the good atmosphere. Special thanks go to Rasmus Bro, Lars Nørgaard, Karin Kjeldahl and Thomas Skov for inspirational and educational discussions about mathematics and chemometrics. And in particular, I will like thank Tina Salomonsen and my officemate Peter I. Hansen for friendship, support and many instructive NMR discussions.

And finally, thank you Kristoffer.

## Summary

Over the past years, nuclear magnetic resonance (NMR) spectroscopy has been increasingly used in food science and more recently, NMR based *metabonomics* has emerged as a field of great importance not only to medical scientists but also to nutrition scientists. The advantages of NMR spectroscopy are that the technique is non-destructive, fast and requires in most cases limited sample preparation and above all, NMR spectra provide detailed analytical profiles from the intact biological matrix. The NMR spectra are highly informative because  $^1\text{H}$  NMR spectra show proton signals of all constituents in a complex mixture. NMR spectroscopy is by definition a quantitative spectroscopic tool because the intensity of a signal is directly proportional to the number of resonant nuclei. Because of the quantitative nature of NMR data, it is well suited for multivariate data analysis.

Exploitation of complex NMR spectra of biological materials requires advanced data analytical techniques as well as qualified NMR spectral knowledge combined with biological understanding. Combining investigations of complex NMR spectra with advanced multivariate modelling (chemometrics) makes it possible to extract underlying common structures in the data. The complexity of NMR data makes it of prime importance to utilize data reduction techniques in order to access the latent chemical information in the data.

In this Ph.D. project, various multivariate data analytical methods have been applied to liquid-state  $^1\text{H}$  NMR spectra of biofluids, alcohols and hydrocolloid solutions as well as to  $^1\text{H}$  high resolution magic angle spinning (HR-MAS) NMR spectra of semi-solids such as wheat kernels. This work has resulted in the four papers described in the following.

The combination of NMR spectroscopy and chemometrics has tremendous potential in the metabonomics fields for exploring patterns of biomarkers of

diseases and food intake in biofluids. In this Ph.D. study, the ability of multivariate curve resolution (MCR) to separate mixtures into pure spectra and concentrations in a ternary experimental design of  $^1\text{H}$  NMR spectra of alcohol mixtures has been demonstrated. Under certain conditions, MCR offers resolution of the spectra into 'true' underlying components (i.e. the pure chemical spectra) as opposed to the principal component analysis (PCA) solution. However, MCR remains a research tool as it may provide multiple solutions and is prone to noise and interferences [Paper I].

NMR spectroscopy in combination with chemometrics can be used to explore the protein metabolism of single wheat kernels. Extreme climate events are becoming an increasingly important factor in crop growth and yield. NMR spectra of methanol extracts of wheat flour were found to be a good marker for drought treatment in mature kernels. The effect of different drought events on the protein fractions in grains of winter wheat was examined using NMR spectroscopy.  $^1\text{H}$  HR-MAS NMR spectra of single wheat kernels showed considerable differences between early and late harvest. Visualisation of the data by PCA trajectories of  $^1\text{H}$  HR-MAS NMR spectra of single wheat kernels showed good contrast of the metabolic development during grain filling. For the first time, grain filling data has been analysed by complex multiway modelling using parallel factor analysis (PARAFAC) resulting in an excellent overview of the data. The PARAFAC approach combined with the multiparametric NMR method has great potential for the study of dynamic metabolic events as the results are more direct and intuitive [Paper II].

In metabonomic studies with human volunteers, it is common knowledge that volunteers frequently do not report all their food supplements. Therefore, it would be highly desirable to obtain objective methods that can monitor a diet intervention. Animals can be exposed to much more controlled environments and diet and they can be much more identical with respect to gender, genetic and physics. In this study, *in vivo* investigations using rats as animal study-models made it possible to study biomarkers for onion intake. NMR spectroscopy in combination with chemometrics has shown to be an excellent tool for diverting diet groups by their urine profile. It was demonstrated that it was possible to distinguish between NMR spectra of urine from rats with a normal diet and rats on an onion diet. Searching for influential signals in the spectra with respect to differentiating between different onion products, interval extended canonical variates

analysis (iECVA) was applied. This revealed two signals from onion biomarkers, which were identified as dimethylsulfone (an oxidation product of dimethylsulfoxid) and 3-hydroxyphenylacetic acid. Being able to detect and maybe even identify specific diets could add new possibilities to human intervention studies [Paper III].

The classical use of NMR spectroscopy to food systems is the detailed quality control of ingredients and composition of hydrocolloids. In this study, this research was taken to the limit by investigating if NMR spectroscopy in combination with chemometrics can be used to predict degree of blockiness in pectins. This is important because the distribution of the free methyl ester groups is critical for the functionality of pectins and development of a fast and reliable method to examine the methyl ester distribution along the pectic polymer is most desirable. Until now, the only alternative analytical method was enzymatic degradation of the polymer combined with various spectroscopic techniques. In this study, a nearly complete assignment of the  $^1\text{H}$  NMR spectrum of pectin solutions was carried out and some very interesting correlations were shown between the methoxy signal and the H-2 and H-3 signals which were found to be due to the (nuclear Overhauser effect) NOE effects from the spatial structure. The results indicate that the pectic polymer has a three-fold helical conformation [Paper IV].

The collected work of this Ph.D. study shows examples of successful applications as well as the possibilities and limitations of applying chemometrics to complex NMR data.

## Resumé

Gennem de sidste årtier har anvendelsen af kernemagnetisk resonans (NMR) spektroskopi inden for fødevareforskning været tiltagende og i den seneste tid har forskning baseret på NMR metabonomic studier vist sig at være betydningsfuld for medicinske- såvel som ernæringsforskere. Fordelen ved NMR spektroskopi er at metoden er ikke-destruktiv, hurtig og i de fleste tilfælde kræver begrænset prøveforberedelse. Men vigtigst af alt indeholder NMR spektre detaljeret analytisk information om den intakte biologiske matrice. NMR spektre er yderst informative eftersom  $^1\text{H}$  NMR spektre indeholder protonsignaler fra alle bestanddele i en kompleks blanding. NMR spektroskopi er per definition en kvantitativ spektroskopisk teknik eftersom signalintensiteten er direkte proportional med antallet af de kerner, der bidrager til signalet. NMR data kvantitative karakter, gør metoden meget velegnet til multivariat dataanalyse.

Udforskning af komplekse biologiske NMR spektre kræver avancerede dataanalytiske teknikker samt kvalificeret NMR spektralviden kombineret med biologisk forståelse. Ved at kombinere komplekse NMR spektre med avanceret multivariat modellering (kemometri) er det muligt at ekstrahere fælles underliggende strukturer fra data. På grund af kompleksiteten af NMR data er det nødvendigt at anvende datareduktions teknikker for at finde den underliggende kemiske information i data.

I dette ph.d.-projekt er forskellige multivariate dataanalyseteknikker blevet anvendt på væske  $^1\text{H}$  NMR spektre af biologiske væsker, alkoholer og hydrokolloidopløsninger, samt på  $^1\text{H}$  'højopløst rotation omkring den magiske vinkel' (HR-MAS) NMR spektre af halvfaste materialer som hvedekerner. Dette arbejde har resulteret i de fire artikler beskrevet nedenfor.

Kombinationen af NMR spektroskopi og kemometri har et enormt potentiale indenfor metabonomics området til mønstergenkendelse af biomarkører i kropsvæsker fra sygdomsforløb og kostindtag. I dette ph.d.-studium er det blevet vist, at 'multivariat kurve tilpasning' (MCR) kan anvendes til at adskille blandinger i rene spektre og koncentrationer i et trekants design af  $^1\text{H}$  NMR spektre alkoholer. Under særlige omstændigheder kan MCR opløse spektre af blandinger i 'rene' underliggende komponenter (dvs. spektre af rene kemiske forbindelser) i modsætning principal komponent analyse (PCA) løsningen. MCR forbliver dog et forskningsværktøj, da metoden kan resultere i flere løsninger og er følsom overfor støj og interferenser [Paper I].

NMR spektroskopi kombineret med kemometri kan anvendes til at undersøge proteinmetabolismen i hvedekerner. Forekomsten af ekstreme klimaperioder en faktor af stigende vigtighed i forhold til afgrødevækst og udbytte. NMR spektre af metanolekstrakter af hvedemel viste sig at være gode markører for tørkebehandling i modne kerner. Effekten af forskellige tørkebehandlinger på proteinfraktioner i vinterhvedekerner blev undersøgt ved hjælp af NMR spektroskopi.  $^1\text{H}$  HR-MAS NMR spektre af enkelte hvedekerner viste betydelig forskel på tidlig og sen høst. Afbildning af data ved hjælp af PCA trajektorier af  $^1\text{H}$  HR-MAS NMR spektre af enkelt hvedekerner viste store forskelle på den metaboliske udvikling under kernefyldning. For første gang er kernefyldningsdata blevet analyseret ved hjælp af kompleks multivejsmodellering i form af parallel faktor analyse (PARAFAC), hvilket resulterede i et godt overblik over data. PARAFAC metoden kombineret med multi-parameter NMR metoden har et stor potentiale til at undersøge dynamiske metaboliske ændringer eftersom resultaterne er mere direkte og intuitive [Paper II].

Det er kendt viden at frivillige personer i metabonomic studier ikke altid rapporterer alle kosttilskud. Det er derfor ønskeligt at opnå objektive metoder, der kan overvåge kostintervention. I dette studium har *in vivo* undersøgelser af rotter som dyremodeller gjort det muligt at undersøge biomarkører for lögindtag. NMR spektroskopi i kombination med kemometri viste sig at være en fremragende metode til at adskille kostgrupper ud fra urinprofilerne. Det blev demonstreret, at det er muligt at skelne mellem NMR spektre af urin fra rotter, der har fået normal kost, fra rotter, der har fået lög i kosten. Interval udvidet canonical variat analyse (iECVA) blev anvendt til at finde signalerne i spektrene der har indflydelse på adskillelsen. Dette afslørede to signaler fra lögbiomarkører, som blev



identificeret som dimetylsulfon og 3-hydroxyphenyldikesyre. Dimetylsulfon er et oxidationsprodukt af dimetylsulfoxid (DMSO). At kunne bestemme og måske endda identificere specifik kost giver nye muligheder indenfor humane koststudier [Paper III].

Den klassiske anvendelse af NMR spektroskopi til fødevarer systemer er til detaljeret kvalitetskontrol af ingredienser og sammensætning af hydrokolloider. I dette studium er dette forskningsområde blevet udfordret ved at undersøge om NMR i kombination med kemometri kan anvendes til at bestemme graden af blokstruktur i pektiner. Dette er vigtigt eftersom fordelingen af frie estergrupper er afgørende for funktionaliteten af pektiner og udviklingen af en pålidelig hurtigmetode til bestemmelse af fordelingen af methylestergrupper er derfor meget attraktiv. Indtil nu har den eneste alternative metode været enzymatisk nedbrydning af polymeren kombineret med forskellige spektroskopiske teknikker. I dette studium er  $^1\text{H}$  NMR spektret af pektinopløsninger blevet næsten fuldstændigt tilordnet og nogle meget interessante korrelationer blev set mellem metoxysignalet og H-2 og H-3, hvilket skyldes 'Overhauser kerneeffekt' (NOE) fra molekylets rumlige struktur. Disse resultater indikerer at pektin polymeren er drejet i en 3-fold helisk struktur [Paper IV].

Dette Ph.D. studie viser eksempler på succesfulde anvendelser samt de muligheder og begrænsninger, der findes ved at anvende kemometri på komplekse NMR data.

## List of Publications

### *Refereed journal papers*

#### **Paper I**

H. Wining, F.H. Larsen, R. Bro, S.B. Engelsen, *Quantitative analysis of NMR spectra with chemometrics*, Journal of Magnetic Resonance **190** (2008) 26–32.

#### **Paper II**

H. Wining, N. Viereck, B. Wollenweber, F. H. Larsen, S. Jacobsen, I. Søndergaard, S. B. Engelsen, *Exploring abiotic stress on asynchronous protein metabolism in single kernels of wheat studied by NMR spectroscopy and chemometrics*, Journal of Experimental Botany **1** (2009) 291-300

#### **Paper III**

H. Wining, E. Roldán-Marín, L.O. Dragsted, N. Viereck, M. Poulsen, C. Sánchez-Moreno, M.P. Cano, S.B. Engelsen, *An exploratory in vivo metabonomic investigation reveals dimethylsulfone as a robust biomarker for onion intake*, **submitted** (2009)

#### **Paper IV**

H. Wining, N Viereck, L. Nørgaard, J. Larsen, S.B. Engelsen, *Quantification of the degree of blockiness in pectins using <sup>1</sup>H NMR spectroscopy and chemometrics*, Food Hydrocolloids **21** (2007) 256–266

### **Additional publications:**

#### **I**

H. Wining and S.B. Engelsen, *A ternary full-rank experimental design as viewed by chemometrics and NMR spectroscopy*, Magnetic Resonance in Food Science: From Molecules to Man, Eds: I.A Farhat; P.S Belton and G.A. Webb. The Royal Society of Chemistry, Oxford UK, **310** (2007) 205-213

## II

H. Winning, N Viereck, T Salomonsen, J Larsen, S.B. Engelsen, *Quantification of blockiness in pectins -A comparative study using vibrational spectroscopy and chemometrics*, Carbohydrate Research (2009), doi: 10.1016/j.carres.2008.10.015

## III

N. Viereck, H. Winning, H.F. Seefeldt and S.B. Engelsen, *<sup>1</sup>H HR MAS NMR: Profiling metabolites in single cereal kernels (of wheat and barley)*. Magnetic Resonance in Food Science: Challenges in a Changing World (2009). Eds: Guðjónsdóttir M, Belton P S and Webb G A. The Royal Society of Chemistry, Oxford UK. In Press

# Table of Contents

<b>Preface .....</b>	<b>3</b>
<b>Summary.....</b>	<b>4</b>
<b>Resumé.....</b>	<b>7</b>
<b>List of Publications.....</b>	<b>10</b>
<b>List of Abbreviations .....</b>	<b>15</b>
<b>1. Introduction .....</b>	<b>17</b>
Content outline.....	18
<b>2. Nuclear magnetic resonance .....</b>	<b>21</b>
Introduction.....	21
Applications of NMR spectroscopy .....	22
<i>NMR spectroscopy in metabonomics .....</i>	<i>22</i>
<i>NMR spectroscopy in food science.....</i>	<i>23</i>
Theory .....	23
<i>The resonance frequency .....</i>	<i>25</i>
<i>Chemical shift .....</i>	<i>26</i>
Solid state NMR spectroscopy .....	26
HR-MAS .....	27
CP-MAS .....	30
The NMR spectrometer.....	30
<i>Field strength and spectrometer frequency .....</i>	<i>32</i>
Pre-measuring steps .....	32
<i>Tuning and matching .....</i>	<i>32</i>
<i>Shimming.....</i>	<i>32</i>
<i>Receiver gain.....</i>	<i>32</i>
<i>Number of scans.....</i>	<i>33</i>
<i>Recycle delay.....</i>	<i>33</i>
<i>Water suppression .....</i>	<i>33</i>
The nature of NMR data .....	34
Spectral post-processing .....	34
<i>Phase and baseline correction.....</i>	<i>35</i>
<i>Chemical shift alignment .....</i>	<i>35</i>
<i>Intensity scaling .....</i>	<i>37</i>

Normalization .....	38
Binning, smoothing and data reduction.....	38
Quantitative NMR spectroscopy .....	39
Advantages and drawbacks.....	43
Sensitivity problems.....	43
Alternative NMR techniques .....	44
Other NMR active nuclei .....	44
Two dimensional NMR.....	44
SNIF NMR.....	45
<b>3. Chemometrics .....</b>	<b>47</b>
Multivariate data analysis.....	47
Unsupervised data exploration .....	48
PCA .....	48
MCR.....	52
PARAFAC.....	53
Supervised data exploration .....	54
PLS .....	54
iPLS .....	55
Validation .....	57
Classification methods.....	58
SIMCA.....	58
ECVA and iECVA .....	58
PLS-DA .....	59
Multivariate tricks of the trade.....	62
Orthogonalization.....	62
Forward selection .....	62
Correlation plot.....	63
<b>4. NMR as a quantitative high-throughput technique.....</b>	<b>65</b>
Process analytical technology.....	66
Visualization of NMR data.....	67
Signal overlapping .....	69
<b>5. NMR in the study of metabolites in intact food matrices.....</b>	<b>71</b>
Solid state NMR spectroscopy .....	71
HR-MAS NMR spectroscopy in the study of solid foods.....	72
Wheat quality.....	72
Quantitative aspects of HR-MAS NMR spectroscopy .....	73
Considerations concerning experimental setup.....	75
Three way data.....	75
<b>6. The dynamic metabolic events of NMR uroscopy in nutri-</b>	
<b>metabonomics .....</b>	<b>77</b>
Metabonomics.....	77
Uroscopy .....	78

<i>Biological variations in urine</i> .....	79
<i>Nutri-metabonomics</i> .....	79
Monitoring diet interventions by NMR spectroscopy.....	80
<i>The Diogenes study</i> .....	81
Exploration of biomarkers for onion intake.....	86
<b>7. NMR spectroscopy for advanced quality control of food</b>	
<b>ingredients</b> .....	<b>89</b>
Pectin: a case study.....	89
<i>Functionality</i> .....	90
<i>Blockiness in pectin</i> .....	91
<i>iPLS exploits functionality in the spectra</i> .....	92
<i>Spatial pectic conformation</i> .....	94
<b>8. Conclusion</b> .....	<b>97</b>
<b>9. Perspectives</b> .....	<b>101</b>
<b>Reference List</b> .....	<b>105</b>

## List of Abbreviations

2D	Two-dimensional
ALS	Alternating least squares
$B_0$	External magnetic field
$\delta$	Chemical shift
C	Concentrations profile
COW	Correlation optimized warping
COSY	Correlation spectroscopy
CP-MAS	Cross polarisation magic angle spinning
cS	Spectrometer constant
CVA	Canonical variates analysis
DE	Degree of esterification
DOSY	Diffusion ordered spectroscopy
E	Residuals
$\Delta E$	Energy difference
ECVA	Extended canonical variates analysis
FID	Free induction decay
FT	Fourier transforms
$\gamma$	Gyromagnetic ratio
GalA	Galacturonic acid
GC	Gas chromatography
$^1\text{H}$	Proton
$h$	Planks constant
HGA	Homogalacturonan
Hz	Hertz
HPLC	High-performance liquid chromatography
HR	High resolution
HR-MAS	High resolution magic angle spinning
HSQC	Heteronuclear single quantum coherence
I	Quantum number or signal intensity
IR	Infrared

iECVA	Interval extended canonical variates analysis
iPCA	Interval principal component analysis
iPLS	Interval partial least squares regression
$k$	Boltzmann constant
K	Kelvin
LC	Liquid chromatography
LF-NMR	Low field nuclear magnetic resonance
LV	Latent variables
MCR	Multivariate curve resolution
MHz	Megahertz
MS	Mass spectrometry
MRI	Magnetic resonance imaging
$\nu$	Larmor frequency
$N$	Population
NIR	Near infrared
NMR	Nuclear magnetic resonance
NOE	Nuclear Overhauser effect
NLR	Non-linear least squares regression
P	Loading matrix
PARAFAC	Parallel factor analysis
PAT	Process analytical technology
PC	Principal component
PCA	Principal component analysis
PLS	Partial least squares regression
PLS-DA	Discriminant partial least squares regression
ppm	Parts per million
$r^2$	Squared correlation coefficient
RG I	Rhamnogalacturonan I
RMSECV	Root mean square error of cross-validation
$s$	Standard deviation
S	Pure component spectra
SIMCA	Soft independent modelling of class analogy
S/N	Signal to noise ratio
SNIF-NMR	Site-Specific natural isotope fractionation-NMR
$\sigma$	Shielding constant
$T$	Temperature
T	Score matrix
TOCSY	Total correlation spectroscopy
TSP	Trimethylsilyl proprionic acid



# 1. Introduction

Traditionally, food analysis involves identifying and classifying food constituents into very broad categories of molecules such as protein, fats, carbohydrate and solids. With new high resolution and high throughput modern techniques, the multiple nuclear magnetic resonance (NMR) spectroscopy has become a popular tool for food authenticity and other food studies.

NMR spectroscopy is an extremely versatile measurement technique and has become one of the most valuable tools for obtaining structural and dynamic information of biological molecules. The ability to measure liquids as well as solids without any upfront separation steps makes it a unique analytical method compared to various destructive chromatographic analytical methods. NMR spectroscopy provides a rapid non-destructive identification and is by definition a quantitative spectroscopic tool because the intensity of a resonance line is directly proportional to the number of resonant nuclei. Over the past years, NMR spectroscopy has been increasingly used in food science and increasing in the *metabonomics* field where the study of biofluids has become of increasing interest to food and nutrition scientists.

NMR spectroscopy is able to detect large numbers of metabolites simultaneously and can obtain a far more detailed molecular picture of food composition, food consumption and of the molecular consequences of different diets compared to the ordinary vibrational spectroscopy methods. Exploitation of complex NMR spectra of biological materials requires the quantitative properties of NMR spectroscopy is completely exploited. Chemometric or multivariate methods are routinely utilized in other areas of spectroscopy (most notable in NIR spectroscopy; see for example Additional publication II) for analysis of complex mixtures. However, the challenges in data analysis of NMR spectra are very different, where spectral sensitivity and spectral pre-processing have to cope with e.g. signal overlap are of vital importance.

The aim of this thesis is to demonstrate the new possibilities of NMR spectroscopy research of food and nutrition and to characterise the potential in the use of chemometrics in combination with NMR spectroscopy. The thesis also emphasizes the importance of understanding the limitations and pitfalls of quantitative multivariate NMR spectroscopy. An example of this is the instrumental parameter setup which influence the resulting data and thereby the subsequent conclusions. There also lies great risk in the data analysis including a potential risk of overfitting of the chemometric models due to insufficient validation or simply lack of information in the samples. A great challenge also lies in the sampling and the sample preparation technique.

### ***Content outline***

The present thesis is based on three peer-reviewed papers and one submitted for peer review. Additional work has been done providing the basis for these results. The main body of the thesis is divided into an NMR theory part, a chemometric theory part and an application part describing the four selected applications of multivariate NMR spectroscopy.

Chapter 2 serve as a basic introduction to NMR spectroscopy, introducing high field liquid and solid state NMR spectroscopy. A brief introduction to the NMR spectrometer is given together with a description of the critical instrument parameters with a special view to quantitative NMR spectroscopy. The chapter sum up the advantages and drawbacks of quantitative NMR spectroscopy in food and nutrition research.

Chapter 3 introduces a number of useful multivariate methods in the data analysis of complex NMR data. Besides the basic chemometric methods such as principal component analysis (PCA) and partial least squares regression (PLS). Much emphasis will be made to the advantages of using interval based approaches such as interval PCA (iPCA), interval PLS (iPLS) and interval extended canonical variates analysis (iECVA). In addition, recommendations and cautions are given on order to avoid overfitting including proper validation and common sense.

In Chapter 4, the potential of using NMR spectroscopy for exploring biofluids in the metabonomics fields and in process analytical technology (PAT) is characterized. An example from [Paper I] is given, concerning the

usefulness of multivariate curve resolution (MCR) to separate mixtures into pure spectra and concentrations in a ternary experimental design of  $^1\text{H}$  NMR spectra of alcohol mixtures.

Chapter 5 illustrates the possibilities and advantages of measuring intact food matrices. In the study of intact biological matrices there are many critical parameters which have to be dealt with. In addition, the limitations of semisolid state NMR spectroscopy are described and an alternative method of managing these measurements is given. [Paper II]

Chapter 6 deals with emerging problems in NMR metabonomics studies and what to be aware of and how to solve these. This is demonstrated in a case study using rats as animal models for diet interventions where advanced chemometric data-mining revealed specific biomarkers for onion intake [Paper III]

Chapter 7 explore the capability of NMR spectroscopy to elucidate block structures of pectins (a common food ingredient) and describe how interval based chemometric modelling can be used to obtain knowledge about informative regions of the NMR spectrum. [Paper IV]

Finally, the conclusions of this Ph.D. study and the perspectives of future work and ideas are collected in Chapters 8 and 9.



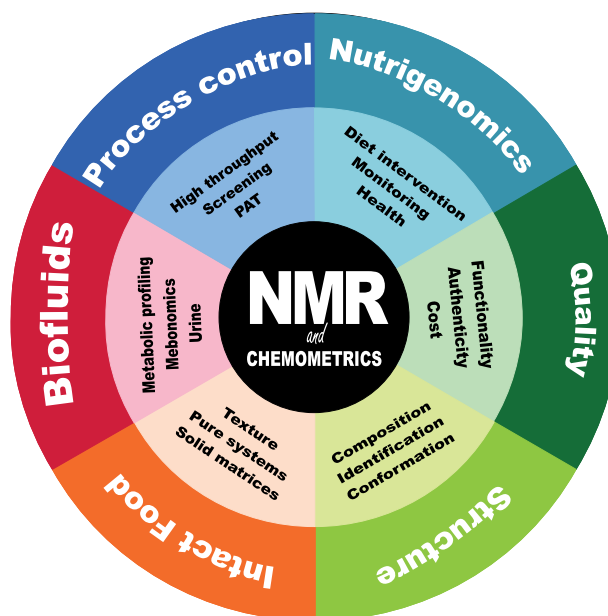
## 2. Nuclear magnetic resonance

### *Introduction*

Since the introduction of NMR spectroscopy by Rabi and Breit [1,2] in 1931 and its further development by Bloch and Purcell [3], NMR spectroscopy has become one of the most valuable tools for obtaining structural and dynamic information of various compounds. In 1952 the Swiss and the American physicists Bloch and Purcell shared the Nobel Prize in physics for discovering the basis of magnetic resonance imaging (MRI). About 50 years later, Lauterbur and Mansfield shared the Nobel Prize in 2003 for conducting the work that led to present MRI technology. Continuous improvements in NMR instruments and techniques have resulted in many new applications. This led to the selection of Richard R. Ernst in 1991 as the Nobel Laureate in Chemistry for his work in developing Fourier transformed (FT) NMR spectroscopy. And in 2002, Kurt Wüthrich was selected as Nobel Laureate in chemistry for the development of NMR spectroscopy for determining the three-dimensional structure of proteins in solutions. Hence, NMR spectroscopy is a well developed unique analytical method.

NMR spectroscopy has a long tradition in structure elucidation of biological as well as synthesized compounds. However, in the past years NMR spectroscopy has been increasingly used in a quantitative manner. Although originating from physics, it is in chemistry that NMR spectroscopy has attracted the greatest interest and now further finds application in biology, medicine, geology, materials science and recently, in food science. Until recently, most work was focused on clinical or pharmaceutical applications such as drug discovery [4,5], drug assessment [6], clinical toxicology [7,8] and clinical chemistry [9,10]. However, over the past few years, NMR based *metabonomics* has emerged as a field of increasing interest to food and nutrition scientists [11] which will be discussed in the following paragraph.

The unique versatile applications of NMR spectroscopy are illustrated in Figure 1.



**Figure 1** Nuclear magnetic resonance (NMR) is a unique versatile analytical method able to perform chemical identification and quantification, to elucidate chemical connectivity, to measure intact matrices, and is a highly valuable tool in the analysis of biofluids, in process control (process analytical technology (PAT) and nutritional diet intervention studies.

## *Applications of NMR spectroscopy*

### **NMR spectroscopy in metabonomics**

*Metabonomics* was defined as “the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification” by Nicholson and co-workers at Imperial College (London) in 1999 [12].  $^1\text{H}$  NMR spectroscopy has emerged as a promising non-invasive technique for metabonomics studies due to its ability to simultaneously detect a large number of compounds in a rapid and high-throughput manner that requires little sample manipulation [13]. The very closely related *metabolomics* is defined as the comprehensive and quantitative analysis of all metabolites of a biological system which are identified and quantified [14]. Where the goal of *metabolomics* primarily is to

quantify the numerous small molecules found in biological fluids under different conditions, *metabonomics* is the study of the changes in the metabolic profile of a complex biological system in response to different stimuli. The work presented here belongs to metabonomics and therefore the term *metabonomics* will be employed throughout this thesis. A thorough explanation of NMR spectroscopy in metabonomics studies are given in Chapter 6.

### **NMR spectroscopy in food science**

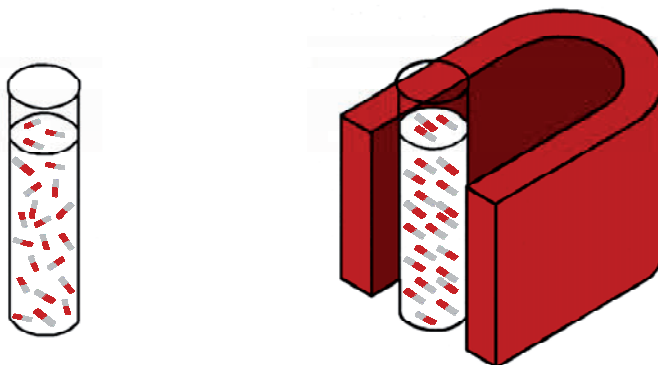
High resolution liquid state  $^1\text{H}$  NMR spectroscopy has many applications in food science. Examples are the analysis of natural food products such as milk [15], wine [16-18], fruit juices [19,20], ciders [21], olive oil [22], port wine [23] and beer [24,25]. The ability of NMR spectroscopy to handle liquid as well as solid samples makes it a very unique analytical tool. Solid-state NMR spectroscopy has also shown many successful applications which will be further elaborated in the section on solid-state NMR spectroscopy.

In order to fully appreciate the potentials of NMR spectroscopy within food science are, the fundamental concepts of NMR spectroscopy will be briefly explained in the following paragraph. The NMR spectrometer is described together with a number of important instrumental settings and parameters. This is followed by a description of the necessary spectral post-processing (after Fourier transformation), and their manifestation in the final NMR spectra. Last, the quantitative properties of NMR spectroscopy will be described. This chapter will end with a conclusion concerning the advantages and drawbacks of the use of NMR spectroscopy within food science and metabonomics.

### **Theory**

NMR spectroscopy is based on the magnetic properties of the atomic nucleus. The magnetic properties of a nucleus are based on the concepts of nucleus spin and the magnetic moment of the nucleus. About two-thirds of all isotopes, among these  $^1\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$ ,  $^{17}\text{O}$ ,  $^{23}\text{Na}$  and  $^{31}\text{P}$ , possess nuclear magnetic moments, and are characterized by an angular momentum quantum number  $I = \frac{1}{2}$  (or increments of  $\frac{1}{2}$ ). Here, only nuclei with  $I = \frac{1}{2}$  are considered. These nuclei can be considered as small magnetic dipoles. At normal conditions, the magnetic dipole axes of the nuclei are randomly distributed in all directions. However, when a nucleus with a spin is

exposed to a stationary external magnetic field ( $B_0$ ), the magnetic dipoles align parallel ( $I = \frac{1}{2}$ ) or anti-parallel ( $I = -\frac{1}{2}$ ) to the external magnetic field with a small dominance in the parallel direction (Figure 2).



**Figure 2** Schematic illustration of the alignment of the protons in an external field. At normal conditions, the magnetic dipole axes of the nuclei are randomly distributed in all directions (left) but when exposed to a stationary external magnetic field, the magnetic dipoles align (right) [26]

The relative distribution of the population of nuclei in the two energy states is given by the Boltzmann distribution (Equation 1) where  $N$  is the fraction of the population of nuclei in each energy state,  $T$  is the absolute temperature and  $k$  is the Boltzmann constant:

$$\frac{N_{-\frac{1}{2}}}{N_{\frac{1}{2}}} = \exp\left(-\frac{\Delta E}{k \cdot T}\right) \quad (1)$$

The result of this uneven distribution is a weak net magnetization aligned exactly along the axis of  $B_0$ . It is this weak net magnetization that is measured by NMR spectroscopy. It is normal to ascribe the external magnetization as  $z$  in an  $x$ - $y$ - $z$  coordinate system, where the  $x$ - $y$  plane is orthogonal to the external magnetic field and is in fact where the signal is detected after perturbation with a  $90^\circ$  radio frequency pulse.

The net magnetization will start to precess around the direction of the magnetic field ( $z$ ) with a frequency which is unique for a nucleus depending on the local chemical environment (see section on high field NMR



spectroscopy). Though, when the sample is at equilibrium there is no observable signal. By applying radio frequency pulses at exactly the frequency of the nucleus, the net magnetisation can be flipped from the z-axis into the x-y plane and thus obtain an observable signal. After the pulse, the system will return to equilibrium, so-called relaxation. Consequently, the readable magnetization in the x-y plane decay to zero in a certain time which is recorded by the spectrometer in the form of a free induction decay (FID). The time domain is widely used directly in low field (LF) NMR relaxometry, operating at the lower range of magnetic field strength (frequency < 100 MHz) typically using permanent magnets. LF-NMR spectroscopy is used within food research to detect distribution and mobility of water and fat and physical conditions within a sample like solid-fat ratio in meat, edible oils, moisture or oil content etc.

### **The resonance frequency**

When operating at higher field strength (operating at a  $^1\text{H}$  NMR frequency of > 200 MHz) it becomes possible to distinguish between same types of nuclei within the same spectrum and thereby extract chemical information because the frequency range is higher. The population difference  $N_{-1/2} - N_{1/2}$  and thereby the energy difference  $\Delta E$  depends on the magnetic field strength (Equation 2), where  $h$  is Planck's constant and  $\nu$  is the resonance frequency or the Larmor frequency. The Larmor frequency depends on the specific nucleus and the magnetic field strength  $B_0$  and  $\gamma$ , the magnetogyric ratio of the nucleus. The magnetogyric ratio is a proportionality constant that describes the spin state energies of a given nucleus in an external magnetic field.

$$\Delta E = h \cdot \nu = \frac{h \cdot \gamma \cdot B_0}{2\pi} \quad (2)$$

At a field of 60 MHz, the population excess is only 0.001% (Equation 1), but at a field of 600 MHz, the population excess is 0.01%, resulting in a detection level in the micromolar range. This is utilized in high field (or high-resolution, HR) NMR spectroscopy. Indeed, this makes high resolution NMR spectroscopy a useful analytical method.

During an NMR experiment, the signal is measured in the time domain, i.e. as a function of time. The signal is then Fourier transformed to obtain the

spectrum in the frequency domain. Thereby, the FIDs from all components in a sample which superimpose in the time domain are sorted out according to frequency by the Fourier transformation.

### **Chemical shift**

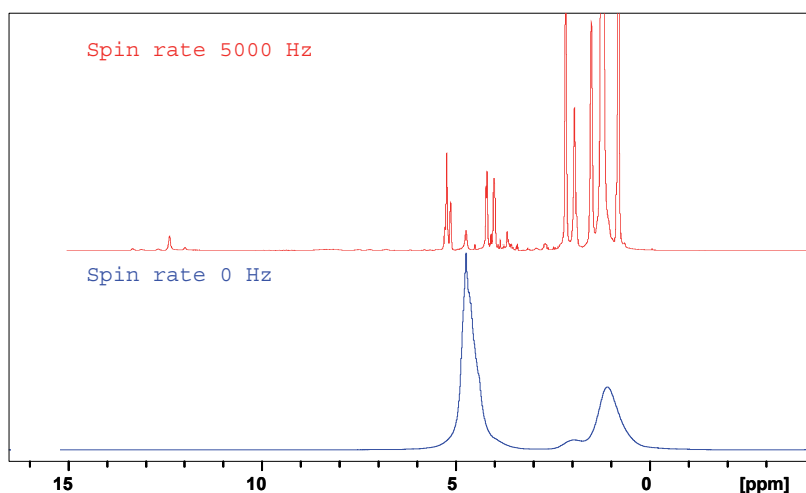
The local field experienced by a magnetic nucleus is modified by the interaction of local electron density, an effect known as shielding. In a molecule, electrons are hindered in their rotation around a particular nucleus by the presence of other atoms and are therefore not capable of exerting their maximum degree of shielding. Thus, local differences in the shielding reflect the fact that different nucleus in different chemical environments experience different effective magnetic fields. This is the why the resonance frequency of a nucleus depends on its chemical environment. It is common practice to define an entity called chemical shift ( $\delta$ ) as the frequency difference between the nucleus and a reference nucleus. This definition allows the resonance frequency of a signal to be expressed independent of the field strength of the magnet field strength which is used to measure it. The chemical shift is expressed in parts per million (ppm). The most widely used reference substance for  $^1\text{H}$  NMR spectroscopy is TSP-d4 (per-deuterated 3-trimethylsilyl propionate sodium salt), as it exhibits almost complete shielding. The difference between the nucleus and a reference with maximum shielding (per definition 0 ppm) will result in positive value on the axis, as most compounds have a smaller shielding than TSP.

### ***Solid state NMR spectroscopy***

When the mobility of a spin system decreases the relaxation rate will increase. Since the width of an NMR signal depends on the relaxation rate, NMR signals become broader as molecular mobility is restricted. This is the case in solid and semi-solid systems (cereals, cheese, meat etc.). Indeed, this has important practical implications. The broader the line, the harder it is to detect the signal, and the greater the chance of overlap between adjacent lines to give a poorly resolved spectrum. The molecular mobility further decreases when molecular weight increases, and as a result, it is only the freely mobile, low molecular weight substances that are detected in liquid NMR spectra of for example foods. This is the reason why, the many NMR spectroscopy studies of food have focused on liquids or liquid extracts rather than analysis of solid food.

### HR-MAS

Fortunately, the continuous advancement within NMR instrumental technologies today makes it possible to obtain well-resolved  $^1\text{H}$  spectra of semi-solid samples. High resolution magic angle spinning (HR-MAS) NMR spectroscopy [27-29] has become an extremely versatile tool to study heterogeneous systems. HR-MAS NMR spectroscopy relies on fast spinning of the sample at the magic angle to average out magnetic susceptibility differences in the sample in order to obtain resonance line widths approaching those of liquid state NMR spectroscopy. Accordingly, the use of the MAS results in a well-resolved NMR spectrum. This is a result of the sample being spun fast around an axis which makes an angle to the applied magnetic field of  $54.7^\circ$  (the so-called magic angle) in order to simulate molecular motion and the dipolar interactions can therefore be averaged out. Consequently, the combination of the magic angle and fast rotation of the sample eliminates dipole-dipole couplings and chemical shift anisotropy (both physical expressions include  $3\cos^2(54.74) - 1 = 0$ ) which induce significant line broadening in the NMR spectra of solids. These effects are averaged out in solutions due to fast isotropic motions. Indeed, HR-MAS NMR spectroscopy makes solids appear liquid-like, and the spectral resolution similar to liquid-state NMR spectroscopy.



**Figure 3**  $^1\text{H}$  HR-MAS NMR spectra of mozzarella cheese with no spin and a MAS spin rate of 5000 Hz.

Figure 3 show an example of the difference in resolution of a spectrum of mozzarella cheese recorded with no spin and a spectrum of the same cheese recorded using HR-MAS spin rate of 5 kHz (unpublished results).

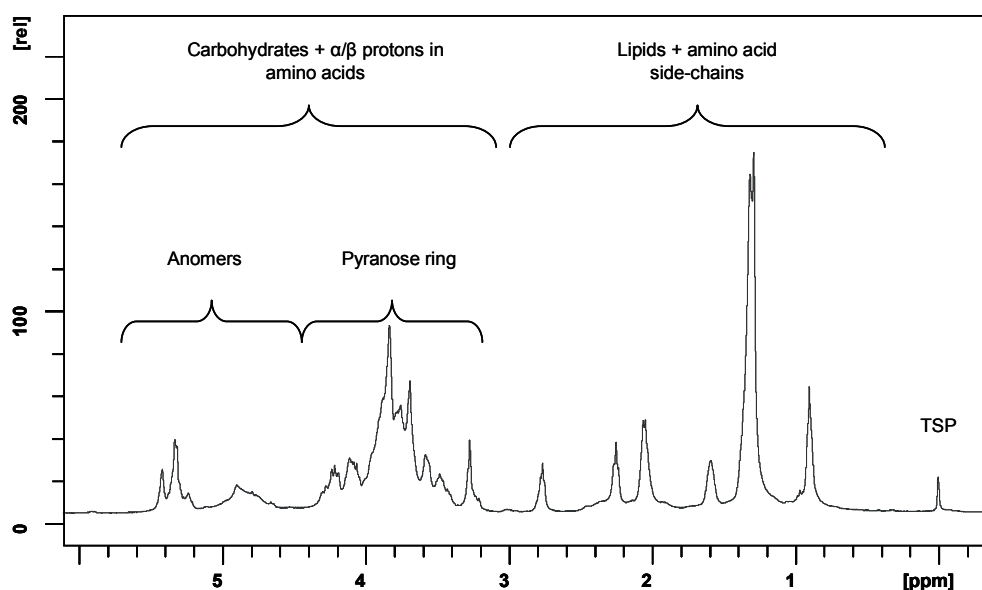
The HR-MAS NMR technique has been widely used to explore biological tissues from animals and plants. The domain of application of HR-MAS NMR spectroscopy is extremely diverse and includes the analysis of molecules in membranes [30], mesoporous materials [31], cells [32-35] and biopsies [36-42]. Numerous advantages attributed to HR-MAS NMR spectroscopy have motivated its use in the metabonomics field, in particular for human pathological biopsies [43]. However, a practical evaluation of HR-MAS NMR spectroscopy on a wide experimental range (*in vivo* animals, fresh or frozen tissues) reveals logistic, sanitary, and analytical disadvantages. The main drawback concerns the high difficulty to obtain correct metabolite quantification because of the sample preparation. HR-MAS NMR spectroscopy has in principal a straightforward sample preparation, but in practice, it is complicated due to the manual character of the sample preparation. The sample is placed in a small rotor (4 x 20 mm) which is subsequent filled with deuterated solvents. Furthermore, an insert is included which has to keep the sample in place and to ensure that no air bobbles is present, see Figure 4.



**Figure 4** Picture of an HR-MAS rotor together with an insert, screw, cap and some whole wheat kernels.

This sample preparation makes HR-MAS NMR spectroscopy slow and laborious in practice. Indeed, the preparation is critical because HR-MAS NMR spectra are highly sensitive to the shape, design and placement of the sample and how well the rotor is filled. This makes some samples difficult to measure, because the matrix to matrix variation is larger than the variation between, for example, treatments.

Another disadvantage of this method is that immobile parts of the sample remain unobserved. Thus, it is possible that the recorded spectra not are representative of the total sample. HR-MAS NMR spectroscopy is not a real solid-state method and therefore not really suited for very solid material like, for example, dry mature wheat kernels. All the protons are not mobilized and therefore invisible why immobile starch and fibres are not possible to quantify. Protons from mobile carbohydrates and  $\alpha$ - and  $\beta$ -protons from amino acids, as well as protons from anomers are also visible. Furthermore, signals from lipids are clearly visible. See Figure 5.



**Figure 5** HR-MAS NMR spectrum of a single wheat kernel. Lipid signal dominate although they represent less than 4 % of the sample material

Under HR-MAS NMR conditions, the samples are subjected to high centrifugal forces that may cause irreversible damage. This can for example result in unwanted transport of water molecules between compartments.

Recently, it has been possible to study the asynchronous protein metabolism in single wheat kernels using HR-MAS NMR spectroscopy during grain filling (Paper II). However, some difficulties arise when measuring mature dry single kernels because of the lack of mobility of the protons. Instead,  $^{13}\text{C}$  cross-polarization magic angle spinning (CP-MAS) NMR spectroscopy would perhaps be a feasible method to yield a carbon profile of the protein metabolism in the wheat kernels.

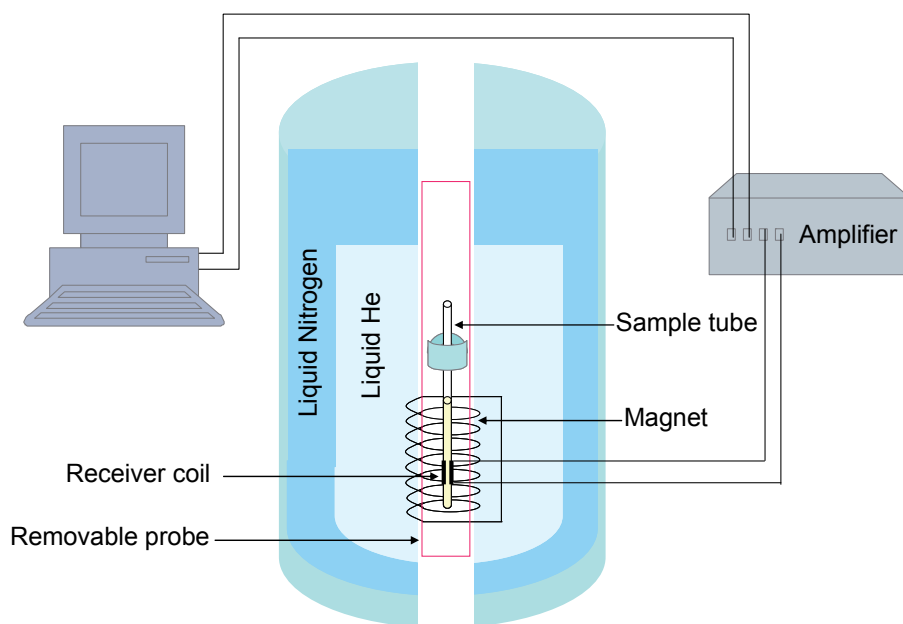
### **CP-MAS**

If real solid state measurements are to be performed, the CP technique [44,45] combined with MAS is the most common NMR experiment. The magnetization in CP-MAS spectroscopy is transferred from  $^1\text{H}$  by dipolar coupling to the directly attached  $^{13}\text{C}$  nuclei. This results in an increase in inherent sensitivity of carbons attached to one or more protons and, because relaxation via the  $^1\text{H}$  pathway is rapid rather than the slow  $^{13}\text{C}$  route, experiments can be repeated more rapidly, with consequent increase in signal-to-noise ratio (S/N). Thus, combining the techniques CP, MAS and high power decoupling of protons enables one to acquire fairly well-resolved  $^{13}\text{C}$  NMR spectra of solid samples.

Applications of CP MAS spectroscopy in food science have so far mainly been applied to carbohydrate research. Especially starches have been exploited with CP-MAS spectroscopy [46-52] but also meat [53], gluten [54], canola seed oil composition [55] and polysaccharides in sugar beet cell walls [56] has been investigated.

### ***The NMR spectrometer***

A high resolution NMR spectrometer is basically constructed with a super conducting electro magnet, a probe which contains the sample in either a tube or a rotor, a radio frequency transmitter to pulse and receive the emitted response, preamplifiers, to boost the analogue NMR signal and a computer to collect, digitize and Fourier transform the signals (Figure 6).



**Figure 6** A schematic illustration of the NMR spectrometer with a removable probe containing the receiver coil which collects signal from the sample. The super conducting magnet is cooled by liquid nitrogen and liquid helium and the receiver is connected to an amplifier and a computer.

The high field NMR spectrometer is a delicate instrument. It is highly sensitive to outer magnetic fields and extremely influenced by mechanical vibrations and therefore demands constant fine-tuning. In order to obtain the highest quality spectra, several adjustments should be made before the measurement. Therefore, an NMR spectrometer demands highly qualified personnel to run the samples. These pre-measuring steps are processing is described in the following paragraph. A major and in many cases prohibitive disadvantage of high field NMR spectroscopy is that it is one of the most expensive analytical techniques to employ, both in terms of the initial instrumental costs and running costs. The super conducting magnet demands as low temperature as possible. Liquid nitrogen cools the instrument down to about 77 K and the inner core which consist of liquid helium cools the system further down to 4 K. NMR spectrometers thus use large amounts of liquid nitrogen as well as liquid helium which is very costly. Therefore, the prize of a high field instrument required for

metabonomics studies, cost can be a major consideration in the development of large scale metabonomics screenings.

### **Field strength and spectrometer frequency**

NMR spectrometry can be considered as an indirect spectroscopy, as it requires an external magnetic field. Hence, the magnetic field strength is an important parameter for NMR spectroscopy. The high fields available from superconducting magnets have been improved the sensitivity of the NMR experiments, and the increased chemical shift dispersion which has made it possible to study more complicated chemical systems [57]. The spectrometers are often referred to by their proton frequency rather than by magnetic field strength. As the strength of the magnetic field is increased, the resolution and S/N ratio are also increased, which leads to higher information content, allowing the detection and characterization of smaller amounts of material and more complex molecules. Other important acquisition parameters are discussed in the following.

### ***Pre-measuring steps***

#### **Tuning and matching**

The sensitivity of the NMR probe varies with the optimal signal on a given sample at the given field frequency and there is only one frequency at which the probe is most sensitive to the chosen nucleus. This frequency has to be adjusted on each sample, a process called tuning in NMR terms. This is done by tuning the probe circuit. Matching involves ensuring that the maximum amount of power arriving at the probe base is transmitted up to the coil and reflected back at the amplifiers. Modern NMR probes have automatic tuning and matching to the chosen nucleus.

#### **Shimming**

Shimming is a process in which minor adjustments are made to the magnetic field within the probe surrounding the sample until the field homogeneity is with the sample in place. Optimal field homogeneity enhances the spectral resolution and thereby the spectral information increases from the sample.

#### **Receiver gain**

The optimal signal dynamics may also be optimized on each sample. Receiver gain is an important parameter that is used to match the amplitude of the FID to the range of the digitizer, i.e. receiver gain is a scaling factor.



The gain may automatically be set, but it can cause problems when measuring many samples in a quantitative series with different receiver gains due to non-linear response of the amplifiers. Then, the variation due to the different receiver gains is larger than the quantitative variation between the samples. As a minimum, the gain factor should be divided out before further data analysis.

### **Number of scans**

The number of collected scans, the so-called signal averaging, depends mostly on the decision of how long the experiment must take. The higher number of scans, the better the S/N. The basic principle of signal averaging is straightforward; if many spectra are added together, the signal intensity will increase. The noise is random and when adding together, the noise signals will also increase in intensity, but more slowly than the signal increases. In fact, the summation of  $n$  spectra, the signal will become  $n$  times bigger compared to a single spectrum, but the noise will increase by  $\sqrt{n}$  only.

### **Recycle delay**

Another important parameter for NMR spectroscopy is the recycle delay. Since the spin system must be allowed to recover entirely from the pulse before another pulse can be applied, the recycle delay is introduced as the time between the collections of each spectrum. If the relaxation is slow and the delay between the pulses is not long enough, the magnetization will not recover fully to equilibrium between the pulses. The recycle delay must be 5 times the longitudinal time. Ideally, if the acquisition time is long enough, the FID will show an exponential decay reaching zero. In order to improve the S/N, an exponential function can be multiplied the FID before the Fourier transformation. The FID will thereby be more rapidly decaying which will result in a broader signal but with less noise.

### **Water suppression**

Biological samples often contain large amounts of water which result in a very strong  $^1\text{H}$  NMR signal from water that dominates the spectra to an extent where compounds in low concentrations becomes impossible to observe. In many cases, protons with a chemical shift close to the chemical shift of the water protons will disappear under the huge water signal. The simplest and most robust water suppression method is pre-saturation where the water signal is turned out of phase by a long radio frequency pulse,

affecting only the water signal. The method is simple and reasonably effective, but has the disadvantage of also partly suppressing other signals in vicinity of the water peak. Other, more advanced techniques such as WATERGATE [58,59] are very effective but demands optimization of the specific area of suppression. Since water is not truly separated from the sample before measuring NMR spectroscopy, residual signals from water need subsequently to be removed from the spectra before any data analysis. However, in the process of removing such signals there is a chance that potentially important metabolite signals are also removed or partially suppressed which consequently deteriorate subsequent quantitative data analysis [60].

### *The nature of NMR data*

An NMR spectrum is highly informative as every single peak is a signal arising from protons attached at exactly the same position in a molecule and in exactly the same environment. The signals are normally very narrow and contain information about neighbouring protons which manifest as splitting of the signal. Moreover, the signals in an NMR spectrum are often baseline separated. Proton NMR spectroscopy are rapid experiments since the proton is the most abundant NMR nucleus providing inherently high S/N, and therefore useful for screening of many samples. However, protons relax relative slowly wherefore repeated experiments (number of scans) can only be performed with a time delay of typically 5 seconds between each scan. However, in a matter of minutes, it is possible to obtain informative spectra with thousands of variables containing narrow, baseline separated signals, with good dispersion and a high S/N.

### *Spectral post-processing*

Experimental variations in the recording of NMR spectra within a dataset leads to difficulties in the quantitative determination of peak height and will complicate the interpretation and analysis by quantitative pattern recognition methods. Minor experimental variations include small differences in the NMR spectral phasing, small shifts in line positions (chemical shift shift), small variations in line shapes and baseline. These variations result either from chemically induced changes or instrumental instability and field inhomogeneity [61]. For this reason, post-processing of the NMR data is required if optimal quantitative NMR spectroscopy is to be

carried out. All the necessary processing steps are described in the following.

### **Phase and baseline correction**

Until recently, spectral variations was sought to be manually corrected which is highly time intensive for large data sets. Commercial NMR software now includes multiple methods to automatically correct the baseline inconsistency and phase adjustment. Several numerical methods have been proposed for rapid and automatic adjustment of both the phase and frequency variation present within large NMR data sets. For example, PCA has been introduced for simultaneously phase correction of single resonances in a series of NMR spectra [62-66]. In this way, phase and frequency correction can be applied across the entire NMR data set simultaneously. This method appears to work well for phasing and frequency shifting, but only of spectral regions that contains well resolved resonances, not multiple overlapped spectra which is the case in complex NMR spectra of biological matrices. For this reason, complex NMR spectra are often manually phased corrected. However, baseline offset are most often automatically corrected.

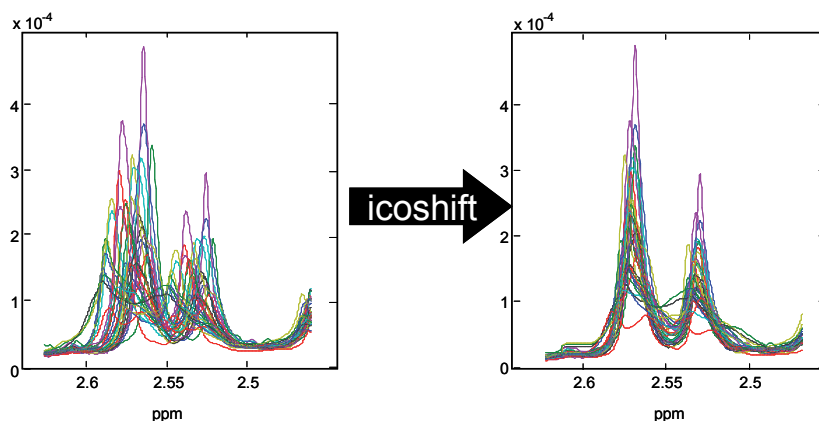
### **Chemical shift alignment**

A major problem in NMR spectroscopy is the chemical shift variability (chemical shift shift) induced by pH or temperature differences or small instrumental drifts. While chemical shift variations may be relevant from a chemical point of view, they will deteriorate the bilinear nature of NMR data, making any modelling unnecessarily complex. Shift of NMR signals are most often observed in pH sensitive signals which is why only a part of the spectrum is affected. In order to remove the peak shifts, different procedures have been developed.

One way of handling shifts variability in metabonomics research was by the use of binning or bucketing of the spectra to smooth-out small chemical shift shifts. The binning procedure will be discussed later in this section. The most widespread method to eliminate small global chemical shift shift is to perform rigid shift (co-shift) of the spectra according to the TSP signal. Another option is to globally shift the spectra according to a another important and non-sensitive signal such as the  $\alpha$ -glucose peak [67] or to a reference spectrum. These global methods are easy, fast and reliable but remove only shifts induced during measurements and will not to remove

local shifts induced by e.g. pH. Many processing methods have been developed for alignment of chromatographic data which in contrast to most spectroscopic techniques have inherently unstable x-axis (elution time) [68-71]. Correlation optimized warping (COW) has been employed for signal alignment. The latter algorithm is a time dynamic programming that selects predetermined spectral segments of equal length and by linear interpolation stretches and shrinks them to fit a reference spectrum. The evaluation function used to find the best fit warping is the correlation coefficient. COW has also shown promising results for correcting chemical shift shifts in NMR spectra of apple juice [72]. However, COW is not always successful for NMR data due to the complexity of the spectra and the relatively large chemical shift shifts which is why the simpler co-shift method often is preferred. COW and co-shift cannot handle frequency shift crossing, where signals in a data matrix are shifted both upwards and downwards of the spectrum compared to the reference spectrum. Other alignment techniques are available such as genetic algorithm [73], beam search algorithm [74], and fuzzy warping [75]. Recently, interval shifting has been proposed as a practical tool for alignment of NMR spectra [18,76]. In this approach, selected intervals are aligned whereas the rest of the spectrum remains unchanged.

In the metabonomics field, especially when measuring urine, signal shifting is a major problem which leads to difficulties sorting out the important signals. In urine, the chemical shift shifts are due to pH variation or small differences in the experimental conditions, but also reflect an individual variation, depending on the metabolism of the individual. The pH variations can largely be handled by adding buffer, but still, some misalignment will remain. As an example, Figure 7 show 32 urine samples from rats visualising misalignment of the two doublets from the citrate signal but otherwise well aligned spectra. Interval shifting (icoshift) was applied to the specific signals resulting in much better aligned signal in Figure 7. The consequence of the improved alignment was a noticeably reduction of the number of components in the subsequent predicting model and an improved prediction.



**Figure 7** Interval based co-shifting (icoshift) applied to the doublet arising from citrate signals resulting in perfect aligned signal (yet unpublished in-house algorithm)

When measuring less complex samples, a rigid shift of the spectra referenced to the TSP signal or another constant signal is often preferred [67]. This is also performed in Paper I-IV where the co-shift referenced to the TSP signal was adequate to remove small misalignments.

### Intensity scaling

From a biological point of view, metabolites present in high concentrations are not necessarily more important than those present in low concentration [77]. Scaling methods are data pre-treatment approaches that divide each variable by a factor (a column operation) such that each variable in the analysis have mean zero and unit variance. If concentration determination is the final objective, no scaling of the NMR data should be done, since the relative intensities of the NMR resonances are proportional to the concentration of the observed nuclei. On the other hand, the dominant resonances within the NMR spectra may not necessarily be the spectral features that reveal systematic variation occurring within the analysed samples [61]. Therefore, down weighting of those variables that are the least stable can become a necessary step. A number of scaling methods are used for NMR data including mean-centring [78] (see next chapter), auto-scaling, range scaling [79], pareto scaling [80] and vast scaling [81].

### **Normalization**

Under ideal conditions, NMR peak integrals are directly proportional to concentrations. However many biological, experimental and instrumental variables can affect absolute NMR peak intensities. Normalization methods are data pre-treatment approaches that divide each sample by a factor (a row operation). One common method of normalization involves setting each observation (spectrum) to have unit total intensity by expressing each data point as a fraction of the total spectral integral [82]. However, minor perturbations in the concentration of abundant metabolites will affect the scaling of all other metabolites. An alternative normalization approach, probabilistic quotient normalization, was recently introduced to reduce scaling artefacts resulting from normalization by the total spectral integral [83]. Torgrip et al. (2008) have introduced the histogram matching approach which can reduce normalization related variance [84].

### **Binning, smoothing and data reduction**

A common pre-processing approach is to smooth the data set by binning or moving average, where the horizontal axis is divided into (equal or unequal) regions and thereby a sum of the spectral intensity in each region is calculated [85]. Broader bins can be defined to cover for peak shifts and the resulting 'binned' data are less sensitive to the alignment problem. In the past, binning has been standard practice in dealing with NMR data, not least because the binned spectra contain far fewer data points than the original data. High resolution NMR spectra are most often recorded in at least 32 k ( $k = 1024$ ) data points and the FID is subsequent zero filled once to 64 k prior to Fourier transformation. This results in more detailed spectra with enhanced resolution but also in a very heavy data load. The most common bin size for a 64 k data point  $^1\text{H}$  NMR spectra was 0.04 ppm, because this corresponds to 256 equal size bins in the  $\delta$  range -0.2-10.00. (This corresponds exactly to the number of columns allowed in the old Microsoft Excel spread sheet.) However, the penalty is considerable loss of spectral resolution and complicated interpretations. Minor, but important, signals may be included in the same bin as major, but non-significant, signals. At best, this complicates the interpretation; at worst, important information may be seriously diluted and overlooked. Nowadays, binning should not be a necessary step due to computational load. Instead, most applications use the full spectral resolution as acquired, being aware that some regions of the spectrum will suffer from poor peak alignment (Paper I-IV).

The size of the NMR data set to be analysed can advantageously be reduced by removal of unused or undesired spectral regions, such as the water signal due to insufficient water suppression or removal of the reference (TSP) signal which hold no chemical information about the sample. Removal of spectral areas with low or no viability or particular noisy regions can in some cases improve the subsequent modelling.

### *Quantitative NMR spectroscopy*

NMR spectroscopy is by definition a quantitative spectroscopy because the intensity of a resonance line is directly proportional to the number of resonant nuclei and NMR spectroscopy can thereby determine the direct physical context of a substance without referencing to another substance. Since the intensity  $I$  of a signal is directly proportional to the number of nuclei  $N$  evoking the signal, the linear relationship is given by:

$$I = cS \times N \quad (3)$$

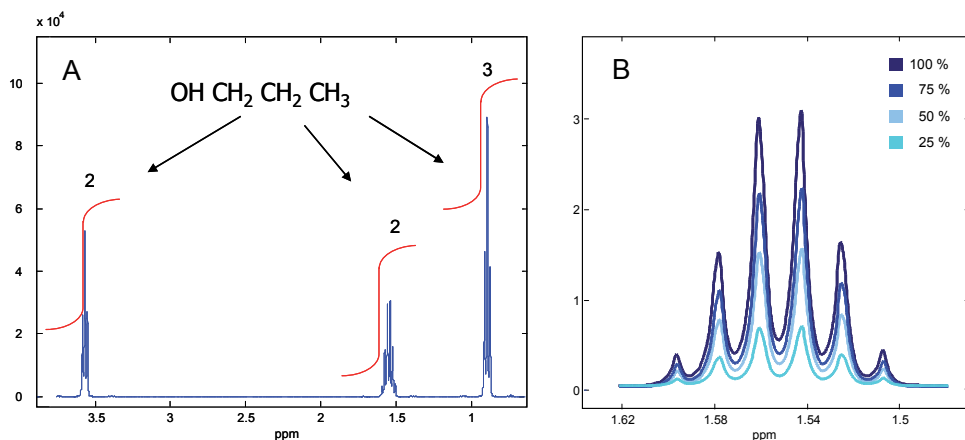
The proportionality constant  $cS$  results from parameters of the spectrometer, termed “spectrometer constant”, and the sample. The accuracy of quantification depends on the noise level of the spectrum, on the line shape, the quality of shimming and phase-, baseline- and drift corrections. These parameters were considered in the previous section.

Compared to the more common optical spectroscopic measurement techniques, e.g. IR and NIR which measures bond vibrations, NMR spectroscopy gives far more detailed information, since  $^1\text{H}$  NMR spectroscopy simultaneously detect all proton-containing compounds in, for instance, biological material, such as carbohydrates, amino acids, organic and fatty acids, amines, esters, ethers and lipids. Therefore, NMR spectroscopy is widely used for the analysis of complex mixtures and hundreds of metabolites can be detected without any upfront separation. And compared to chromatographic separation techniques, NMR spectroscopy is faster, better quantitative and more reliable, only the sensitivity is worse. The advantages of NMR spectroscopy over for example the common used high-performance liquid chromatography (HPLC) techniques are many. Quantitative NMR analysis is often more accurate and precise than standard HPLC methods without isolation of the impurity, no

expensive chemical reference substances are necessary. NMR spectroscopy is less time consuming (no equilibration time), easy to perform and more specific leading to a high reproducibility [86]. Wishart et al. (2008) made a comparison study on NMR spectroscopy, gas chromatography-mass spectrometry (GC-MS) and liquid chromatography mass spectrometry (LC-MS) in order to quantify all the metabolites that can be commonly detected in the human cerebrospinal fluid metabolome. Out of the total 308 compounds, NMR spectroscopy was the most versatile method, able to detect 53 compounds, having GC-MS second best detecting 41 compounds [87]. However, GC-MS and LC-MS are generally much more sensitive compared to NMR spectroscopy which has a detection limit of about 1  $\mu$ M. But the chromatography techniques have major drawbacks as the techniques are slow, require separation and demand that the compound is polar, which makes the techniques unable to detect the more abundant compounds such as sugars and some amino acids and organic acids. The fact that NMR spectroscopy is complementary to chromatographic methods makes it an attractive tool in drug analysis as well as in analysis of biological samples. A rather new hyphenated technique combines the methods in LC-NMR/MS, where the sample after chromatographic separation is transferred to an NMR spectrometer. In this way, the sensitivity is optimised and the technique is efficiently used in screening of e.g. pharmaceutical products.

An example of the highly quantitative nature of NMR spectroscopy is highlighted in (Paper I). The paper elucidates the characteristics of being able to detect intramolecular quantification, i.e. within a molecule and intermolecular quantification i.e. between molecules. The intra-molecular quantification is illustrated in Figure 8 where the peak integral of the CH<sub>3</sub> peak is confirming the theory of being 2/3 of the CH<sub>2</sub> peak in propanol. The left plot in Figure 8 illustrates the intermolecular quantitative nature of NMR spectroscopy. The intermolecular quantification is further described in Chapter 4 where the illustrative experimental design is fully recovered in the subsequent modelling.





**Figure 8** Illustration of intramolecular (A) and intermolecular (B) quantisation of propanol NMR spectra

Modern quantitative NMR spectroscopy could find a great use in the food industry in monitoring quality control, detection of adulterated or contaminated food products or as “metabolic” fingerprinting of process streams and raw materials.

In exploratory NMR spectroscopy, all the chemical compounds are generally not identified. Only their spectral pattern and intensities are statistically compared and used to identify the relevant spectral features that distinguish sample classes. Fingerprinting techniques involve collecting spectra of unpurified samples under standardized conditions and ignoring, initially, the problem of making individual assignments of peaks in the resulting complex and overlapping NMR spectra. The detection of adulterated or contaminated food products exploits the fact that certain chemicals or certain concentrations of chemicals are quite characteristic for a given product. Similarly, chemical composition characteristics can also be exploited to distinguish between food products with desirable characteristics that cannot otherwise be detected. Food quality assessment also impacts food quality control. In this approach, the focus is on attempting to identify and quantify compounds in the sample which have nutritional or functional properties. This is usually done by comparing the NMR spectrum to a spectral reference library obtained from pure compounds [88]. Quantitative  $^1\text{H}$  NMR analysis have already been employed in the analysis of a number of food products such as wine [16-18],

fruit juices [19,89], ciders [21], olive oil [22], flour [90], cheese [91-96] port wine [23] and beer [24,25]. Table 1 shows examples of the use of liquid proton NMR spectroscopy in food studies.

**Table 1** Examples of the application of high field  $^1\text{H}$  NMR spectroscopy to complex food systems

Product	Quality control	Adulteration/ authenticity	Metabolic profiling
Wine	Effect of climate, soil, and cultural on grape and wine quality [16,97]	Origin characterisation of Italian wines [17]  Characterization of Slovenian and Italian wine [98]	Metabolite fingerprints of grape berry [99]  Wine fingerprinting [18]
Port wine	Characterization of different ages of wine [23]		
Beer	Quality control of beer [24,25]	Discrimination between production site [100]	Fingerprinting of beer and wine [101]
Juice	Quantisation of formic acid in apple juice [20]  Analysis of malic and citric acids in fruit juices [102]	Fraud prevention of orange juice using grape juice [103]  Adulteration of orange juice [104,105]	Discrimination between apple varieties [19]
Cider		Discrimination of different cider apple juices [21]	Characterisation of cider apple juice [106]
Coffee	Quality control of different coffee producers [107]		
Vegetable oil	Classification of olive cultivars [22]	Adulteration of olive oil [108]  Origin control of pistachio oil [109]	
Fish oil	Oxidation of fish oil [110]	Authentication of origin of salmon [111]	
Balsamic vinegar		Discrimination of vinegars and balsamic vinegar [112]	
Milk/ dairy	Discrimination of yoghurt based on fruit content [113]	Origin of breeding of cow and buffalo milk [94]  Origin of buffalo milk and mozzarella cheese [93]	
Cheese		Production chain of Asiago d'Allevo cheese [96]	
Potato	Prediction of sensory texture and quality of potatoes [114]	Authentication study of pomodoro di Pachino [115]	
Tomato	Quality sorting of tomatoes [116]		Carotenoid profiles in tomato [117]
Rhubarb			Rhubarb's stalk characterization [118]

Besides liquid-state  $^1\text{H}$  NMR spectroscopy, also alternative techniques are potential in food analysis studies. This is discussed in the end of this chapter.

In the metabolomics field, NMR spectroscopy (together with MS) is the number one choice of analytical method. Identification of biomarkers can involve the application of a range of techniques, but  $^1\text{H}$  NMR spectra of blood and other biofluids, even though very complex, allow many resonances to be assigned directly based on their chemical shifts, signal multiplicities and by adding authentic material. In addition, further information can be obtained by using spectral editing techniques or interrogation of spectral databases of authentic substances [119].

### ***Advantages and drawbacks***

NMR spectroscopy has some distinct advantages over other instrumental methods: It is a high-throughput technique, quantitative, it is non-invasive and non-destructive and allows extraction of both chemical and physical information, and it is possible to get additional structural information about impurities, isomers, etc. Although NMR spectroscopy has many advantages, including relatively high-throughput sampling, and new sensitivity improvements, the technique has some major drawbacks. Some of problematic drawbacks are considered in the following.

### **Sensitivity problems**

Sensitivity is a significant issue as already mentioned. The disadvantage of low sensitivity is based on the small difference of the spin populations in the two energy states; the population difference, which is given by the Boltzmann equation (Equation 1). In metabolomics research, the sensitivity can become a problem as many metabolites are present below mg/ml concentrations which can make them difficult to detect. It is possible to, increase the number of scans (accumulation) which will increase the S/N ratio and somehow compensate for low sensitivity.

Higher field strength can also compensate for the low sensitivity as already briefly mentioned. A ten times increase of the field strength will result in a ten time increase of the percentage of the protons be detected. Another problem is that NMR spectroscopy often requires a relatively large sample amount (500  $\mu\text{l}$ ). A possible approach of maximization of concentration and

optimizing solvent volume is to use microcoil technology where e.g. a flow probe only uses 150  $\mu\text{l}$ . Nano-NMR probes or microcryo probes can actually increase the S/N by a factor of about 10 for  $^1\text{H}$  measurements. Finally, careful and precise (gradient) shimming techniques can increase the quality of the spectra and increase the final sensitivity.

### ***Alternative NMR techniques***

#### **Other NMR active nuclei**

Besides the most commonly observed nuclei,  $^1\text{H}$  and  $^{13}\text{C}$ , other nuclei can be studied by NMR. The sensitivity will depend on the magnetogyric ratio, which is inherent to the nuclei. Furthermore, when the natural abundance of an isotope is low, the actual sensitivity becomes very low indeed. In food science, particularly interesting is  $^{23}\text{Na}$  and  $^{31}\text{P}$ , which can be measured via dedicated or multinuclear broadband NMR probes. The isotope 23 of sodium is the natural abundant, stable isotope, and with a fairly high magnetogyric ratio, the relative receptivity is approximately 9, which defines a medium sensitivity nucleus (where  $^1\text{H}$  nuclei have a relative receptivity of 100). However,  $^{23}\text{Na}$  is a spin  $3/2$  nucleus and is therefore quadrupolar. As a result, the signal width increases with asymmetry of the environment. Though, the signal from sodium in NaCl is narrow, with line widths of app. 10 Hz.  $^{23}\text{Na}$  NMR spectroscopy has been used in food studies primarily to characterize salted, brined or cured meat products [120-122] and fish [123], pickled vegetables [124] and to study the saltiness in soups [125]. Phosphorus 31 is 100 % natural abundant, the relative receptivity is approximately 7 and being a spin  $1/2$  nucleus, the NMR spectra are easy to interpret.  $^{31}\text{P}$  NMR spectroscopy has been used in the determination of the composition of olive oils [126], in the study of egg-white proteins [127], in the study of phytic acid degradation in cereals and bread [128] and to the post mortem metabolism in porcine meat [53].

#### **Two dimensional NMR**

Besides being able to detect multiple nuclei in liquids and solids, NMR spectroscopy also has the potential of measuring two (or more) nuclei at a time. The first dimension represents the traditional measurement of chemical shifts and couplings, but the second dimension adds an additional frequency axis, a concept first introduced by Jeener at a lecture in 1971. By introducing a second frequency dimension, the magnetic interactions between nuclei through structural connectivity, spatial proximity or kinetic

interchange are measured [129]. The most common two dimensional (2D) NMR experiments in food science include homonuclear correlation spectroscopy (COSY), which measures direct scalar coupled protons, homonuclear total correlation spectroscopy (TOCSY), which measures long range successive scalar coupling and heteronuclear experiments (e.g. heteronuclear single quantum coherence (HSQC)), which correlates  $^1\text{H}$  and another directly attached NMR active nucleus (in NMR food studies most often  $^{13}\text{C}$ ). By obtaining information about coupled and correlated nuclei, chemical structures can be elucidated. 2D NMR spectroscopy has been used in Paper II, Paper III and Paper IV.

### **SNIF NMR**

Since site-specific natural isotope fractionation NMR (SNIF-NMR) was introduced already in the 1980s, it has become a highly used analytical method in food applications [130]. The relative deuterium concentration and specific deuterium-site locations in a molecule can be determined using SNIF-NMR, and this can provide information about the chemical pathway of formation and, in some cases, information about the geographic origin. This is particular of interest for authentication and certification. SNIF-NMR has been applied to the analysis of wine [131], fats and oils [132], fishes [133,134], milk and dairy products [135], coffee [136], syrup [137] and honey [138], all products which have been authenticated.

To sum up, the highly informative multivariate spectra is, due to the quantitative nature of NMR spectroscopy, well suited for analysis using multivariate data analysis which is described in the following chapter.



### 3. Chemometrics

#### *Multivariate data analysis*

*Too much data - too little information! [139]*

Data is not the same as information. In fact, the more data we have, the less information we may have. Science demands often one-to-one relationships between a cause and the effect, and this is most conveniently investigated using simple univariate data analysis. Univariate causality has obviously worked well, but is now a hindrance to the study of more complex systems characterised by hidden many-to-many relationships. New methods of analysis are generating megavariable data sets which require new evaluation methods. Thousands of different, potentially important properties can be measured at the push of a computer button. To find new relationships between multiple blocks of such data structures requires new ambitious exploration, pattern recognition and data mining strategies and methods.

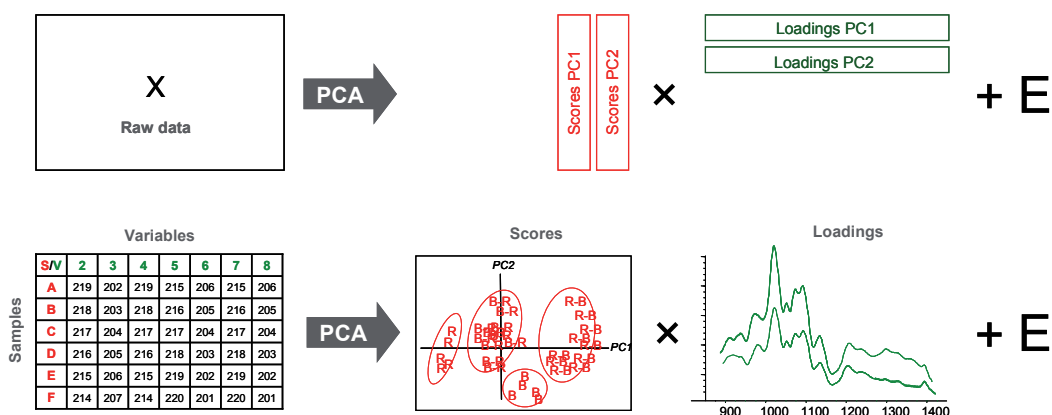
Multivariate data analysis is based on the extraction of latent components or underlying common structures in the data. Exploitation of complex NMR spectra of biological materials require that the quantitative capabilities of NMR spectroscopy to be fully exploited. Therefore it has become more and more common to combine investigations of complex NMR spectra with advanced multivariate modelling (chemometrics). The complexity of NMR data makes it of prime importance to utilize data reduction techniques in order to access the latent chemical information in the data. Chemometrics provides a powerful toolbox for bringing quantitative mathematical approaches to complex NMR data. The purpose of many NMR metabonomics studies is to investigate the data for class information (origin of material, effect/response of treatment). Unsupervised methods such as pattern recognition methods are employed to reduce the complexity and to visualize the classification in two-dimensional graphical representations. In complicated classification studies, supervised classification methods such as discriminant PLS and ECVA may also become useful. Another purpose of

NMR studies may be the need for quantifying a component or a pattern of components for the prediction of a given response or functionality. For this purpose, supervised calibration methods are required.

## Unsupervised data exploration

### PCA

Principal Component Analysis (PCA) [140] is *the* fundamental method and most commonly applied unsupervised chemometric method. In PCA the data matrix, for example composed of a set of NMR spectra, is resolved into principal components, PC's. The first principal component is defined by the spectral profile (loading) in the data which describes most of the variation and the second PC is the profile describing the second most of the variation orthogonal to the first etc. The PC's are composed of so-called scores and loadings. Loadings contain information about the spectral variables (chemical shifts) while the scores hold information on the amount or importance (pseudo-concentrations) of the loading vector in the sample set. Variation in the data which is not explained by the model is described in the residuals (E). (Figure 9)



**Figure 9** The data matrix X is decomposed into scores and loadings and the residual (E)

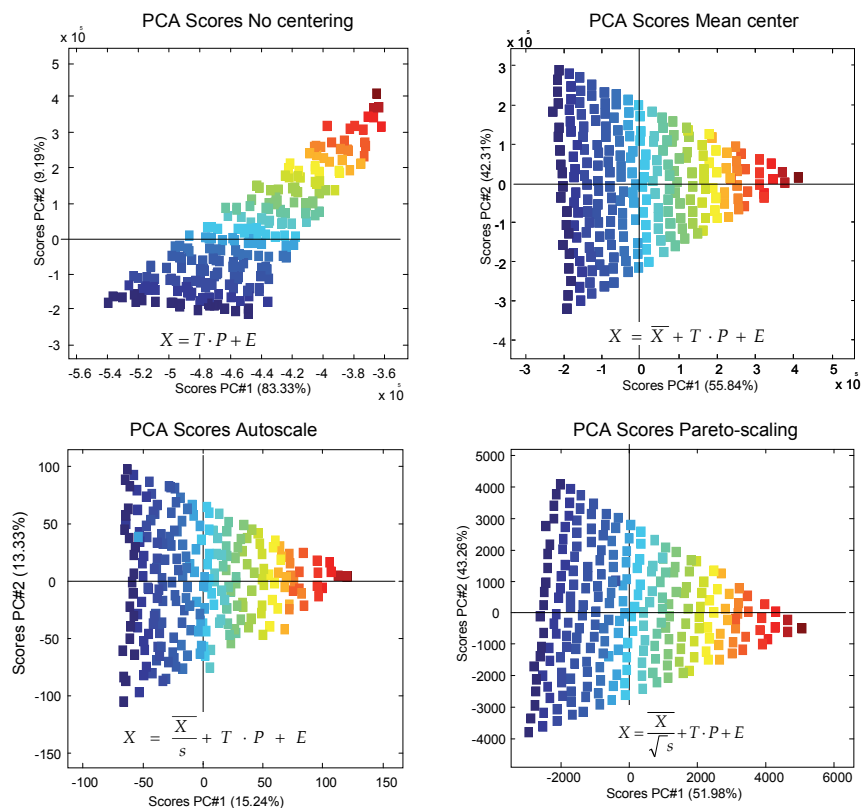
The result of the PCA model is represented as a score plot and a loading plot. The scores are plotted against each other in a scatter plot giving a 'map' of all the samples. The score plot is often displayed as a function of the two first principal components, as they explain most of the variance in the data.



The loadings are for spectroscopic data normally plotted against the unit of the measurement method, i.e. the ppm axis for NMR data. The loadings indicate which parts of the spectrum that represent the main variation among the samples. The scores can be considered as concentrations of multivariate, so called *latent*, variables (LV's). For a given principal component, the loading vector is a spectral profile and the score for each sample is the amount of that particular loading in the sample in a least squares sense. Thus the loadings weighted by the sample score values will provide an approximation of the spectrum of that sample. In PCA, the loadings will not resemble real NMR spectra due to the orthogonality constraints imposed, but the peaks in the loadings remain indicative of large spectral variations.

As the number of spectral components in a dataset is typically much lower than the number of chemical shifts, the whole data set can mostly be represented by few (typically less than 10) components that represent all the chemical variation in the data. The strength of PCA is to provide a quick unsupervised view of the samples in the score plot. It is thereby possible to identify samples that exhibit deviating features (outliers) or discover trends and groups in the samples. Samples that are close in the score plot are spectrally similar with respect to the variation shown and samples that are apart are different.

It is common to carry out simple data transformations prior to the PCA modelling. The minimal step is to mean centre the data, i.e. the mean spectrum is subtracted from the individual sample spectra. This simple pre-transformation ensure that the first principal component describes the variation between the samples rather than the direction of the overall variance. Mean centring is therefore used to focus on the varying part of the data and leaves only relevant information for analysis [78]. By mean centring the common feature of the samples are removed and as a result the PCA score plot is centred on origo and will map the covariance between the samples. However, the 'NMR spectral' loadings will become more difficult to interpret after mean centring because they will resemble difference spectra rather than real spectra. Figure 10 shows four score plots of the same sample set with different pre-transformations.



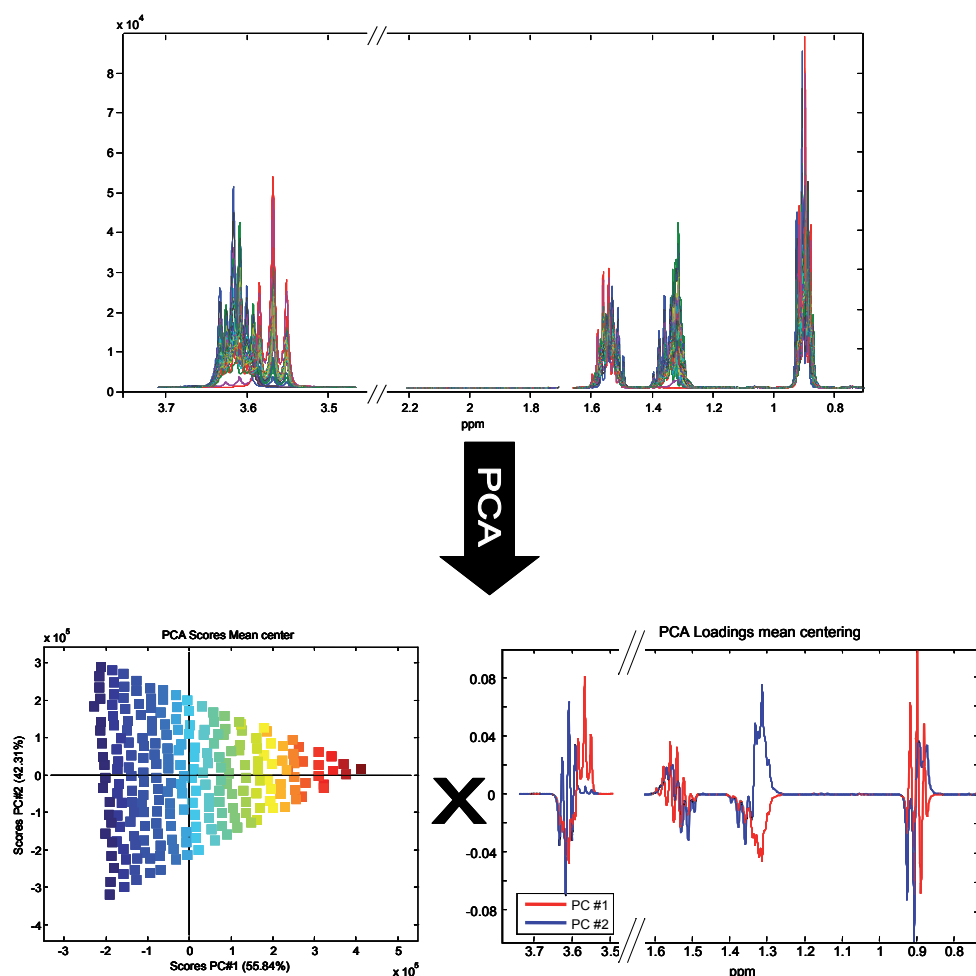
**Figure 10** Score plots of raw (A), mean centred (B), autoscaled (C) and pareto-scaled (D) PCA models of an NMR spectral ensemble (data from Paper I).  $X$  is the data matrix,  $T$  and  $P$  is the score and loading matrix respectively.  $E$  is the residual and  $s$  is the standard deviation.

Where the mean centred PCA model the covariance among the samples, autoscaled PCA models the correlations amongst the samples. Autoscaling is a more extreme transformation technique where all variables are divided with the standard variation and the data are scaled to the same unit. Autoscaling is essential when dealing with variables with different units. Pareto scaling [80] is very similar to autoscaling. However, instead of the standard deviation, the square root of the standard deviation is used as the scaling factor. Large fold changes are thereby decreased more than small fold changes, thus the large fold changes will be less dominant in the data. These transformations are normally applied as the last step after possible pre-processing methods. Since NMR data show very different signal

intensities, when modelling the data, small signals might be overshadowed by larger signals unless scaling is used.

Example:

A mean centred PCA model performed on 231 NMR spectra of alcohol mixtures display perfect recovery on the experimental design (Figure 11).



**Figure 11** top, NMR spectra of a ternary experimental design of 231 propanol, butanol and pentanol mixtures. Bottom, scores and loadings plot of the first two principal components from a PCA model calculated on mean-centred NMR spectra. The score plot is coloured according to the propanol content (red 0% propanol, blue = 100% propanol). The first two principal components explain together 97.8% of the total spectral variation. (Data from Paper I)

As obvious in the example, NMR data often have large regions containing no chemical information (baseline) and sometimes it can be beneficial to calculate models only on the information rich regions of the spectra. This can for instance be done by looking at the loading and choosing the most important spectral regions. Several interval methods exist which calculate models on sub intervals of the spectra. These methods will be further discussed in the following sections.

### **MCR**

An alternative method to PCA is to decompose the data matrix by Multivariate Curve Resolution (MCR) using Alternating Least Squares (ALS) [141]. PCA is designed for efficient and robust data exploration and classification for which reason loadings and scores are constructed to be orthogonal. So in general, it is not possible to obtain direct estimates of pure component NMR spectra and concentration. MCR-ALS, on the other hand, can resolve the spectral data into the 'true' underlying components, i.e. the pure spectra. Instead of abstract orthogonal loadings, MCR can ideally provide loadings that do in fact estimate the real spectra. When this is the case, it also follows that the scores in the MCR model will then be the corresponding relative concentrations. The drawback of MCR is that the solution is not unique. In most cases, a number of equally well-fitting solutions can be found. For this reason, it is often necessary to presume that the concentrations as well as the spectra are non-negative. Such a constraint can help eliminating the so-called rotational ambiguity [142] and provide the sought uniqueness. Unfortunately, even with non-negativity imposed, uniqueness is not guaranteed. Huo *et al.* (2006) have proved that MCR-NLR (non linear least squares regression), a multivariate curve resolution method was able to obtain unique pure spectra and pure decay profiles from diffusion-ordered spectroscopy (DOSY) NMR data [143]. The MCR equation is similar to the PCA equation, but instead of scores and loadings, MCR resolve the data matrix in concentrations (C) and spectra (S) plus the residual (E).

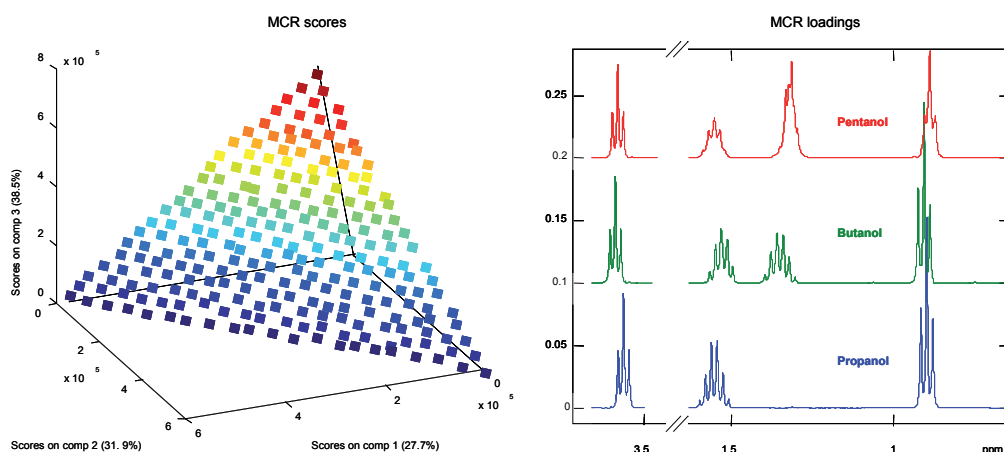
MCR

Data matrix **X** is the product of  
concentration profiles **C** and  
pure component spectra **S**

$$\mathbf{X} = \mathbf{C} \cdot \mathbf{S} + \mathbf{E}$$

Example:

MCR has proven very useful in resolution of overlapped NMR spectra of a mixture of three similar alcohols (Figure 11 top). A three component MCR model was able to extract the pure alcohol spectra and their concentration profiles (Figure 12).



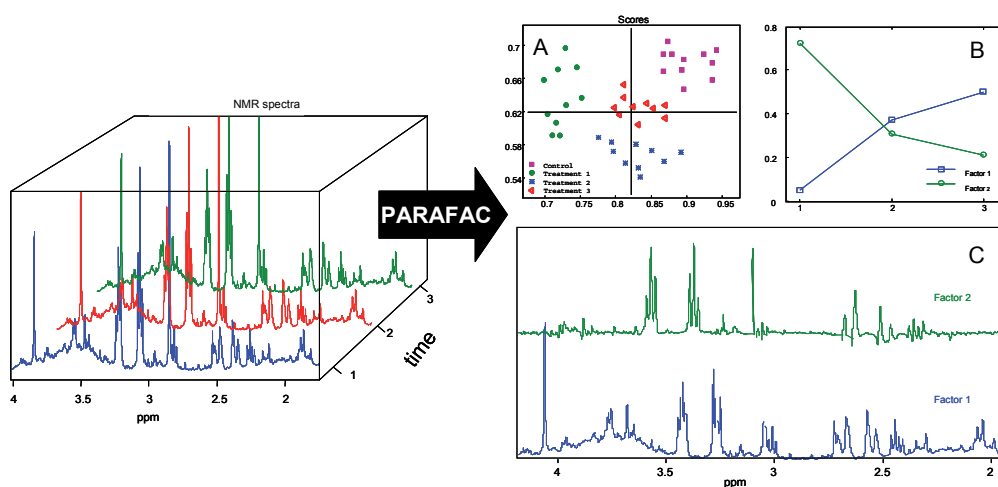
**Figure 12** Scores of the first three components (left) and loading plot (right) of the three pure components from the MCR model, obtained on the NMR spectra. The first two components explain together 98.1% of the variation. The loading plot shows pure profiles of the three alcohols. Data from Paper 1

For obvious reason, mean centring and autoscaling is not used in MCR studies.

## PARAFAC

In exploratory studies investigating changes over time for a set of samples, the data are naturally arranged as a three-way data set. The first dimension represents the samples, the second dimension represents the time and the third dimension represents the measured profile, for example NMR spectra (Figure 13). Parallel factor analysis (PARAFAC) [144-146] can be considered

as a multiway extension of MCR able to handle three-way and higher order data. The PARAFAC model is based on the decomposing of the data into trilinear components in a similar way to the bilinear components extracted in MCR. When such higher-order data are available, the so-called second-order data advantage provides a unique solution and e.g. the pure analyte spectra will be found in a mixture [146]. Dyrby et. al (2005) proved that multi-way curve resolution by PARAFAC applied to 2D diffusion edited  $^1\text{H}$  NMR spectra was a valuable tool for analyzing lipoprotein main fractions in human plasma samples [147].



**Figure 13** A schematic PARACAC model showing time development in NMR spectra yielding scores (A) and two sets of loadings; a time development loading (B) and an NMR spectral loading (C)

## *Supervised data exploration*

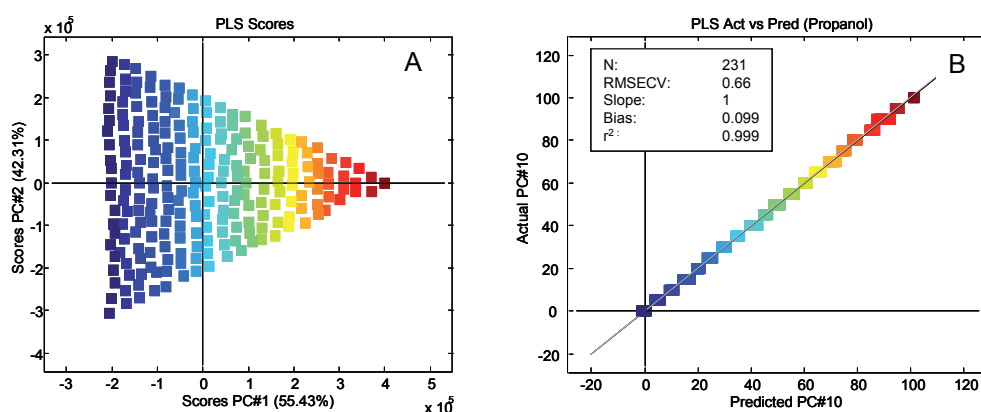
### **PLS**

If reference data are available it is possible to develop calibration models between the NMR spectra and a given response variable. This can be done for example with the regression method called partial least squares regression (PLS) [148]. PLS regression is the second basic algorithm of chemometrics. In PLS, the bilinear data matrix (X) is resolved in to linear components (latent variables) just as in PCA. However, in PLS focus is not only to describe the variation between the samples, but also to emphasize variance in data which co-vary (or is correlated when autoscaled) with the response variable (y). Thus, PLS regression is a supervised method which

may be utilized to develop prediction models that can replace the reference method by a much faster and perhaps a more precise and accurate NMR method. PLS works quite similar to PCA, but its scope is to regress (or force) the result in a given direction (reference method) and is thus called a supervised method.

Example:

A PLS model obtained on 231 NMR spectra of alcohol mixtures shows perfect predicting model for propanol content with very high correlation coefficient ( $r^2$ ) (0.999) and low root mean square error of cross validation (RMSECV) of 0.66 % (range 0-100%) in Figure 14. The validation is further described in the paragraph “Validation”.



**Figure 14** PLS score plot (A) of a model calculated on mean-centred NMR spectra. The score plot is coloured according to the propanol content (red = 100 % propanol and blue = 0 % propanol). The first two PCs represent 99.77% of the variation (B). Predicted versus actual/measured plot of the PLS model on propanol which has a prediction error of only 0.66% and a correlation coefficient of 0.999 (right)

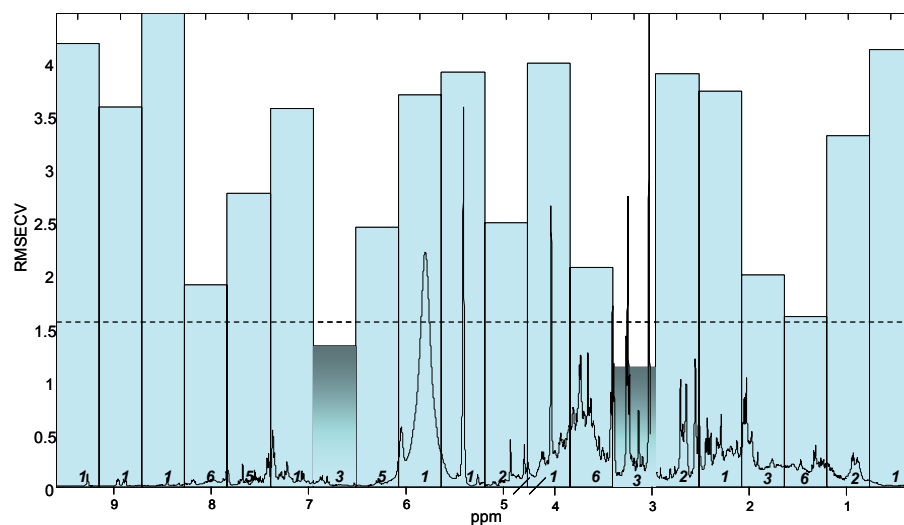
As previously mentioned, NMR spectra often show regions which lack chemical information. These regions only containing noise can disturb the PLS model at best. At worst, spurious correlation can be observed especially when using severe pre-transformation such as autoscaling. An example of this will be shown later in this chapter.

## iPLS

Assignment of an NMR spectrum of a biological sample is highly complex as resonances from all protons give rise to signals in the NMR spectrum. To

investigate the influential areas of the spectra, the region selection method interval PLS (iPLS) [149] has proven very efficient. iPLS is an extension of PLS which develops local PLS models on a number of subintervals of the full spectrum region. The main advantage of iPLS is that it provides an overall picture of the relevant information in different spectral subdivisions, thereby removing interferences from other regions. iPLS reveals areas of the spectra which hold information about the reference  $y$  and is thus useful for interpretations and assignment.

iPLS is particular useful for NMR spectra because of the complex nature of the spectra. NMR spectra may hold information about hundreds of different compounds and the background of larger peaks will dominate the chemometric modelling as the algorithms normally search for largest variation in the spectra. Dividing spectra into subintervals will remove interferences from the larger more varying signals. Mean centring is reasonable to choose as minimum pre-processing in iPLS. For reasons mentioned earlier, caution should be taken when using autoscaling as pre-processing of the spectra.



**Figure 15** iPLS plot of the prediction of onion intake from NMR spectra of rat urine. Dotted line is RMSECV for the global model (7 LV's) which is 1.55 %. Italic numbers are optimal LV's in interval model. Two intervals can markedly improve the calibration model. (Data from Paper III)



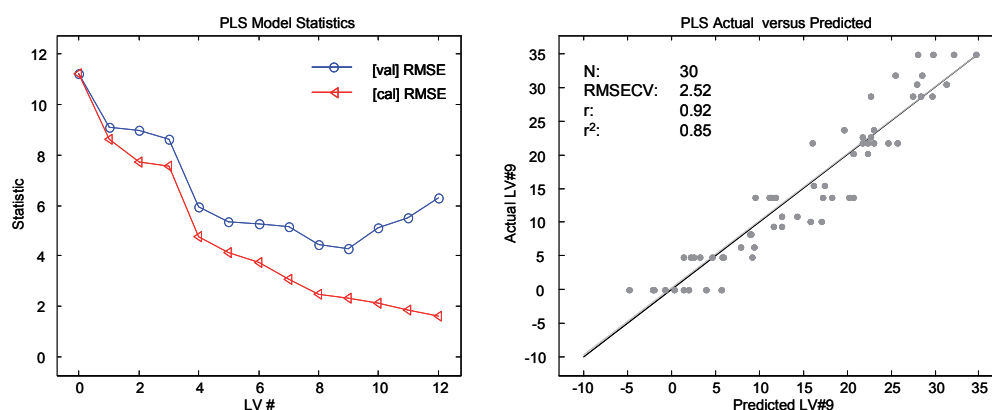
Example:

A mean centred iPLS model of 32 NMR spectra of urine from rats on an onion diet reveals relevant intervals with respect to onion biomarker in urine (Figure 15).

A more evolved application of iPLS is backwards iPLS [150] where the least significant interval is left out, one at a time. Then the model is then recalculated yielding a predicting error which is compared to a model where the second least significant interval is left out. By this procedure, intervals are excluded and the model ends up only with the remaining most significant intervals. Adding this sophisticated level to iPLS increases the danger of overfitting and thus increases the demand for efficient validation methods.

## Validation

The number of LV's in a PLS model can be determined using cross-validation where the sample set is divided into a number of segments which in turn are excluded "one at a time" before re-entering into the model in order to estimate the prediction error, root mean square error of cross validation (RMSECV). The models statistics show the development of the RMSECV as a function of the number of LV's. The optimal number of LV's is chosen from the first minimum of the RMSECV curve (Figure 16).



**Figure 16** Plot showing root mean square error of validation and calibration (A) and actual versus predicted plot showing model performance (B)

The RMSECV is then used to determine the number of components used in the prediction model. The predicted versus measured plot is generally used to evaluate the model performance.

Validation of PLS-based classification models is crucial especially in metabonomic applications. Calibration models performed on metabonomic data are prone to serious modelling and validation problems [151-153]. These problems will be discussed in the following sections.

### ***Classification methods***

The goal of classification is to assign new objects to the class to which they show the largest similarity. So an object belonging to the same class show a particular class pattern. As mentioned in the previous paragraph, the most simple and basic classification method is PCA. Visualizing data by PCA in a two-dimensional map of the sample, hidden relationships or grouping of the samples might be revealed. However, often PCA modelling is not sufficient to reveal grouping.

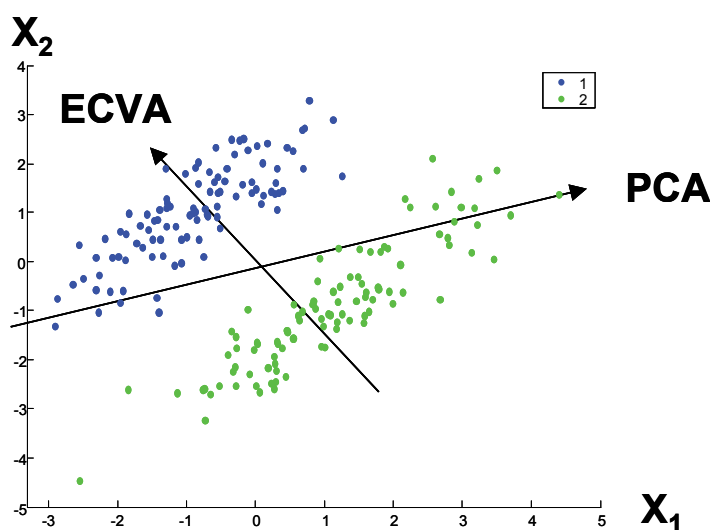
#### **SIMCA**

Soft independent modelling of class analogy (SIMCA) is an unsupervised classification method based on PCA model residuals and hotellings. Samples belonging to one group are modelled by PCA, and new samples are compared to the class models and assigned to classes, according to their analogy to the training samples. An example is NMR measurements coupled with SIMCA which offered a powerful mixture-analysis tool classification of 176 kinds of green, black, oolong and other tea infusions. The result was clear classification reflecting the fermentation and processing of each tea, and revealed marker variables that include catechin and theanine peaks [154].

#### **ECVA and iECVA**

Extended canonical variates (ECVA) [155] is a relatively new classification tool. Canonical Variates Analysis (CVA) [156,157] is a method for estimation of directions in space that maximizes the differences between the groups. However, CVA cannot deal with highly collinear data such as spectroscopic data where the number of variables is larger than the number of samples. The ECVA method is based on the standard CVA, and by a transformation of an eigenvector problem to a regression problem, it is possible to use PLS

to solve the inner part of CVA and thereby allowing for the analysis of collinear data. ECVA finds the vector which maximize 'the between class scatter' over the 'within class scatter' [155]. Interval ECVA (iECVA) can analogous to iPLS be an efficient tool to investigate significant areas of the spectra able to classify the samples (Figure 17).



**Figure 17** From PCA to ECVA, maximizing the distance between the groups. Modified from [155]

### PLS-DA

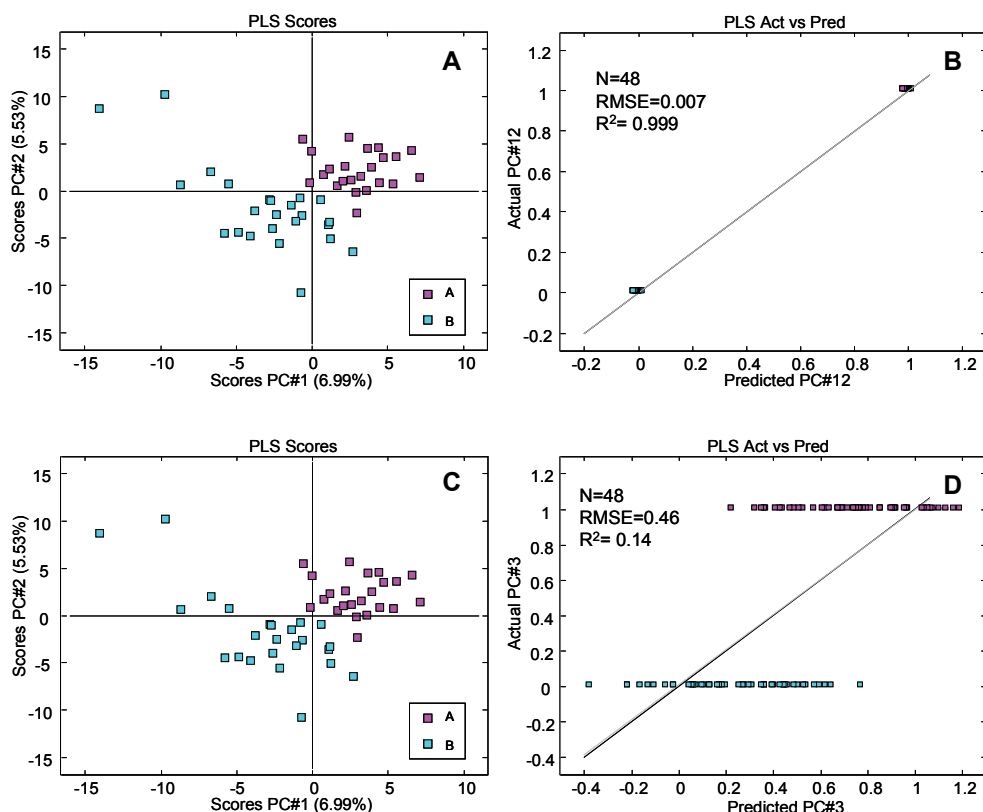
Discriminant PLS (PLS-DA) is a common choice of multivariate data analysis for classification. PLS-DA is a regression model describing maximum separation between two pre-defined classes, but which is more focused on the actual class discriminating variation in the data compared to the unsupervised approaches. The predefined groups are incorporated in the model where a 'dummy vector' is constructed containing zeros and ones which hold information about the sample-groups. Interpretation of the PLS-DA regression coefficients generates information pertaining to the explanation of class differences based on the fact that each variable coefficient is related to a certain NMR spectral region associated with a specific molecular structure. However, PLS-DA is seriously prone to overfit. Westerhuis (2008) states that classifying groups of individuals based on their

metabolic profile using the PLS-DA method eagerly overfits the data due to the low number of samples compared to the large number of variables [158]. This is a major problem in NMR spectroscopy, as the number of variables easily is around 65,000 whereas the number of samples is much smaller. The result of overfitting can be serious misleading, with the added difficulty that the problem is not always obvious. Therefore, the strictest possible validation of chemometric models is crucial in order to obtain correct interpretations.

The result of a validation is easily seen in the actual versus predicted plot which holds information on the model performance. However, it is extremely important to bear in mind that the consequence of the validation is *not* reflected in the score plot which will not be any different when applying validation. A general, but risky, procedure when developing PLS-DA classification models is to make all assumptions and interpretations from the score plots without checking the basic model performance. In the rapid developing metabonomics field it has become worrying common to show PLS-DA score plot as evidence of diversity between two groups when measuring biofluids. This is done when the PLS-DA score plot illustrate two groups without giving any information on model performance and behalf of this concluding effect of treatment/ gender/ diet [159-169]. However, as the score plot is not validated, the result looks too optimistic and serious misinterpretations can be made. The subsequent interpretation of the corresponding loading plot (or regression coefficients) will in biological research almost always make some kind of sense as metabolic pathways are enormous complex for most biological materials.

An example of the danger of false grouping by PLS-DA can be made by a simple test. A matrix consisting of 50 times 65,000 random variables with random groups is modelled by PLS-DA. The score plot of this model (not shown) will give two nicely separated groups. However, when checking the model performance it becomes obvious that no correlation exist in the random data with respect to the random groups. Statistically there is a great chance that one variable out of 65,000 even random variables will show a chance correlation with any 'dummy' y-variable.

Another example of why PLS score plot should not stand alone in PLS-DA classification is shown in Figure 18



**Figure 18** PLS score plot and actual versus predicted plot of a non-validated calibration model (top A, B) and validated calibration model (bottom C,D) obtained on a non informative NMR spectral baseline

In this case, a PLS-DA model is calculated from an autoscaled noisy baseline of an NMR spectrum which holds no chemical information and data are distributed in random groups. There should not be any possibility that these baseline 'spectra' are able to distinguish between the two groups. Nevertheless, when examine only the score plot in Figure 18A, there is a chance to be lead to the conclusion that it is possible to classify the two groups by their spectra. And the non-validated actual versus predicted plot, Figure 18B, shows perfect correlation between the 'real' group value and the group value which is predicted by the PLS model.

However, when applying a simple cross-validation step where the sample set is divided into 4 data splits, leaving one segment out at a time and thereby obtaining an error for each data split the result gets quite different. The actual versus predicted plot for the same but *validated* PLS-DA model reveals that it is not possible to model the data properly Figure 18D. The plot shows no correlation which is a proof of a bad model performance. The prediction error is almost 0.5, the exact average of the two class values 0 and 1. This is clear evidence that there is no correlation between the noisy baseline and random groups. And most important, the score plot of the validated model, Figure 18C, is exactly the same as for the non-validated model, since score plots rarely show the validated (predicted) scores.

However, it is strictly important to notice that even with proper cross-validation; PLS-DA is still prone to overfit.

### ***Multivariate tricks of the trade***

In some cases the basic chemometric methods are not sufficient to model the data at hand. The standard chemometric tools need tailoring for NMR data since they contain shifting signals and baseline inhomogeneity (variations) that need to be addressed specifically. These 'artefacts' disturb the PCA model to an extent where the 'real' chemical information is not described in the first PC combinations. Kemsley *et al.* (2007) showed that neither PCA nor PLS are sensitive enough when the most relevant spectral information is concentrated in a few very small peaks or even a single peak. This is difficult when set against a background of many other larger peaks which vary in ways which is unrelated with the groups of interest [170].

### **Orthogonalization**

In order to examine the dependence on indirect correlations to Y, orthogonalization of the data matrix with respect to Y can be applied. Thus, the vector describing the variation due to Y can be extracted from the spectra prior to model development. In this way the predictions can be tested if they are due to an indirect correlation with another confounded property of Y [171].

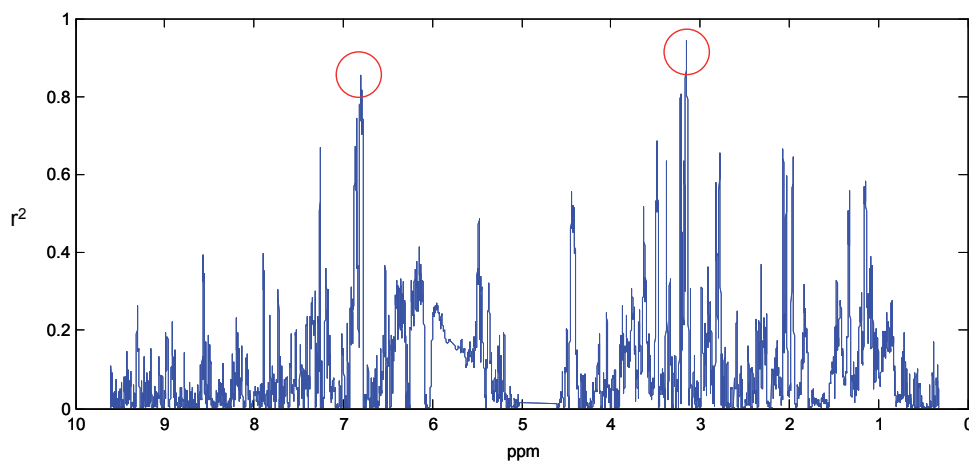
### **Forward selection**

Due to the nature of NMR spectra, the large baseline regions without chemical information and many non significant variables, variable selection

can become necessary. In forward selection, the variables are added to the model one at a time. At each step, each variable that is not already in the model is tested for inclusion in the model. Forward selection has drawbacks, including the fact that addition of new variables may render one or more of the already included variables non-significant. An alternate approach is backward selection where all the variables are included from the start. Then the least significant variable is dropped. These approaches are similar to iPLS and backwards iPLS, using intervals instead of variables.

### Correlation plot

When searching for influential areas of the NMR spectrum with respect to a response variable ( $y$ ), correlation plots can be carried out. The principle is the same as PLS which finds correlations but with another presentation of the data. Each variable is correlated to the response variable yielding a plot of all the correlations coefficients, see Figure 19. It is convenient to square the coefficients in order to obtain only positive values. Obviously, variables closest to one are best correlated with  $y$ . Using correlation plots; it is possible to obtain correlations between variables in order to check if other variables are correlated such as intramolecular proton signals.

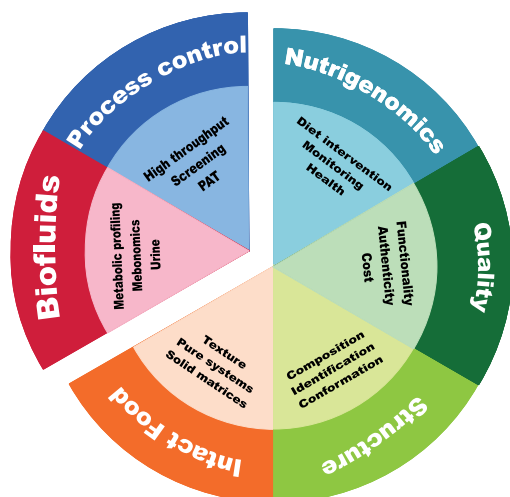


**Figure 19** Correlation plot between 32 NMR spectra of rats urine and onion dose, clearly pointing two variables out around 3 and 7 ppm

In conclusion, NMR spectroscopy combined with multivariate techniques has great potential in resolving many problems within health, food and

nutrition research. In the following, a number of significant applications will be highlighted.





## 4. NMR as a quantitative high-throughput technique

Nuclear magnetic resonance spectroscopy is a unique versatile analytical method which has been exploited in this Ph.D. study. Many, very different applications of NMR spectroscopy have been carried out which give an idea of the possibilities and limitations of NMR spectroscopy combined with chemometrics within food and metabonomic studies. All the applications presented can be united in one figure (Figure 1, Chapter 1), illustrating the different properties of NMR spectroscopy used throughout this thesis and the work can be illustrated as included in one or two 'pieces' of the pie. This representation will be used in the following four chapters concerning Paper I-IV. The ability of  $^1\text{H}$  NMR spectroscopy to perform chemical identification and quantification, to elucidate chemical connectivity, to measure intact matrices, to study of biofluids, in process control and in nutritional diet intervention studies will be examined in the following four chapters. In the following chapter, advanced chemometric modelling for exploring complex NMR data is introduced and the results from Paper I is presented.

Due to its non-destructive nature, NMR spectroscopy can be used to analyse the composition of metabolites under or close to physiological conditions [172,173]. NMR spectroscopy has many advantages including relatively fast high-throughput sampling. High field  $^1\text{H}$  NMR spectra of food or biofluids typically contain several thousand resolvable signals, potentially providing

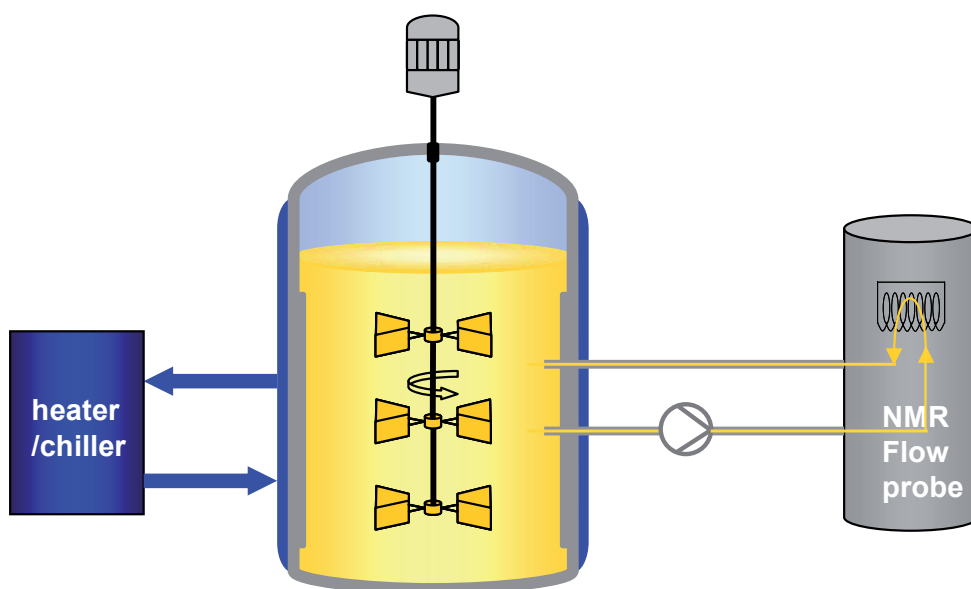
structural and quantitative information on hundreds of compounds from a single, non-destructive measurement in a few minutes [174]. The combination of NMR spectroscopy and chemometrics has tremendous potential in the nutrigenomics and metabonomics fields for exploring patterns of biomarkers in biofluids for diseases and food intake, but a future in process analytical technology (PAT) in the food and medico industries is also highly possible.

### **Process analytical technology**

PAT is a US Food and Drug Administration (FDA) initiative to improve drug substance manufacturing. The definition of PAT is: A system designed for analyzing and controlling manufacturing through timely measurements (i.e. during processing) of critical quality and performance attributes of raw and in-process material and processes with the goal of ensuring final product quality [175]. The quality control part of the process is to supply quantitative and qualitative information about the process and to optimize the efficient use of energy, time and raw materials. In this way, it is possible to detect early changes in quality and regulate the process. Quality control should be built-in using on-, in-, and/or at-line measurements and thereby reducing production time. This will prevent rejects and re-processing, and reduce human errors. In order to achieve this, FDA has recognized four categories of tools; 1) Multivariate tools for design, data acquisition and analysis, 2) Process analyzers, 3) Process control tools and 4) Continuous improvement and knowledge management tools [175]. The analytical tools have largely relied on vibrational spectroscopy and chromatographic methods despite difficulties with peak assignment and reliable quantification. However, NMR spectroscopy integrates well with the establish procedures and can provide useful technology for the PAT challenge [176]. In the capacity of being quantitative, as well as qualitative, NMR spectroscopy provides the fundamental requirements for process control. NMR spectroscopy can monitor reaction progress because it affords accurate concentration of all major compounds and NMR spectroscopy can be used to directly determine the structure of reaction intermediates and impurities.

Furthermore, NMR spectroscopy is perfectly suitable for multivariate data analysis to fingerprint process streams and raw materials. However, the delicacy of the NMR instruments provides a great challenge before implementing in a process. As already mentioned, the NMR spectrometer is

highly sensitive to outer magnetic fields. High resolution NMR spectroscopy is therefore not well suited for inline monitoring; only at-line or on-line monitoring of diverted process streams is practical. A solution for making on-line measurements possible is a chemical reactor with a sampling device where a pump circulates the mixture through an NMR flow cell. The flow is then stopped and the measurement takes place (Figure 20). However, this is only possibility for measuring liquid samples with a low viscosity, including food such as juice, beer and wine etc. Apart from relatively minor off-line quality control applications it has so far failed to be developed as a sensor in the industrial sector [177].



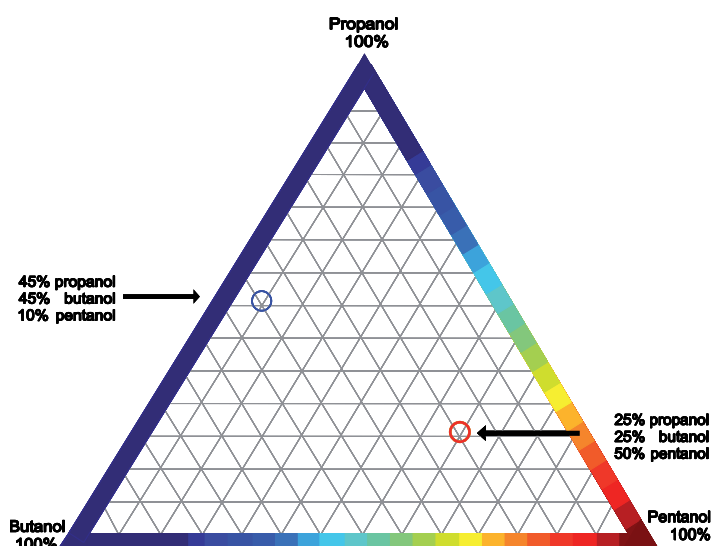
**Figure 20** Sketch of an apparatus for sampling chemical reactions from a chemical reactor and recording NMR spectra. Modified from [176]

### Visualization of NMR data

The greatest advantage of chemometrics is the simplicity by which even large data structures are analysed and visualized which thereby adds an exploratory dimension to modern NMR science. The signal intensity is proportional to the concentration (obey Beer's law) which is a necessary assumption for the use of basic bilinear chemometric. Applying quantitative chemometric methods such as PCA and PLS to multivariate high resolution

NMR data from mixtures will give an idea of how it performs on more complex data, such as metabonomic data.

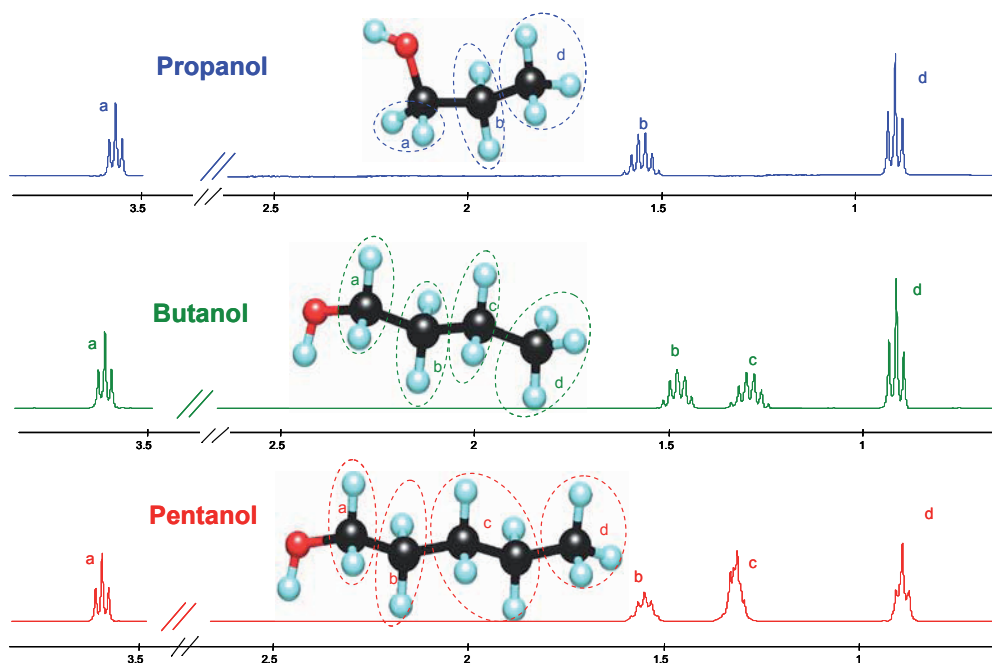
All biological material consists of complex mixtures of hundreds of compounds. Therefore, it would be highly desirable to be able to extract pure spectra of the components in biofluids, such as urine. In order to establish this, advanced state-of-the-art chemometric is introduced (Paper I). MCR is a relatively new chemometric method for analysing NMR spectra, commonly known from ordinary spectroscopic methods, but which can provide radically more information than current approaches. MCR is applied to a ternary experimental design in order to show how simple even large data structures can be analysed in its exploratory nature (Figure 21).



**Figure 21** Ternary experimental design of a sample set consisting of 231 alcohol mixtures of the three alcohols propanol, butanol and pentanol each one in concentration between 0-100 percent. The triangle is coloured according to pentanol. (Data from Paper I)

Figure 22 shows how close the selected three alcohols are in their chemical structure visualized in the  $^1\text{H}$  NMR spectrum of the pure compounds. The pure NMR spectrum of each alcohol shows that the three alcohols have signals at almost exactly the same positions. Quantification is only possible in case the signals are well separated. In contrast to chromatographic

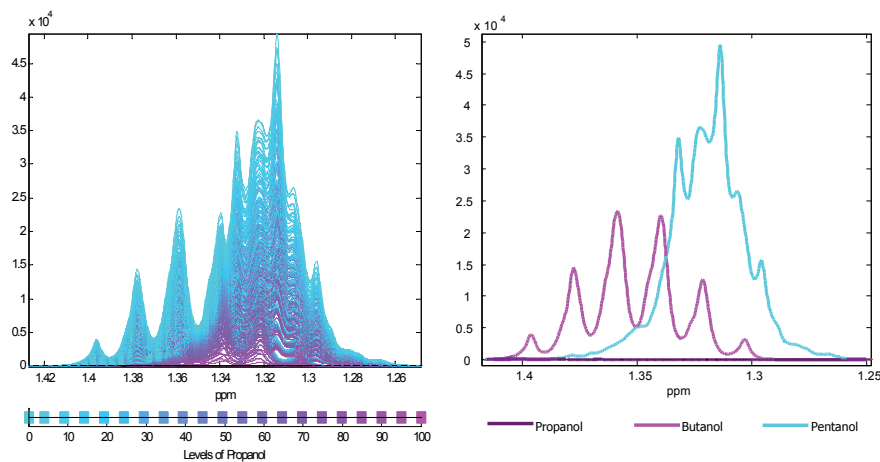
methods, there is only little possibility to influence separation of the signals because the chemical shift is directly related to the molecular structure.



**Figure 22** Chemical structure and  $^1\text{H}$  NMR spectra of propanol, butanol and pentanol. The spectral window is reduced to only including areas with signals. The OH proton does not contribute to the spectra due to exchange with deuterium from the  $\text{D}_2\text{O}$  added. (Data from Paper I)

### Signal overlapping

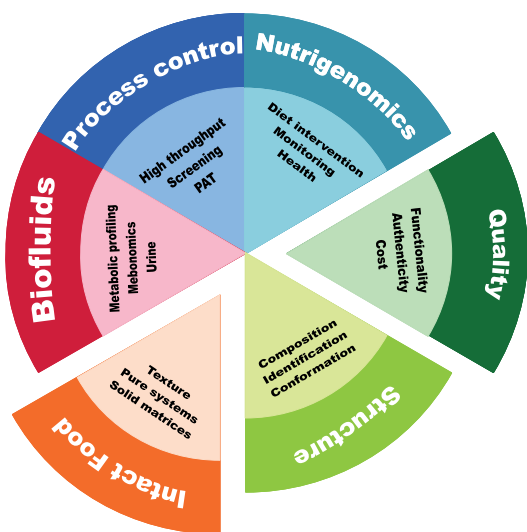
Because the chemical shift range for hydrogen is rather limited, signal overlaps are common when analyzing biological samples. PCA and MCR can be useful in the investigation of highly overlapping data from NMR studies. MCR provides loadings that resemble the pure spectra of compounds and scores which represent the corresponding concentrations. The powerful ability of MCR to separate mixtures into pure spectra and concentrations from a ternary experimental design of  $^1\text{H}$  NMR spectra of alcohol mixtures is shown in Figure 23. This may be considered a model example of how chemometrics can be used to visualise causal relationships among samples (Paper I).



**Figure 23** Heavily overlapped CH<sub>2</sub> signals of alcohol mixtures of propanol, butanol and pentanol (left) and three loadings extracted by MCR resembling pure NMR spectra of the CH<sub>2</sub> signal from propanol (at zero level), butanol and pentanol (right).

The MCR solution is not unique but by repeating the estimation of the MCR model many times from different random starting, it may be verified if the same fit and solution is obtained (results not shown). Hence, the solution can be assumed to be unique. These results are encouraging and imply that NMR spectra of complex mixture can be separated mathematically into the underlying constituents. However, the main reason that the results are as good as they are is the presence of pure samples in the model sample set. The presence of pure samples adds selectivity in the data. Selectivity means samples or variables for which only one analyte is present. This is one of the key prerequisites for obtaining uniqueness in MCR models.

Compared to PCA, the information obtained by MCR models is far more detailed. The loading from PCA are due to orthogonality constraint difficult to interpret and the scores can only be used visual for detection of grouping and trends in the data. However, The MCR scores can be compared with the 'true' value (i.e. the concentration of the three alcohols) by plotting them against each other. This can act as a control of the preciseness of the model. In this case the three correlation coefficients of the three concentration profiles was higher than 0.99 indicating an almost perfect MCR model.



## 5. NMR in the study of metabolites in intact food matrices

The ability of NMR spectroscopy to measuring solid food material makes it a superior analytical method for determining quality and functionality of intact food systems. This chapter concerns mainly Paper II and some additional publications.

### *Solid state NMR spectroscopy*

High resolution solid state NMR spectroscopy was developed in the 1970s but it is not until the recent years that HR-MAS NMR spectroscopy has been recognized as an efficient analytical method for analyzing food systems. Though crystalline compounds have to be measured by CP-MAS NMR spectroscopy (e.g. starch below gelatinization temperature (see below)), foods containing proteins and fats are normally difficult to crystallize and therefore give poorly resolved solid state (CP-MAS) spectra. However, HR-MAS NMR spectroscopy can overcome these problems in semi solids and is, due to higher sensitivity of  $^1\text{H}$  compared to  $^{13}\text{C}$ , much faster. Using HR-MAS NMR spectroscopy it is possible to obtain the many advances of measuring the intact food matrix. Furthermore, the method is non-destructive leaving the sample intact for subsequent analysis. Only HR-MAS measurements are included in the work presented here.

### **HR-MAS NMR spectroscopy in the study of solid foods**

The numbers of successful food applications of HR-MAS NMR spectroscopy are many and wide spread from beef to flour and juice. For example, dried beef samples of certified origin have shown specific patterns of metabolites which are possible candidates for markers of origin, probably linked to feeding system [178]. As other examples it was found possible to differentiate between durum bread and flours coming from different geographical areas of southern Italy [90,179]. Gill and co-workers showed that HR-MAS NMR techniques enabled resolved NMR spectra of mango pulps [180]. Shintu and Caldarelli used HR-MAS NMR spectroscopy for characterization of parmesan cheese allowing identification of the presence of fatty acids (saturated and unsaturated), amino acids and other small organic molecules which have previously been shown to correlate with organoleptic origin and age characterization [91,181]. They also succeeded in grouping emmental cheese samples according to their geographical origins [92]. Consonni and Cagliani (2008) also studied ripening and geographical characterization of Parmigiano Reggiano cheese [182] and finally, Lamanna et al. (2008) investigated degradation of soft cheese under different packaging conditions [95]. These are all examples of authenticity characterization of different foods and the area is rapidly developing as farmers and producers increasingly need to protect against fraud and adulteration. As described in Chapter 2, recently the NMR manufacturer developed a dedicated instrument for authenticity and adulteration, the SNIF-NMR technique.

### **Wheat quality**

Measuring intact raw material such as single wheat kernel is a challenge for HR-MAS NMR spectroscopy. In wheat, proteins are the most important component with respect to end-use quality [183]. However, the quantity of protein per grain is mainly under environmental and genetic control. Extreme climate events are becoming recognised as an important factor in the effects on crop growth and yield. High temperature or drought affect the balance of protein fractions why it is important both to understand the environmental constraints on crop quality and to predict how the quality will change with environmental changes and with extreme climate events [184].

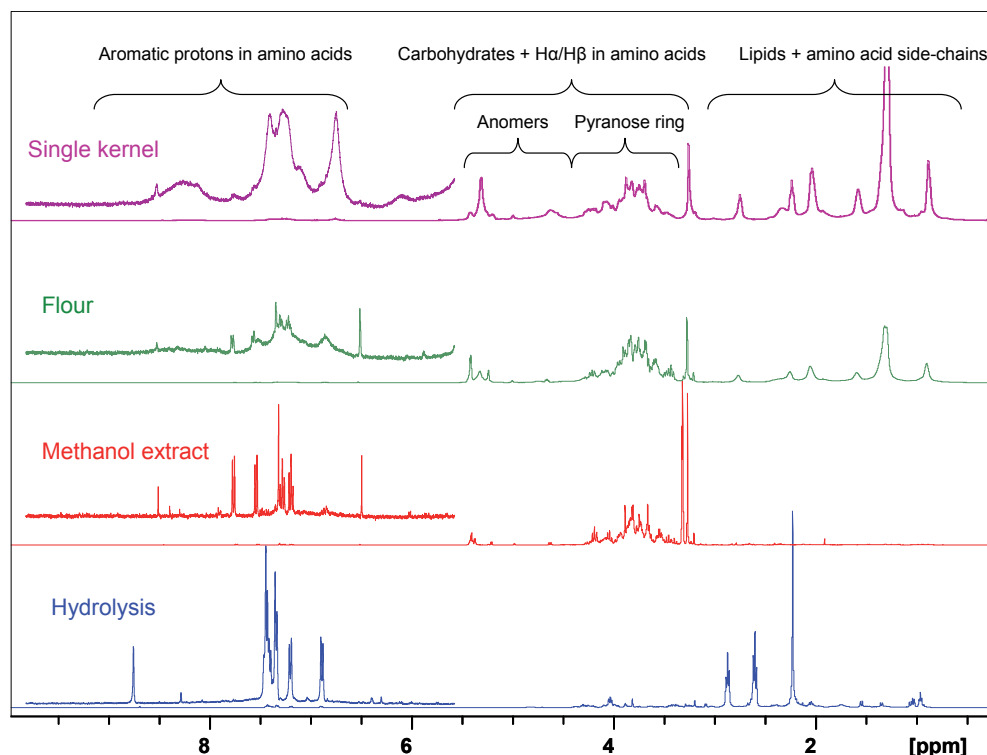


In the work concerning wheat quality presented in this thesis, the implications of one or more drought events on protein quality in developing wheat grains was investigated using information from chemical protein analysis, liquid state  $^1\text{H}$  NMR spectroscopy of methanol extracts, and  $^1\text{H}$  HR-MAS NMR spectroscopy of single kernels and flours followed by unsupervised exploratory chemometric data analysis. HR-MAS NMR spectroscopy showed the possibility of studying the effect of different drought events on the protein synthesis in grains of winter wheat. Furthermore, the  $^1\text{H}$  HR-MAS NMR spectra of single wheat kernels showed considerable differences between early and late harvest (Paper II).

### **Quantitative aspects of HR-MAS NMR spectroscopy**

The use of HR-MAS NMR spectroscopy on single cereal kernels provides unique possibilities in studies of cereals. Seefeldt *et al.* (2008) studied the temporal and genotypic differences in bulk carbohydrate accumulation in three barley genotypes during grain filling [185]. The barley flours were measured by HR-MAS NMR spectroscopy which gave well-resolved spectra with clear differences between times of harvest. However, they experienced a problem of measuring starch below the gelatinization temperature and fibres because immobile  $\beta$ -glucan and rigid starch remained immobilized and thus not visible in the spectra. The solid character of the single kernels due to water immobility makes it difficult to mobilize semi crystalline parts of the kernel. Especially in mature kernels, which have almost solid character, the rigid, semi crystalline matrix will lead to few broad signals. Only signals from protons that are mobile by themselves, protons from fat, protons that are naturally in contact with cellular water or protons that are accessible to the soaking water are visible in the  $^1\text{H}$  HR-MAS NMR spectra.

The presentation of the sample or the fraction of e.g. a wheat seed (single kernel, flour or extract) induces great variation as the HR-MAS NMR technique, not necessarily provides signals from all protons in the solid sample. Therefore, flour or to a greater extend extracts of the single kernels must be chosen in order to mobilize the semi crystalline parts of the kernels. Figure 24 shows the NMR spectra of four different fraction of wheat; whole single kernel, flour, methanol extracts of flour and sour hydrolysis of flour, all showing different signals. This is an example of how different the information to obtain when measuring different fractions.



**Figure 24**  $^1\text{H}$  NMR spectra of hydrolyzed wheat and wheat methanol extract together with  $^1\text{H}$  HR MAS NMR spectra of wheat flour and single kernel wheat normalised to similar noise level.

Furthermore, the preparation of the sample is also important because of the difficulty to obtain correct quantification due to varying sample size etc. The size of the kernel material studied and the relatively small dimension of the MAS sample holder (rotor) are not really compatible. HR-MAS NMR rotors have a fixed inner volume, whereas the sample size of a wheat kernel can vary markedly resulting in unequal concentrations between replicates. For quantitative purposes, it is thus very important to keep these factors as constant as possible.

The advantage of using flour or extracts instead of single kernels is that the variations in size of the kernel, the spike differentiation and the different position of the spike amongst kernels are averaged out. And as described, more protons are mobilized in flour and extract. Furthermore, the sample preparation of flour and extract is easier, as the rotor can be filled

completely. However, in this way the advantage of measuring intact biological material is of course lost.

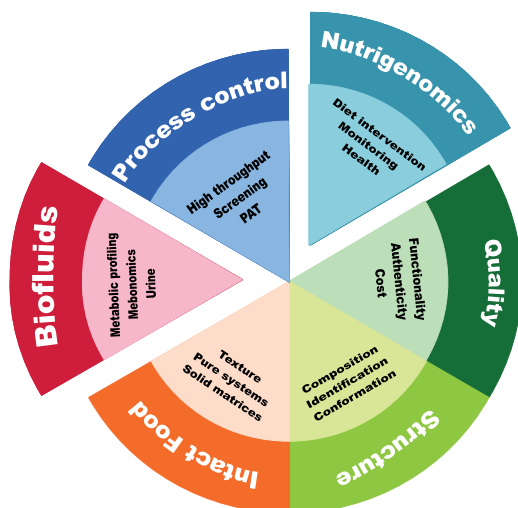
### **Considerations concerning experimental setup**

The experimental parameter setup in HR-MAS NMR spectroscopy is very critical. The spinning rate, solvent and temperature have to be optimised for the specific problem to be studied to ensure reproducible spectra that include all mobile compounds of interest. Two opposite considerations have to be taken into account with respect to the spinning rate. The spinning speed has to be sufficiently high to ensure high quality signals from all expected compounds and for preventing spinning sidebands in the spectra. But, if the material is permeable for solvent, the destructive forces of centrifugation point towards using a low spinning rate [186]. Another important parameter in relation to sample preparation of HR-MAS NMR experiments is the choice of solvent. Dry kernels tend to absorb the added deuterated water ( $D_2O$ ) and exchange it with internal water ( $H_2O$ ). The result of this process is a poorer shimming and in turn increasingly lower quality spectra over time. An alternative solvent which could be used to replace  $D_2O$ , is deuterated DMSO. Viereck et al. (2009) found that HR-MAS NMR spectra of single wheat kernels soaked in  $D_2O$  gave better resolved spectrum in the carbohydrate region, whereas the DMSO spectrum was similar to a fully gelatinized starch spectrum due to the better solubilised starch polymers.  $D_2O$  resulted in a minimum of kernel damage, whereas DMSO left the sample very fragile. The choice of the experimental temperature is also an important parameter to consider. For example, in single kernels which can be regarded as a solid with a very rigid cellular and granular matrix, mobilisation of particularly the carbohydrates (starch and fibres) can be enhanced by heating of the samples. This consideration should be taken together with the specific problem of interest [186].

### **Three way data**

In the field of metabonomics, the dynamic characterization of metabolic changes over time is a fundamental aspect of elucidating the biochemical response of an organism to an external perturbation. It is not the change of a specific metabolite but the dynamic variation of the whole system in time that enters directly in the discrimination among the groups [187]. Exploratory studies of change over time have the advantages that data series are naturally arranged as a three-way data set. The first dimension can represent the treatment, the second dimension can represent the time and

the third dimension can represent the development of the metabolic profile measured by NMR spectroscopy. Castro and Manetti (2007) also used the multiway approach to analyze metabonomic data in a study of maize seeds development [187]. Jansen et al. (2008) used the higher-order advantage in the study of the effect of two doses of hydrazine in urine from rats, investigated by time-resolved metabolic fingerprinting [188]. In the work presented here, for the first time, grain filling HR-MAS NMR data has been analysed by PARAFAC and the result was an excellent overview of the data (Paper II). The PARAFAC approach combined with the multiparametric NMR method showed a great potential in the study of the grain metabolome evolution. In Paper II, it was further shown how visualisation of the data by PCA trajectories of HR-MAS NMR spectra of single wheat kernels can give good contrast of the metabolic development during grain filling. Keun et al. (2004) have previously shown that geometric trajectory analysis of metabolic responses can define treatment specific profiles [189].

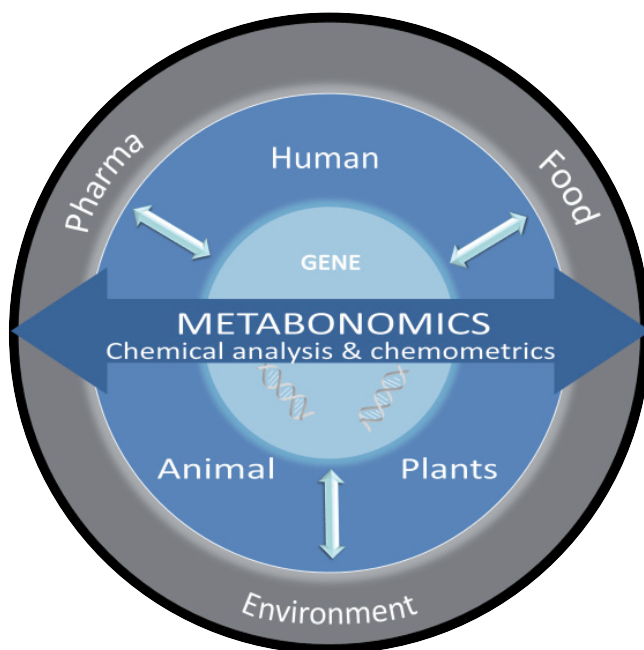


## 6. The dynamic metabolic events of NMR uroscopy in nutri-metabonomics

A promising opportunity exists for nutritional scientists to precisely identify the biological consequences of bioactive ingredients in food which in turn will provide more profound insight to their potentially beneficial consequences. This chapter concerns the usefulness of NMR spectroscopy in urine metabonomic (uroscopy) studies based on Paper III and unpublished results.

### *Metabonomics*

Nowadays, nutrition focuses on improving health of individuals through diet. By analyzing the metabolic content and concentrations of a biofluid using NMR spectroscopy, a metabolic profile can be generated that provides a non-invasive ‘whole body snapshot’. Comparison of these snapshots before and after a dietary intervention may highlight particular metabolites that respond. In this way, potential biomarker identification leads to a greater understanding of biochemical consequences of food, bioactive compounds and drugs. The definition of the *normal* metabolic profile is fundamental in human studies where additional variance is introduced through a number of uncontrolled factors (i.e., genetics, ethnicity, stress, exercise level, and diet) [190]. An illustration of metabonomics is given in Figure 25.



**Figure 25** Metabonomics can be defined as the analysis of metabolites using multivariate data analysis. Living cells produce changes in metabolite pattern under stimuli from nutrients or medicine (pharma) and environment. The genetic composition of the cell is determining for of metabolite changes.

The automated acquisition of large amounts of metabonomics data (and other -omics data) by modern NMR spectrometers results in exploratory and interpretative challenges. However, the abundance of data is not in itself a guarantee for obtaining useful information on metabolic events taking place in an investigated system. The data has to be investigated in a proper way, before any conclusions can be made. This will be discussed in the following.

### **Uroscopy**

Samples for metabonomics studies are mostly derived from body fluids such as blood, urine, spinal fluid, saliva and cerebrum spinal fluid or as biopsies of tissues. However, blood and urine are the most frequently used samples for exploring the systematic modification of metabolome. Compared with blood and plasma, urine samples enable non-invasive monitoring of metabolic changes that manifest in the urine [191]. Urine is a practical diagnostic sample because it is relatively safe and simple to collect, and contains information about the individual. Beyond diagnostic markers to diagnose pregnancy and specific diseases, urine can also provide other

relevant information about the individual, its health, physiological function and response to xenobiotics and diet [192].

### **Biological variations in urine**

Biologic material show generally great sample variability. The intra-individual metabolite variance i.e., the day to day variation of an individual and the inter-individual variation, i.e. within a normal population has to be considered before conclusions are made. Compared to plasma and blood, urine is highly varying in concentration and composition since urine is a body 'waste product'. On the other hand, blood is under constant chemical control by the body because of the vital importance of constant pH and glucose concentrations for instance. In contrast, all excessive metabolites from all body functions are excreted into the urine. Many factors influence the number and the amount of metabolites to be excreted into the urine. In urine from rodents, it has been established that species, strain, genetics, sex, age, hormone concentrations, diurnal cycles, diet, temperature, stress, and gut microflora all contribute to the metabolic composition of the urine [193]. Another factor that can contribute to variation of the urine is the sampling procedure. Sample preparation, time of sampling and storage can also impact on a clinical test result [194,195]. All in all, there is a risk that small and subtle metabolic responses disappear in the large variation caused by biological variation, experimental inhomogeneity or inadequate sample procedure. Keun and co-workers (2002) found that metabonomic urine analysis is an extremely precise analytical tool with evidently high analytical reproducibility of metabonomics and that the reproducibility of the NMR data indicates the reliability of the data acquisition. They also demonstrated that any analytical variation was many times smaller than the detected normal physiological variation in urine composition [174].

### **Nutri-metabonomics**

The changes in metabolic profiles resulting from dietary intervention can be difficult to detect as the contrast between the metabolic profiles produced from healthy individuals are small [196]. Metabonomics provides nutrition research with an alternative to the traditional single-biomarker approaches used to assess health and disease. <sup>1</sup>H NMR metabonomics has been used for several dietary studies of urine and many biomarkers have been identified which has been found interacting in dietary interventions. For example, NMR spectroscopy has been shown by van Velzen (2008) to be able to detect changes in phenolic compounds in urine after grape/wine consumption with

hippuric acid as the strongest biomarker [197]. Hippuric acid has also previously been reported to have metabolic impact of polyphenolic-rich food consumption (from tea) in humans [198-200]. In another study by Solanky and co-workers (2003) was epicatechin, a bioactive flavonoid widespread in many food products like green tea, and cocoa found in rats urine as a function of intake [201]. The identifiable biochemical effects associated with epicatechin intake included decreased urinary concentrations of taurine, citrate, dimethylamine, and 2-oxoglutarate. Solanky et al. (2005) have further investigated changes in the NMR spectral profile of urine collected from controlled dietary intervention studies in order to identify the biochemical effects of a diet rich in soy isoflavones, phytochemicals which are receiving significant attention because of their potential importance to human health and wide bioactivity *in vitro* [202]. Wang et al. used NMR spectroscopy to study the biological responses to chamomile tea ingestion and saw clear differentiation between the samples obtained before and after chamomile ingestion as an increased urinary excretion of hippurate and glycine and depleted creatinine concentration [203]. In these examples, urine NMR metabonomics (uroscopy) have shown useful in studies of biochemical effects of bioactive compounds. These studies confirm that many compounds are fluctuating in urine. However, metabonomics in human nutrition research is faced with the challenge that changes in metabolic profiles resulting from diet may be difficult to differentiate from normal physiologic variation [196]. This problem will be discussed in the following.

### ***Monitoring diet interventions by NMR spectroscopy***

Humans are extremely diverse organisms and it is therefore not surprising that most human metabonomics studies have found that spectral outputs are strongly influenced by inter- and intra-individual variation [196]. Normal physiologic variation may be a strong confounder in human studies and so far, most of the human metabonomics studies have clinical cases and controls. Volunteers in diet interventions studies are generally healthy and therefore in a metabolic homeostasis [202,204]. Still, changes in metabolic profile may be difficult to distinguish from normal physiological variation. The effects of the nutritional treatment tend to be much smaller than the biological variation that exists between the individuals [168,204,205].



One of the main challenges in nutritional metabonomics relies on the separation of specific metabolic signatures due to well-determined foods or diets. The metabolome is itself composed of numerous signatures from the confounding factors, from not only genotype, but also from i.e., gender, aging, physiological status, and lifestyle [193].

### **The Diogenes study**

One of the projects in this Ph.D. study was to investigate a dietary intervention study using uroscopy. Diogenes [206] or the full title “Diet, Obesity and Genes” is a controlled dietary intervention study in obese/overweight families (adults and children) in 8 different European countries, testing the efficacy of diets differing in Glycaemic Index (GI) and protein content. A large numbers of people across Europe are continuing to gain weight and when trying to lose weight, if successful, people subsequently find it very difficult to keep the weight off.

The weight-loss study aims to identify the diet which will be most effective in preventing weight gain and weight regain. Each parent has undergone an eight-week weight loss diet (using a low calorie diet formula), which has been designed to achieve a weight loss of 8% of their original starting weight. Then the parent is offered the opportunity to participate in the investigation of the problems related to weight re-gains. In this part of the study, volunteers have been assigned to one of five different dietary regimes, designed to test the relative effectiveness of GI and protein content in weight control.

The diets were:

- 1: Low Protein, Low GI
- 2: Low Protein, High GI
- 3: High Protein, Low GI
- 4: High Protein, High GI
- 5: Control, medium protein and medium GI

A controlled supermarket set-up was used to provide appropriate foods during the 6 months of the trial so each family could shop according to their diet group. All food from the supermarkets were provided free of charge. All subjects were allowed a 3 week break from the project, during which no recording of the dietary intake was required. As a first result of the Diogenes intervention study, it was established that an increase in dietary protein

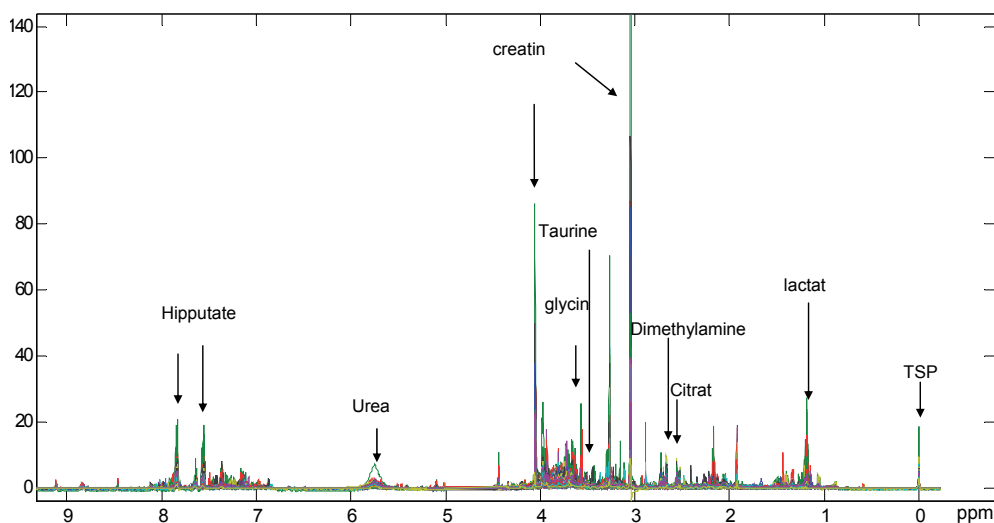
content decreases weight regain after weight loss, whereas there was no effect of GI on weight change [206].

The purpose of our part of the Diogenes study was to investigate the urine metabolome by  $^1\text{H}$  NMR spectroscopy and evaluate the results with multivariate chemometric approaches in order to identify significant biomarker-profiles. Approximately 100 persons in the five diet groups collected 24 h urine four times during the study; at day 0, (after the eight-week weight loss diet (before the diet intervention)) and 1, 3 and 6 months after the beginning of the diet intervention. See Table 2.

Table 2 Overview of urine sample collection days

24h urine collection	After 8 week weight loss	Diet intervention		
Clinical investigation day	Day 0	1 mdr	3 mdr	6 mdr

These urine samples were measured by NMR spectroscopy and Figure 26 shows the 478 NMR spectral urine profiles from this study.

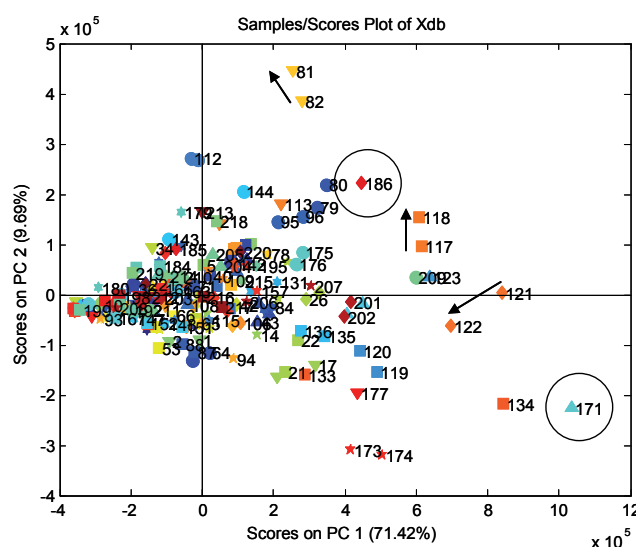


**Figure 26** 478 NMR spectra of human urine from the Diogenes study. Approximately 100 persons were each collecting four 24 h urine samples; at day 0 and 1, 3 and 6 months after the weight loss

The figure shows the great complexity of  $^1\text{H}$  NMR spectra of urine. Many signals seem to show variation between the samples and the problem of concentration variation between the samples is evident. Indeed, this study

turned out to be much more complicated than first assumed compared to the previously mentioned successful uroscopy studies.

Approximately half of the urine samples were prepared in replicates, i.e. two ‘samples’ were collected from the 24 h ‘pool’ and each of these were prepared in NMR tubes and measured, yielding 222 replicates in total. Figure 27 shows a PCA score plot of the 222 replicate NMR spectra of urine from the diet intervention study.

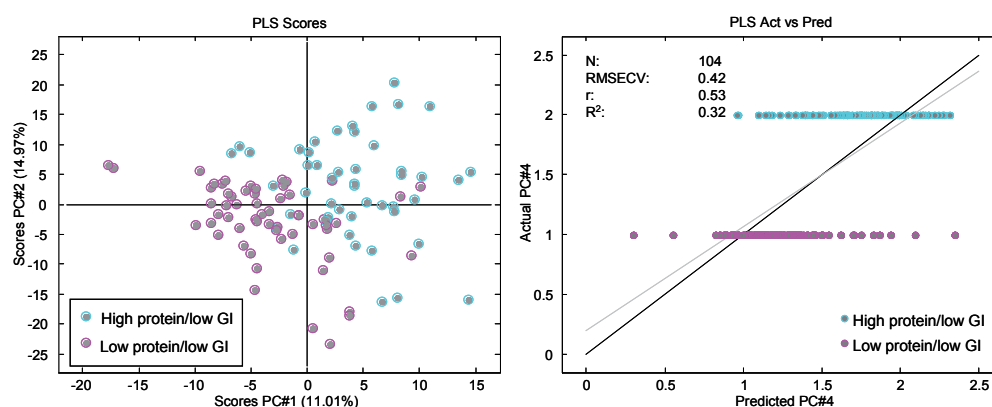


**Figure 27** PCA score plot of 222 replicates of urine spectra (mean centred) from the diet intervention study. The arrows indicate the distance and direction of some of the replicates.

The samples are numbered in pairs, so that e.g. 1 + 2 and 121 + 122 are replicates. This plot reveals the variation within the NMR data between the replicates. This variation could be caused by the sample preparation and/or variation in field homogeneity or other small experimental variations. These variations must be much smaller than the variation between the samples due to the intervention study in order to obtain reliable results. These variations can to a certain extent be minimized using some of the spectra post-(recording) processing and pre- (modelling) processing steps mentioned in Chapter 1 & 2. Inspection of score plots of replicates can also help to test the best pre-processing method as the average Euclidian distance can be calculated as a target function of most efficient pre-processing method. In

Figure 27, a couple of samples (171, 186) have large distance to their replicate sample. These samples, so-called outliers, should be removed from the data set before further analysis. What is even more evident from the score plot is that there are no systematic variations induced in the sample set, as the replicates 'wanders' in different directions (arrows in Figure 27). If systematic variation was induced due to e.g. a time parameter, replicates would show the same 'trend' and split up in the same direction.

The overall aim of the study was to discriminate between any combinations of large vs. small weight loss, large vs. small maintained weight loss or weight gain during the 6 months diet intervention. The discrimination was also concerning normal vs. high intake of protein and low vs. high content of GI in the diet. In order to determine which metabolites that is responsible for possible differences in the diet combinations and to characterize the different metabolic fingerprints, supervised modelling is applied. As mentioned in Chapter 2, the PLS-DA model should always be used with caution as it tends to overfit the data. However, sometimes it is practical to start out with PLS-DA exactly because of its ability to overfit. If PLS-DA is not able to model the data, no relations probably exist between the NMR spectral urine profiles and the diet groups. Figure 28 shows the PLS-DA model obtained on the supposed most extreme diet groups, having high protein/low GI and low protein/low GI. Unfortunately, no systematic variation seems to exist in the data that can discriminate between the two diet groups.



**Figure 28** PLS score plot of 104 NMR spectra of urine from high/low protein diet and low glycemic index together with actual versus predicted plot containing model specifications. Day 0 is excluded because the diet has not yet started which result in a total of 104 NMR spectra

From the plot it seems reasonable that removal of the worst predicted samples together with the most outlying samples would be able to improve the predictions. Unfortunately, this is not the case. The case here is a typical example of an overfitted model; in the score plot there seems to be coherence between diet group and their urine profile, but even with half of the 'worst' samples removed the model performances remain unsatisfactory. All the previously mentioned multivariate tricks and pre-processing approaches were systematically investigated, however, still without any model improvements. In conclusion, we found that it was not possible to model gender, age, weight gain/weight loss, smokers/non-smokers or even the day of investigation (0, 1, 3 or 6 month).

This study illustrates some of the most common problems in NMR measurements of urine. The urine metabolome in this study was found to vary more between the samples compared to the changes of interest induced by the nutritional intervention. This is the major problem in uroscopy studies. A challenge remains in uroscopy in eliminating the 'noise' due to random daily variations. However, Assfalg et al. (2008) found evidence of different metabolic phenotypes in humans. This was only made possible when measuring urine from 22 individual persons, 40 times over a period of 3 months. This amount of data is apparently sufficient to average out and eliminate noise due to daily variations [207]. Therefore, it is important to define the inter- and intra-individual sample variance within a normal population before biochemical or clinical conclusions are made.

Another reason why we did not obtain significant results was the lack of control of the subjects. In diet intervention studies one more problem is often seen. Self-reported dietary intake is often used for establishing diet-disease associations. Food frequency questionnaire has been the dietary assessment method used most frequently in large-scale studies, primarily because it is easy to administer, is less expensive than other dietary assessment methods, and provides a rapid estimate of usual intake [208]. However, there exists a great problem in using food frequency questionnaires. For example, McKeown et al. (2001), observed that the correlation between 24 hour urinary nitrogen excretion and dietary intake as evaluated by the food frequency questionnaire was as low as 0.25 [209]. Instead, biomarkers do not always correspond to the self-reports of food intake and thus random measurement errors of the biomarker are not likely to be correlated with

those of the dietary assessment method [210]. This indicates that there exist great bias between what people say they have eaten and the actually dietary intake.

Indeed, it is known that human volunteers in dietary studies do not report all their medication or food supplements [211]. Therefore, it would be highly desirable to test or check what test-persons really eat under a diet intervention. *In vivo* investigations using animals as study-models make it possible to study biomarkers for different food intake. Animals can be exposed to much more controlled environments (temperature, humidity and exercise) and diet (controlled amount of food and water) and they can be much more identical with respect to gender, genetic and physics (height and weight). These stringent conditions are not possible with humans as test persons who volunteer to follow diet intervention studies.

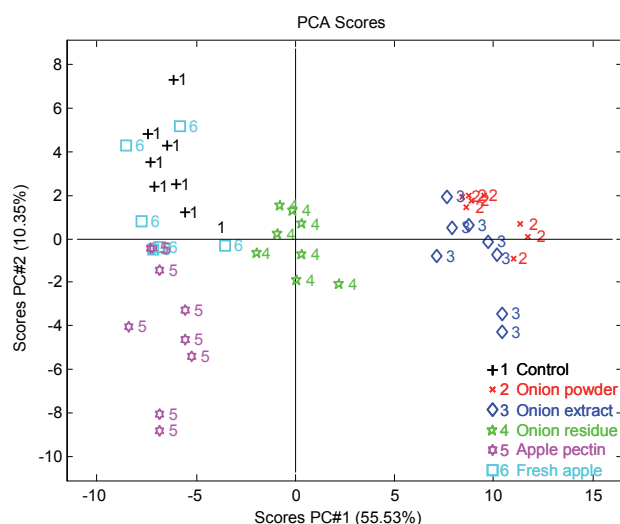
The results of the Diogenes study were not convincing and it was not possible to make conclusions. Therefore, results have not been published yet. There are many reasons why the study did not give the result we expected. The diet intervention itself was uncontrolled as each family could shop free of charge according to the diet group, but it is still unknown what the individual person really haven eaten. Only the *average* amount of protein and the GI of the diet were controlled. Furthermore, the three urine snapshots (per person) from the diet intervention during the six month do not cover the variation from day-to-day. And since the diet was not more controlled, there is a possibility that the day of urine investigation was the day where the person did not hold strict too the diet or exactly match the period where the person took the allowed break from the project. Furthermore, taken into account that test person does not always follow their diet strictly, another source of uncontrolled variation is induced. In the following, a successful uroscopy application using rats as animal study-models to study biomarkers is described (Paper III)

### ***Exploration of biomarkers for onion intake***

The concept of developing a nutritional enhanced food product requires deep understanding of the mechanisms of protection, including identification of the biologically active molecules in order to demonstrate the efficacy of these molecules. Fruits and vegetables contain several bioactive

compounds and onions and apple have been described to have potentially beneficial effects on health.

In this work, biomarkers for onion and apple intake were investigated by NMR spectroscopy of urine from rats. The result regarding onion biomarkers are presented in Paper III but the results regarding apple biomarkers are not yet published. In this study, 48 male rats were divided into 6 diets groups; one control group, 3 groups were given onion (by-product/extracts/residue) and the last 2 groups were given apple (raw apple and apple pectin). NMR spectroscopy showed in combination with chemometrics to be an excellent tool for classifying diet groups by their urine profile. It turned out to be possible to distinguish between urine from rats with a normal diet and rats on an onion or apple diet in a simple PCA plot of the NMR spectra of a selected interval found by iPLS (Figure 29).



**Figure 29** PCA score plot of 32 NMR urine spectra of rats on six different diets

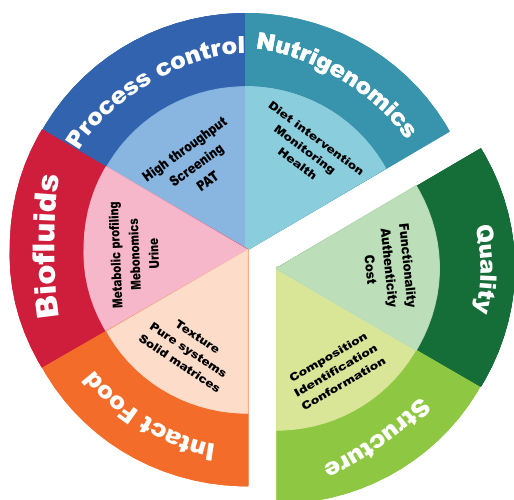
This impressive result illustrates how much easier it is to use animal models because the variation between samples will reflect the diet intervention rather than the confounding factors.

Further investigations were carried out regarding biomarkers from onion intake. In order to investigate the spectra for signals able to distinguish between the control feed and the three different onion fed groups (by-product, extract and residue), iECVA was carried out on the urine spectra.

The ECVA score plot (not shown) obtained on the selected intervals gives four distinct groups. Two interesting intervals were found by iECVA which was able to improve the misclassifications rate significantly. The two intervals, 6.50-6.95 ppm and 2.98-3.42 ppm revealed signal from onion biomarkers in urine from the three onion feed groups (by-product, extract and residue). The two onion biomarkers were identified as dimethylsulfone and 3-hydroxyphenylacetic acid. Therefore, it was possible to identify an objective biomarker for onion intake but not for the different onion products.

Being able to detect and maybe even identify specific diets could add interesting possibilities e.g. in human intervention studies where urine could be used to detecting biomarker of different food intake. With respect to the health beneficial effect of onion, developing dietary supplements or nutritionally enhanced functional foods including onion fractions could be valuable. The concept of developing nutritionally enhanced functional foods requires the understanding of the mechanisms of prevention and protection if potential nutritional effects of an onion supplement should be utilized and the effect documented. The identification of biologically active molecules as potential biomarkers leads to a greater understanding of biochemical pathways and potentially allows objective quantification of onion intake in human studies.





## 7. NMR spectroscopy for advanced quality control of food ingredients

NMR spectroscopy has a long tradition in structure elucidation of biomolecules and can provide information on molecular structure, dynamics, and interactions with unbiased detection. Hydrocolloids are a group of complex carbohydrates wherein the colloid particles are dispersed in a water solution. A hydrocolloid can take on different states, e.g., gel or sol and are employed in food mainly to influence texture or viscosity. NMR spectroscopy provides great help in understanding complex carbohydrate structures and thereby is possible to analyse their functionality. In Paper IV and one of the Additional Paper I on spectroscopic analyses of functionality of pectins have been carried out.

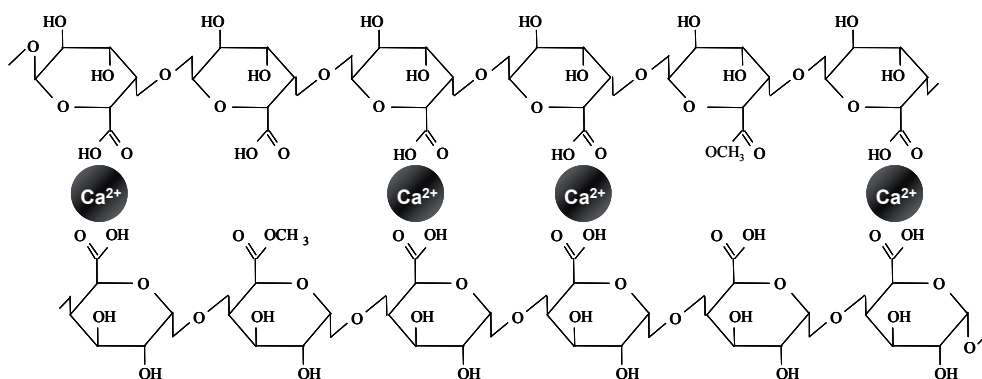
### ***Pectin: a case study***

Pectin is commercially produced mainly from citrus fruit peel and apple pomace, which are by-products from citrus or apple juice production. From an application point of view, citrus and apple pectins are largely equivalent. Because of its excellent gelling, thickening, and stabilizing properties, extracted pectin (E440) is extensively utilized in the food industry as gelling agents in jams, confectionary, and bakery fillings, and stabilisers in yoghurts and milk drinks [212,213]. Pectin is a natural part of human diet through the

daily intake of pectin from fruit and vegetables. In human digestion, pectin passes through the small intestine more or less intact and pectin is thus considered as a soluble dietary fibre. Consumption of pectin has been shown to reduce cholesterol levels in blood, as the increase of viscosity in the intestinal tract, leads to a reduced absorption of cholesterol from bile or food [214]. In the large intestine and colon, microorganisms degrade pectin and liberate short-chain fatty acids which have positive prebiotic influence on health [215].

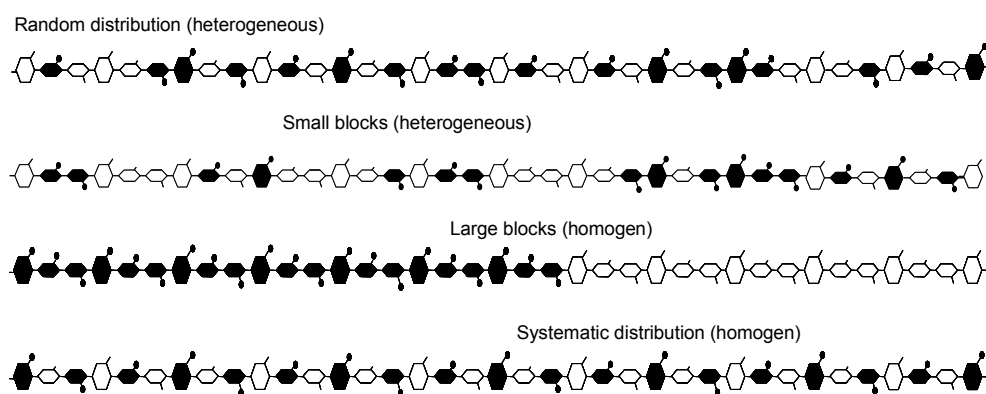
### Functionality

Pectic molecules include primarily homogalacturonan (HGA) and rhamnogalacturonan I (RG I). HGA is composed of unbranched  $\alpha$ -1,4-linked galacturonic acid (GalA) residues, whereas RG I is composed of a backbone of repeating  $\alpha$ -1,2-L-rhamnose(Rhap)- $\alpha$ -1,4-D-GalA disaccharide units. The GalA residues in the HGA and RG I backbone may be methyl esterified and/or O-acetylated [216]. Side-chains, mainly consisting of arabinan and/or galactan, are attached to the RG I backbone at the C-4 position of the Rhap residues. These side-chains are called 'hairy regions' [216,217]. Regions of free (deesterified) GalA units bind calcium (Figure 30) and may form multiple calcium bridges, which create a domain of strong intermolecular association between the galacturonan chains, resulting in increased viscosity. Multiple regions of predominant free GalA residues in the presence of an excess of calcium ions, causes pectin gelling or precipitation [218].



**Figure 30** A fraction of homogalacturonan (HGA) which is partly methyl esterified at C-6 position

Calcium binding studies has showed that the calcium sensitivity of high esterified pectin occurs in the presence of inhomogeneous methyl ester distribution. Therefore, the functionality of pectin is not only depending on the number of methyl esters, the degree of esterification (DE), but also on the distribution of the methyl ester groups along the pectin polymer. Native pectins are varying in DE depending of biological source, but for citrus pectin the DE is approximately around 73%. The DE for a given pectin can be adjusted chemically or enzymatic, depending on the end use [219]. The de-esterification method has great influence on the subsequent distribution of the free ester groups. The de-esterification can be blockwise or random depending on the method. Random and blocky methyl ester distributions are interpretations as illustrated in Figure 31, but there exist no precise definition of how many contiguous free GalA units constitute a block.



**Figure 31** Schematic example of a 3-folds conformation of HGA chains with different distribution of free GalA units

Besides the DE and the distribution of the ester groups also the helical conformation of the pectic polymer is important. Braccini et al. (1999) found that both the 2- and the 3- fold conformation for polygalacturonic acid are almost equally favorable. Thus, depending on the environment (neighboring ions, solvent, etc.), one or the other form may exist [220].

### Blockiness in pectin

The degree and pattern of methyl esterification has important commercial implications. Previous investigations of the methyl ester distribution of

pectin have included (among others)  $^1\text{H}$  NMR spectroscopy together with enzymatic degradation [221,222]. However, studies using enzymatic degradation have the disadvantage that they tend to be destructive with major reduction in the molecular weight of the pectin molecules. An alternative to the enzymatic degradation of the polymer is to use spectroscopy such as NMR spectroscopy and vibrational spectroscopy. With respect to pectins, near-infrared (NIR) and infrared (IR) spectroscopy has proved capable of monitoring the DE of pectin powders [223,224] using full-spectrum regression methods such as PLS. In the search of pectin derivatives with improved functionality properties it would be advantageous to obtain a rapid spectroscopic method which can quantify both the DE and the distribution of methyl esters. For this reason, a sample set of 31 pectin samples with four different ester distributions were produced at CP Kelco (Denmark), (formerly Copenhagen Pectin Factory). This was done using enzymes which de-esterifies in a random or blockwise manner, respectively. While the ester groups give rise to specific group frequencies or chemical shifts, signals relating to their spatial distribution along the pectic polymer are bound to be more subtle why pattern recognition methods such as PCA and PLS is applied in order to extract the information.

The designed set of pectin with block and random distribution of free GalA units has been exploited using NMR, IR, Raman and NIR spectroscopy [225] (Paper IV). NMR spectra were obtained on 0.5 % (w/w) pectin solutions whereas IR, Raman and NIR spectra were recorded on pure pectin powder. Prediction models were developed for random and blockwise distribution. It turned out to be possible to predict random deesterification more easily than blockwise for all the spectroscopies. This is possibly due to the fact that the random enzyme simply removed the blockwise distribution pattern leaving only signal from random distribution. Large de-esterified block leaves apparently no specific spectral signals.

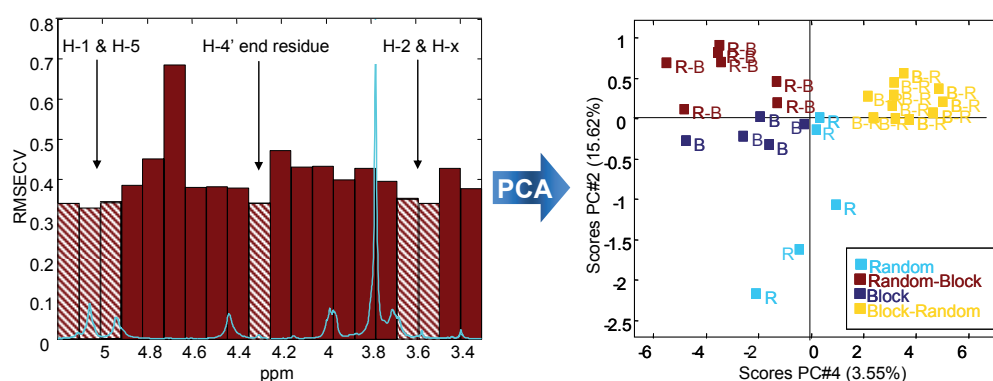
#### **iPLS exploits functionality in the spectra**

Using the variable selection method iPLS, specific regions of NMR, IR and NIR spectra were revealed in which the prediction models of random GalA distribution could be significantly improved. These areas hold quantitative information on random de-esterification which in the NMR spectra corresponds to H-2 in the GalA residue. The most influential spectral area of the IR spectrum ( $970\text{--}870\text{ cm}^{-1}$ ) found by iPLS is dominated by (C-O-H) and ( $\text{CH}_3$ ) in-plane bending. This signal has good potential to be affected by the

randomness, since these vibrations are correlated to DE. The interval in the NIR spectrum (7400-7200  $\text{cm}^{-1}$ ) which showed significance with respect to the random ester distribution is probably due to first overtones of combination bands (C-H stretch and C-H deformations) originating from  $\text{CH}_3$  groups.

In order to investigate the difference between the four different enzyme treatments of the pectin polymers, classification by enzyme treatment was investigated. This study showed that it is possible to extract information not relating to the quantitative information extracted by the regression methods. ECVA classifications were carried out on the vibrational spectroscopy data ensembles. The ECVA revealed one specific signal in the Raman (1045  $\text{cm}^{-1}$ ) spectra and one significant area (1250-1400  $\text{cm}^{-1}$ ) in the IR spectra which was able to classify the pectin samples according to block and random ester distribution. The signal at 1045  $\text{cm}^{-1}$  in the Raman spectrum is assumed to be due to changes in the galacturonic acid backbone structure when the substitution of the galacturonic acid chain with ester groups is changed. The interpretation and assignment of the signal in the IR spectrum were more difficult.

The significant area with respect to classification by enzyme treatment in the NMR spectra was revealed by iPLS (Figure 32).

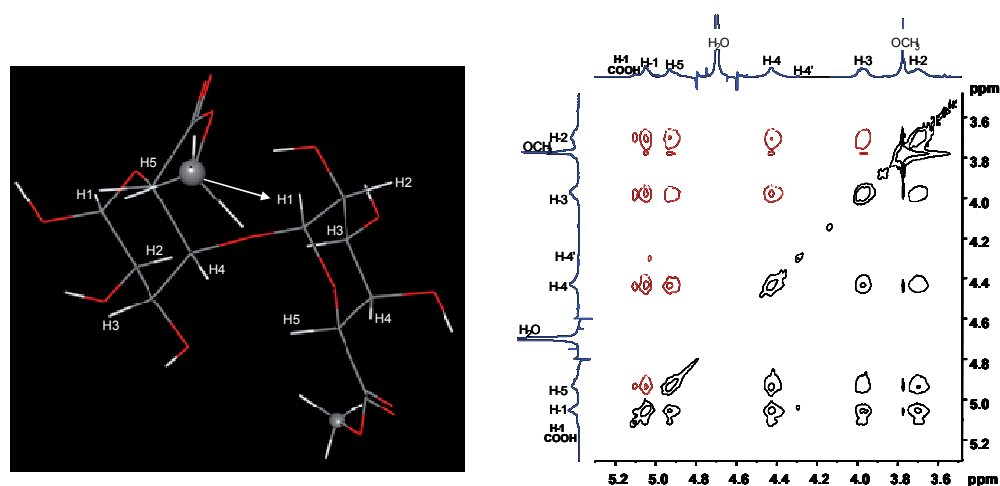


**Figure 32** iPLS plot discriminated by enzyme treatment and the corresponding PCA score plot (PC2 versus PC5) of the four enzyme treatments obtained on the hatched intervals. Paper IV

These signals are possible due to the H-1, H-5, the H-4 end residue and the H-2. To explain this, the spectral conformation is further exploited.

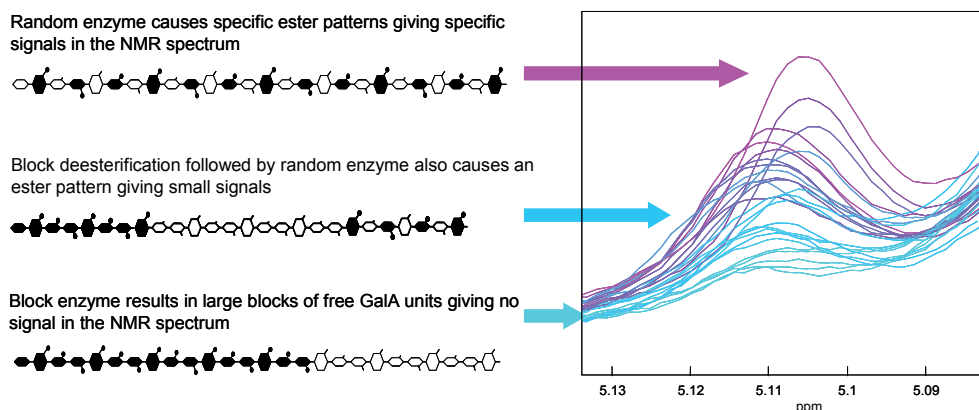
### Spatial pectic conformation

A nearly complete assignment of the  $^1\text{H}$  NMR spectrum of pectin solutions was carried out. A computer model of a 3-fold helical turn of a trimer GalA unit (fully esterified) of the pectic backbone is shown next to a plot from a NOESY experiment (Figure 33). The distances between the methoxy group at one GalA unit (n) and the H-1, H-3, H-4 and the H-5 of the two neighbour GalA units (n-1 and n+1) are measurably very short.



**Figure 33** Molecular model of a 3-fold helical structure of two GalA units illustrating the distances between the methoxy group at one GalA unit (n) and the H-1, H-3, H-4 and the H-5 of the two neighbour GalA units (n-1 and n+1). Right: part of a NOESY spectrum of pectin showing NOE signals from  $\text{OCH}_3$  to H-1 (on de-esterified neighbour unit)

Some very interesting correlations in the NOESY spectrum were observed between the methoxy signal and the H-2 and H-3 signals which were found to be due to the NOE effects from the spatial pectin structure. This indicates that the pectic polymer has a three-fold helical conformation. The shortest distance seems to be between the methoxy group and the H-1 at the neighbouring GalA unit. Examining this signal closer showed that the H-1 signal splits up depending on the esterification. A de-esterified GalA shows in a signal at 5.11 ppm as shown in Figure 34.



**Figure 34** Part of the  $^1\text{H}$  NMR spectra of pectins, illustrating the specific signal caused by random ester distribution. The variation is due to the 3-fold helical structure of the pectic backbone and to the spatially close position of the neighbouring free- and esterified GalA units

When colouring the signal according to the de-esterification, clearly some systematic are emerging. The signal with the highest intensity originates from the random de-esterification. Blockwise distributions of free GalA units give rise to signals with lowest intensity. In the middle are signals from pectin samples which are de-esterified with half block half-random enzymes (Paper IV).

This is a beautiful example of how NMR spectroscopy can be used to elucidate detailed structural conformation. It demonstrates the high utility of using variable selection methods to elucidate interesting regions of the spectra. Subsequently, the versatility of NMR spectroscopy makes it possible to elucidate the structure. This clearly illustrates of the unique combination between NMR spectroscopy and chemometrics.





## 8. Conclusion

This Ph.D, study has contributed with insight and deeper understanding of the possibilities and advantages of the use of NMR spectroscopy within many different food related applications. The major advantages as well as the limitations of NMR spectroscopy in combination with chemometrics have also been characterized. The study illuminates the numerous possibilities of NMR spectroscopy within food and nutrition research, especially concerning the use of chemometrics and data mining techniques in combination with NMR spectroscopy.

Exploitation of complex NMR spectra using chemometrics is not routinely utilized. In the work presented here, the challenges in data analysis of NMR spectra, where spectral sensitivity and spectral pre-processing are affected by e.g. signal overlap have been described. The quantitative aspects of NMR spectroscopy have been exploited and advanced chemometric modelling has been applied. MCR was shown to be a perfect chemometric tool for visualization of NMR spectroscopic data of a ternary experimental design of alcohol mixtures. The MCR was able to separate alcohol mixtures into pure alcohol spectra and corresponding concentrations. Using this design as a simplified simulation of metabonomic applications, these aspects of NMR were initially explored. The quantitative nature of NMR spectra was examined and it was suggested to be potential also in future PAT applications.

In the collected work presented in this thesis, NMR spectroscopy revealed its extraordinary potential, when applied to natural samples and products. NMR spectroscopy was applied for exploiting patterns in the protein metabolism of single wheat kernels. The implications of one or more drought events on protein quality in developing wheat grains was investigated using liquid state  $^1\text{H}$  NMR spectroscopy of methanol extracts and  $^1\text{H}$  HR-MAS NMR spectroscopy of single kernels and flours. HR-MAS

NMR spectroscopy showed the possibility of studying the effect of different drought events on the protein synthesis in whole grains of wheat. Multi-way PARAFAC modelling of the NMR data revealed a great potential in the study of the dynamic grain metabolome evolution. In relation to this, the possibilities and advantages of studying intact biological matrices were described, including critical experimental parameters. In addition, the importance of sampling and the challenges in sample preparation for assurance of quantitative NMR spectra were studied.

When developing nutritional enhanced food products a thorough understanding of the mechanisms of health protection as well as a demonstration of their effectiveness *in vivo* is required. Fruits and vegetables contain several bioactive compounds; onions and apple have been described to have potentially beneficial effects on health. The identification of biologically active molecules as a result of onion intake by rats was investigated in order to obtain objective quantification of onion intake. iECVA revealed specific biomarkers in the NMR spectra of rat urine as a result of onion intake. Two biomarkers for onion intake were identified; dimethylsulfone and 3-hydroxyphenylacetic acid. Being able to identify and quantify onion intake ads great advantage with respect to understanding and monitoring the health beneficial effects of dietary supplements or nutritionally enhanced functional foods including onion. The importance of understanding the limitations and pitfalls of metabonomic data analysis including the limitations of experimental design as well as the risk of overfitting the chemometric models was also established.

The enormous potential of NMR spectroscopy to provide information about the qualitative chemical structure and quantitative content simultaneously was demonstrated in a study of structures of pectins. Quantitative and qualitative information was extracted by chemometric models, including PCA, PLS and IPLS. Especially the interval based chemometric models were shown to be excellent exploratory tools in obtaining knowledge about informative regions of the NMR spectrum. This result was general for the collected work. iPLS and iECVA proved to be particularly well suited to explore multivariate NMR data sets. The reason for this is partially the mega-variate character of NMR data with large baseline regions and partially the huge scaling dynamics in NMR spectra.

In this thesis, the chosen different applications of NMR spectroscopy have been illustrated as one or two 'pieces' in a pie chart. This representation is meant to illustrate the great versatility of NMR spectroscopy and the strong synergetic relation between NMR spectroscopy and chemometrics.



## 9. Perspectives

### **General aspects of NMR and chemometrics**

NMR spectroscopy has a long history and tradition. Applications of the technique have developed from chemical identification and quantification, to elucidation of chemical connectivity, to measurement of intact matrices and to the analysis of biofluids and nutritional diet intervention studies. However, new potential aspects in using NMR spectroscopy combined with chemometrics will be further developed as NMR spectrometers, NMR probes and chemometric modelling are advancing rapidly. New NMR spectrometers installed with cryo probes are becoming more affordable and the more sensitive which makes it possible to obtain even more detailed information in shorter time. This continuous improvement in NMR technology also makes measurements of other NMR active nuclei (besides protons) easier, which provide new possibilities in food and nutrition sciences. Measuring nuclei such as phosphor can offer new insight into metabolic profiles, in foods processes and in human metabolism. Optimal exploitation of such NMR data will demand advanced chemometric methods.

Within chemometrics, new possibilities of handling more and more complex data are continuously emerging. The development of more advanced multivariate techniques also expands the uses of NMR spectroscopy. Multi-way chemometric techniques are well suited for the analysis 2D NMR data. PARAFAC and Laplace inversion have already successfully been used on 2D DOSY NMR spectra [143,226] and these multi-way techniques can further be applied to 2D scalar coupling (COSY, TOCSY and HSQC) NMR data. Also, the development of new chemometric methods alternative to the current basic chemometric PCA modelling for specific handling of complex NMR data is desirable. The PCA has some obvious disadvantages when applied to complex NMR data. PCA is highly sensitive to noise, but not sensitive enough when relevant spectral information is concentrated in a

few, small peaks among many larger signals. Obtaining chemometric models tailored for NMR data would be highly useful for NMR scientists.

#### **Future potential in PAT**

A near future aspect of the use of modern NMR spectroscopy is in process control as alternative method to the already implemented spectroscopic techniques (e.g. NIR and IR). Development of new flow probes can make on-line NMR systems available, which are able to analyse, monitor and control industrial processes with an unprecedented level of information. This could be highly beneficial in the pharmaceutical production, where monitoring and documentation of process control is required for real time release. However, there are still challenges regarding price, delicacy of instrument, including the sensitivity of the magnetic field to ion, and e.g. measurements of moving samples on conveyer belt, which have to be solved, before it can be applied on-line in the industrial sector.

#### **Future potential in cereal science**

As already discussed, NMR spectral pattern recognition has been used in authenticity studies where the origin of the product is revealed in a specific, metabolic profile. Screening of the detailed plant metabolome for bioactive substances could be used to develop a metabolite database of NMR spectra of specific cereals: wheat, barley, rye and oat, including ancient cereal sources: spelt wheat and emmer wheat. This database would be beneficial to food and nutrition scientists, in order to utilize it for developing new, healthy products.

#### **Future potential in uroscopy**

There is an unexploited potential of a standardized analysis of urine within food, health and medical research. Being able to dose medicine on an individual basis with respect to the metabolism of the individual will prevent problems regarding overdose and avoid unnecessary side-effects. However, there are still some problems to be solved. The large day-to-day variations can somehow be handled by intense data collection from each individual and by full cross-over interventions studies, in order to obtain the correct inter-variation. With this knowledge, it will be possible to model the individual variation and remove the information not related to the problem under investigation in the data.

### **Future potential in hydrocolloid production**

NMR spectroscopy is widely used in the hydrocolloid industry, for example in the analysis of alginate. The molecular composition, depending on the source of raw material, is closely related to the functionality of alginate. Salomonsen *et al.* (2008) have developed a rapid spectroscopic method for determining the alginate composition on the basis of solid state NMR spectroscopy [227]. The molecular composition of pectin depends, amongst other parameters, on the natural source of raw material. Thus, it is reasonable to believe that NMR spectroscopy has a high potential in the pectin industry as well.

Finally, the hyphenation of NMR spectroscopy with chromatographic methods is very interesting to explore. Especially, there could be a huge potential in the combining and synergies of the two widely used techniques: NMR and MS. Combining these methods would give great insight and control in the metabonomic research of biofluids.





## Reference List

- [1] I. I. Rabi, Space quantization in a gyrating magnetic field, *Physical Review*, 51 (1937) 0652-0654.
- [2] G. Breit and I. I. Rabi, Measurement of nuclear spin, *Physical Review*, 38 (1931) 2082-2083.
- [3] F. Bloch, W. W. Hansen, and M. Packard, The Nuclear Induction Experiment, *Physical Review*, 70 (1946) 474-485.
- [4] D. B. Kell, Systems biology, metabolic modelling and metabolomics in drug discovery and development, *Drug Discovery Today*, 11 (2006) 1085-1092.
- [5] S. M. Watkins and J. B. German, Metabolomics and biochemical profiling in drug discovery and development, *Current Opinion in Molecular Therapeutics*, 4 (2002) 224-228.
- [6] J. C. Lindon, E. Holmes, M. E. Bollard, E. G. Stanley, and J. K. Nicholson, Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis, *Biomarkers*, 9 (2004) 1-31.
- [7] J. L. Griffin and M. E. Bollard, Metabonomics: Its potential as a tool in toxicology for safety assessment and data integration, *Current Drug Metabolism*, 5 (2004) 389-398.
- [8] L. K. Schnackenberg and R. D. Beger, Monitoring the health to disease continuum with global metabolic profiling and systems biology, *Pharmacogenomics*, 7 (2006) 1077-1086.
- [9] S. H. Moolenaar, U. F. H. Engelke, and R. A. Wevers, Proton nuclear magnetic resonance spectroscopy of body fluids in the field of inborn errors of metabolism, *Annals of Clinical Biochemistry*, 40 (2003) 16-24.
- [10] D. S. Wishart, L. M. M. Querengesser, B. A. Lefebvre, N. A. Epstein, R. Greiner, and J. B. Newton, Magnetic resonance diagnostics: A new technology for high-throughput clinical diagnostics, *Clinical Chemistry*, 47 (2001) 1918-1921.

- [11] M. J. Gibney, M. Walsh, L. Brennan, H. M. Roche, B. German, and B. van Ommen, Metabolomics in human nutrition: opportunities and challenges, *American Journal of Clinical Nutrition*, 82 (2005) 497-503.
- [12] J. K. Nicholson, J. C. Lindon, and E. Holmes, 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data, *Xenobiotica*, 29 (1999) 1181-1189.
- [13] E. J. Saude, D. Adamko, B. H. Rowe, T. Marrie, and B. D. Sykes, Variation of metabolites in normal human urine, *Metabolomics*, 3 (2007) 439-451.
- [14] O. Fiehn, Metabolomics - the link between genotypes and phenotypes, *Plant Molecular Biology*, 48 (2002) 155-171.
- [15] K. Raynal-Ljutovac, G. Lagriffoul, P. Paccard, I. Guillet, and Y. Chilliard, Composition of goat and sheep milk products: An update, *Small Ruminant Research*, 79 (2008) 57-72.
- [16] G. E. Pereira, J. P. Gaudillere, C. Van Leeuwen, G. Hilbert, M. Maucourt, C. Deborde, A. Moing, and D. Rolin, <sup>1</sup>H NMR metabolic profiling of wines from three cultivars, three soil types and two contrasting vintages, *Journal International des Sciences de la Vigne et du Vin*, 41 (2007) 103-109.
- [17] M. A. Brescia, V. Caldarola, A. De Giglio, D. Benedetti, F. P. Fanizzi, and A. Sacco, Characterization of the geographical origin of Italian red wines based on traditional and nuclear magnetic resonance spectrometric determinations, *Analytica Chimica Acta*, 458 (2002) 177-186.
- [18] F. H. Larsen, F. van den Berg, and S. B. Engelsen, An exploratory chemometric study of <sup>1</sup>H NMR spectra of table wines, *Journal of Chemometrics*, 20 (2006) 198-208.
- [19] P. S. Belton, I. J. Colquhoun, E. K. Kemsley, I. Delgadillo, P. Roma, M. J. Dennis, M. Sharman, E. Holmes, J. K. Nicholson, and M. Spraul, Application of chemometrics to the <sup>1</sup>H NMR spectra of apple juices: discrimination between apple varieties, *Food Chemistry*, 61 (1998) 207-213.
- [20] I. Berregi, G. Del Campo, R. Caracena, and J. I. Miranda, Quantitative determination of formic acid in apple juices by <sup>1</sup>H NMR spectrometry, *Talanta*, 72 (2007) 1049-1053.
- [21] G. Del Campo, J. I. Santos, I. Berregi, and A. Munduate, Differentiation of Basque cider apple juices from different cultivars by means of chemometric techniques, *Food Control*, 16 (2005) 551-557.
- [22] M. A. Brescia, G. Alviti, V. Liuzzi, and A. Sacco, Chemometric classification of olive cultivars based on compositional data of oils, *Journal of the American Oil Chemists Society*, 80 (2003) 945-950.

- [23] M. Nilsson, I. F. Duarte, C. Almeida, I. Delgadillo, B. J. Goodfellow, A. M. Gil, and G. A. Morris, High-resolution NMR and diffusion-ordered spectroscopy of port wine, *Journal of Agricultural and Food Chemistry*, 52 (2004) 3736-3743.
- [24] I. F. Duarte, A. Barros, C. Almeida, M. Spraul, and A. M. Gil, Multivariate analysis of NMR and FTIR data as a potential tool for the quality control of beer, *Journal of Agricultural and Food Chemistry*, 52 (2004) 1031-1038.
- [25] D. W. Lachenmeier, W. Frank, E. Humpfer, H. Schafer, S. Keller, M. Mortter, and M. Spraul, Quality control of beer using high-resolution nuclear magnetic resonance spectroscopy and multivariate analysis, *European Food Research and Technology*, 220 (2005) 215-221.
- [26] N. Viereck, P. I. Hansen, and S. B. Engelsen, Undersøgelser af funktionelle faktorer i mejeriprodukter ved hjælp af NMR spektroskopi og kemometri, *Mælkeritidende*, 25 (2003) 616-620.
- [27] E. R. Andrew, A. Bradbury, and R. G. Eades, Nuclear magnetic resonance spectra from a crystal rotated at high speed, *Nature*, 182 (1958) 1659.
- [28] E. R. Andrew, A. Bradbury, and R. G. Eades, Removal of dipolar broadening of nuclear magnetic resonance spectra of solids by specimen rotation, *Nature*, 183 (1959) 1802-1803.
- [29] I. J. Lowe, Free Induction Decays of Rotating Solids, *Physical Review Letters*, 2 (1959) 285-287.
- [30] O. Soubias, M. Piotto, O. Saurel, O. Assemat, V. Reat, and A. Milon, Detection of natural abundance  $^1\text{H}$ - $^{13}\text{C}$  correlations of cholesterol in its membrane environment using a gradient enhanced HSQC experiment under high resolution magic angle spinning, *Journal of Magnetic Resonance*, 165 (2003) 303-308.
- [31] C. Sizun, J. Raya, A. Intasiri, A. Boos, and K. Elbayed, Investigation of the surfactants in CTAB-templated mesoporous silica by  $^1\text{H}$  HR-MAS NMR, *Microporous and Mesoporous Materials*, 66 (2003) 27-36.
- [32] D. Morvan, A. Demidem, J. Papon, and J. C. Madelmont, Quantitative HR-MAS proton total correlation spectroscopy applied to cultured melanoma cells treated by chloroethyl nitrosourea: Demonstration of phospholipid metabolism alterations, *Magnetic Resonance in Medicine*, 49 (2003) 241-248.
- [33] P. Weybright, K. Millis, N. Campbell, D. G. Cory, and S. Singer, Gradient, high-resolution, magic angle spinning  $^1\text{H}$  nuclear magnetic resonance spectroscopy of intact cells, *Magnetic Resonance in Medicine*, 39 (1998) 337-345.
- [34] J. M. Wieruszeski, A. Bohin, J. P. Bohin, and G. Lippens, In vivo detection of the cyclic osmoregulated periplasmic glucan of *Ralstonia solanacearum* by high-

- resolution magic angle spinning NMR, *Journal of Magnetic Resonance*, 151 (2001) 118-123.
- [35] M. S. Chauton, O. I. Optun, T. F. Bathen, Z. Volent, I. S. Gribbestad, and G. Johnsen, HR-MAS  $^1\text{H}$  NMR spectroscopy analysis of marine microalgal whole cells, *Marine Ecology-Progress Series*, 256 (2003) 57-62.
  - [36] L. L. Cheng, M. J. Ma, L. Becerra, T. Ptak, I. Tracey, A. Lackner, and R. G. Gonzalez, Quantitative neuropathology by high resolution magic angle spinning proton magnetic resonance spectroscopy, *Proceedings of the National Academy of Sciences of the United States of America*, 94 (1997) 6408-6413.
  - [37] M. C. Martinez-Bisbal, L. Marti-Bonmati, J. Piquer, A. Revert, P. Ferrer, J. L. Llacer, M. Piotto, O. Assemat, and B. Celda,  $^1\text{H}$  and  $^{13}\text{C}$  HR-MAS spectroscopy of intact biopsy samples ex vivo and in vivo  $^1\text{H}$  MRS study of human high grade gliomas, *NMR in Biomedicine*, 17 (2004) 191-205.
  - [38] D. Moka, R. Vorreuther, H. Schicha, M. Spraul, E. Humpfer, M. Lipinski, P. J. D. Foxall, J. K. Nicholson, and J. C. Lindon, Magic angle spinning proton nuclear magnetic resonance spectroscopic analysis of intact kidney tissue samples, *Analytical Communications*, 34 (1997) 107-109.
  - [39] L. L. Cheng, C. L. Wu, M. R. Smith, and R. G. Gonzalez, Non-destructive quantitation of spermine in human prostate tissue samples using HR-MAS  $^1\text{H}$  NMR spectroscopy at 9.4 T, *Febs Letters*, 494 (2001) 112-116.
  - [40] J. L. Taylor, C. L. Wu, D. Cory, R. G. Gonzalez, A. Bielecki, and L. L. Cheng, High-resolution magic angle spinning proton NMR analysis of human prostate tissue with slow spinning rates, *Magnetic Resonance in Medicine*, 50 (2003) 627-632.
  - [41] C. L. Wu, J. L. Taylor, W. L. He, A. G. Zepeda, E. F. Halpern, A. Bielecki, R. G. Gonzalez, and L. L. Cheng, Proton high-resolution magic angle spinning NMR analysis of fresh and previously frozen tissue of human prostate, *Magnetic Resonance in Medicine*, 50 (2003) 1307-1311.
  - [42] S. J. Barton, F. A. Howe, A. M. Tomlins, S. A. Cudlip, J. K. Nicholson, B. A. Bell, and J. L. Griffin, Comparison of in vivo  $^1\text{H}$  MRS of human brain tumours with  $^1\text{H}$  HR-MAS spectroscopy of intact biopsy samples in vitro, *Magnetic Resonance Materials in Physics, Biology and Medicine*, 8 (1999) 121-128.
  - [43] F. Desmoulin, D. Bon, R. Martino, and M. Malet-Martino, HR-MAS NMR analysis of biological tissues: a critical report, *Comptes Rendus Chimie*, 11 (2008) 423-433.
  - [44] A. Pines, M. G. Gibby, and J. S. Waugh, Proton-enhanced NMR of dilute spins in solids, *Journal of Chemical Physics*, 59 (1973) 569-590.

- [45] A. Pines, J. S. Waugh, and M. G. Gibby, Proton-enhanced nuclear induction spectroscopy - method for high-resolution NMR of dilute spins in solids, *Journal of Chemical Physics*, 56 (1972) 1776-1777.
- [46] M. J. Gidley and S. M. Bociek,  $^{13}\text{C}$  CP/MAS NMR-Studies of Amylose Inclusion Complexes, Cyclodextrins, and the Amorphous phase of starch granules - Relationships Between Glycosidic Linkage Conformation and Solid-State  $^{13}\text{C}$  Chemical-Shifts, *Journal of the American Chemical Society*, 110 (1988) 3820-3829.
- [47] K. R. Morgan, R. H. Furneaux, and R. A. Stanley, Observation by Solid-State  $^{13}\text{C}$  CP MAS NMR-Spectroscopy of the Transformations of Wheat-Starch Associated with the Making and Staling of Bread, *Carbohydrate Research*, 235 (1992) 15-22.
- [48] K. R. Morgan, R. H. Furneaux, and N. G. Larsen, Solid-State NMR-Studies on the Structure of Starch Granules, *Carbohydrate Research*, 276 (1995) 387-399.
- [49] W. R. Morrison, R. V. Law, and C. E. Snape, Evidence for Inclusion Complexes of Lipids with V-Amylose in Maize, Rice and Oat Starches, *Journal of Cereal Science*, 18 (1993) 107-109.
- [50] L. Pizzoferrato, M. Paci, and G. Rotilio, Structural modification and bioavailability of starch components as related to the extent of Maillard reaction: An enzymatic degradation and a solid-state  $^{13}\text{C}$  CPMAS NMR Study, *Journal of Agricultural and Food Chemistry*, 46 (1998) 438-441.
- [51] L. Pizzoferrato, G. Rotilio, and M. Paci, Modification of structure and digestibility of chestnut starch upon cooking: A solid state  $^{13}\text{C}$  CP MAS NMR and enzymatic degradation study, *Journal of Agricultural and Food Chemistry*, 47 (1999) 4060-4063.
- [52] V. Singh, S. Z. Ali, and S. Divakar,  $^{13}\text{C}$  CP MAS NMR-Spectroscopy of Native and Acid Modified Starches, *Starch-Starke*, 45 (1993) 59-62.
- [53] H. C. Bertram, H. Stodkilde-Jorgensen, A. H. Karlsson, and H. J. Andersen, Post mortem energy metabolism and meat quality of porcine M-longissimus dorsi as influenced by stunning method - A  $^{31}\text{P}$  NMR spectroscopic study, *Meat Science*, 62 (2002) 113-119.
- [54] L. Calucci, C. Forte, L. Galleschi, M. Geppi, and S. Ghiringhelli,  $^{13}\text{C}$  and  $^1\text{H}$  solid state NMR investigation of hydration effects on gluten dynamics, *International Journal of Biological Macromolecules*, 32 (2003) 179-189.
- [55] W. C. Hutton, J. R. Garbow, and T. R. Hayes, Nondestructive NMR determination of oil composition in transformed Canola seeds, *Lipids*, 34 (1999) 1339-1346.
- [56] C. M. G. C. Renard and M. C. Jarvis, A cross-polarization, magic-angle-spinning,  $^{13}\text{C}$ -nuclear-magnetic-resonance study of polysaccharides in sugar beet cell walls, *Plant Physiology*, 119 (1999) 1315-1322.

- [57] J. K. M. Sanders and B. Hunter, *Modern NMR spectroscopy, A guide for chemists*, Oxford University Press, Oxford 2008.
- [58] M. L. Liu, X. A. Mao, C. H. Ye, H. Huang, J. K. Nicholson, and J. C. Lindon, Improved WATERGATE pulse sequences for solvent suppression in NMR spectroscopy, *Journal of Magnetic Resonance*, 132 (1998) 125-129.
- [59] M. Piotto, V. Saudek, and V. Sklenar, Gradient-Tailored Excitation for Single-Quantum NMR-Spectroscopy of Aqueous-Solutions, *Journal of Biomolecular NMR*, 2 (1992) 661-665.
- [60] S. A. Fancy, O. Beckonert, G. Darbon, W. Yabsley, R. Walley, D. Baker, G. L. Perkins, F. S. Pullen, and K. Rumpel, Gas chromatography/flame ionisation detection mass spectrometry for the detection of endogenous urine metabolites for metabonomic studies and its use as a complementary tool to nuclear magnetic resonance spectroscopy, *Rapid Communications in Mass Spectrometry*, 20 (2006) 2271-2280.
- [61] T. M. Alam and M. K. Alam, Chemometric analysis of NMR spectroscopy data: A review, *Annual Reports on NMR Spectroscopy*, 54 (2005) 41-80.
- [62] R. Stoyanova and T. R. Brown, NMR spectral quantitation by principal component analysis - III. A generalized procedure for determination of lineshape variations, *Journal of Magnetic Resonance*, 154 (2002) 163-175.
- [63] T. R. Brown and R. Stoyanova, NMR spectral quantitation by principal-component analysis. II. Determination of frequency and phase shifts, *Journal of Magnetic Resonance Series B*, 112 (1996) 32-43.
- [64] H. Witjes, W. J. Melssen, H. J. A. I. Zandt, M. van der Graaf, A. Heerschap, and L. M. C. Buydens, Automatic correction for phase shifts, frequency shifts, and lineshape distortions across a series of single resonance lines in large spectral data sets, *Journal of Magnetic Resonance*, 144 (2000) 35-44.
- [65] R. J. Ober and E. S. Ward, Correcting for Phase Distortion of NMR Spectra Analyzed Using Singular Value Decomposition of Hankel-Matrices, *Journal of Magnetic Resonance Series A*, 114 (1995) 120-123.
- [66] H. Witjes, M. Pepers, W. J. Melssen, and L. M. C. Buydens, Modelling phase shifts, peak shifts and peak width variations in spectral data sets: its value in multivariate data analysis, *Analytica Chimica Acta*, 432 (2001) 113-124.
- [67] J. T. M. Pearce, T. J. Athersuch, T. M. D. Ebbels, J. C. Lindon, J. K. Nicholson, and H. C. Keun, Robust algorithms for automated chemical shift calibration of 1D <sup>1</sup>H NMR spectra of blood serum, *Analytical Chemistry*, 80 (2008) 7158-7162.
- [68] T. Skov, F. van den Berg, G. Tomasi, and R. Bro, Automated alignment of chromatographic data, *Journal of Chemometrics*, 20 (2006) 484-497.

- [69] D. Bylund, R. Danielsson, G. Malmquist, and K. E. Markides, Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data, *Journal of Chromatography A*, 961 (2002) 237-244.
- [70] N. P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, *Journal of Chromatography A*, 805 (1998) 17-35.
- [71] G. Tomasi, F. van den Berg, and C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, *Journal of Chemometrics*, 18 (2004) 231-241.
- [72] F. van den Berg, G. Tomasi, and N. Viereck, Warping: Investigation of NMR pre-processing and correction, in: S. B. Engelsen, P. S. Belton, and H. J. Jakobsen (Eds.), *Magnetic Resonance in Food Science, The Multivariate Challenge*. The Royal Society of Chemistry, Cambridge, UK, 2005, pp. 131-138.
- [73] J. Forshed, I. Schuppe-Koistinen, and S. P. Jacobsson, Peak alignment of NMR signals by means of a genetic algorithm, *Analytica Chimica Acta*, 487 (2003) 189-199.
- [74] G. C. Lee and D. L. Woodruff, Beam search for peak alignment of NMR signals, *Analytica Chimica Acta*, 513 (2004) 413-416.
- [75] W. Wu, M. Daszykowski, B. Walczak, B. C. Sweatman, S. C. Connor, J. N. Haseldeo, D. J. Crowther, R. W. Gill, and M. W. Lutz, Peak alignment of urine NMR spectra using fuzzy warping, *Journal of Chemical Information and Modeling*, 46 (2006) 863-875.
- [76] K. A. Veselkov, J. C. Lindon, T. M. D. Ebbels, D. Crockford, V. V. Volynkin, E. Holmes, D. B. Davies, and J. K. Nicholson, Recursive segment-wise peak alignment of biological <sup>1</sup>H NMR spectra for improved metabolic biomarker recovery, *Analytical Chemistry*, 81 (2009) 56-66.
- [77] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, *BMC Genomics*, 7 (2006) 1-15.
- [78] R. Bro and A. K. Smilde, Centering and scaling in component analysis, *Journal of Chemometrics*, 17 (2003) 16-33.
- [79] A. K. Smilde, M. J. van der Werf, S. Bijlsma, B. van der Werff-van-der Vat, and R. H. Jellema, Fusion of mass spectrometry-based metabolomics data, *Analytical Chemistry*, 77 (2005) 6729-6736.

- [80] L. Eriksson, E. Johansson, N. Kettaneh-Wold, and S. Wold, Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS), Scaling. Umetrics, 1999, pp. 213-225.
- [81] H. C. Keun, T. M. D. Ebbels, H. Antti, M. E. Bollard, O. Beckonert, E. Holmes, J. C. Lindon, and J. K. Nicholson, Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling, *Analytica Chimica Acta*, 490 (2003) 265-276.
- [82] A. Craig, O. Cloareo, E. Holmes, J. K. Nicholson, and J. C. Lindon, Scaling and normalization effects in NMR spectroscopic metabonomic data sets, *Analytical Chemistry*, 78 (2006) 2262-2267.
- [83] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn, Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in  $^1\text{H}$  NMR metabonomics, *Analytical Chemistry*, 78 (2006) 4281-4290.
- [84] R. J. O. Torgrip, K. M. Aberg, E. Alm, I. Schuppe-Koistinen, and J. Lindberg, A note on normalization of biofluid 1D  $^1\text{H}$  NMR data, *Metabolomics*, 4 (2008) 114-121.
- [85] E. Holmes, P. J. D. Foxall, J. K. Nicholson, G. H. Neild, S. M. Brown, C. R. Beddell, B. C. Sweatman, E. Rahr, J. C. Lindon, M. Spraul, and P. Neidig, Automatic Data Reduction and Pattern-Recognition Methods for Analysis of  $^1\text{H}$  Nuclear-Magnetic-Resonance Spectra of Human Urine from Normal and Pathological States, *Analytical Biochemistry*, 220 (1994) 284-296.
- [86] U. Holzgrabe, R. Deubner, C. Schollmayer, and B. Waibel, Quantitative NMR spectroscopy - Applications in drug analysis, *Journal of Pharmaceutical and Biomedical Analysis*, 38 (2005) 806-812.
- [87] D. S. Wishart, M. J. Lewis, J. A. Morrissey, M. D. Flegel, K. Jeroncic, Y. P. Xiong, D. Cheng, R. Eisner, B. Gautam, D. Tzur, S. Sawhney, F. Bamforth, R. Greiner, and L. Li, The human cerebrospinal fluid metabolome, *Journal of Chromatography B- Analytical Technologies in the Biomedical and Life Sciences*, 871 (2008) 164-173.
- [88] D. S. Wishart, Metabolomics: applications to food science and nutrition research, *Trends in Food Science & Technology*, 19 (2008) 482-493.
- [89] I. Berregi, G. Del Campo, R. Caracena, and J. I. Miranda, Quantitative determination of formic acid in apple juices by H-1 NMR spectrometry, *Talanta*, 72 (2007) 1049-1053.
- [90] M. A. Brescia, A. Sgaramella, S. Ghelli, and A. Sacco,  $^1\text{H}$  HR-MAS NMR and isotopic investigation of bread and flour samples produced in southern Italy, *Journal of the Science of Food and Agriculture*, 83 (2003) 1463-1468.



- [91] L. Shintu, F. Ziarelli, and S. Caldarelli, Is high-resolution magic angle spinning NMR a practical speciation tool for cheese samples? Parmigiano Reggiano as a case study, *Magnetic Resonance in Chemistry*, 42 (2004) 396-401.
- [92] L. Shintu and S. Caldarelli, Toward the determination of the geographical origin of emmental(er) cheese via high resolution MAS NMR: A preliminary investigation, *Journal of Agricultural and Food Chemistry*, 54 (2006) 4148-4154.
- [93] M. A. Brescia, M. Monfreda, A. Buccolieri, and C. Carrino, Characterisation of the geographical origin of buffalo milk and mozzarella cheese by means of analytical and spectroscopic determinations, *Food Chemistry*, 89 (2005) 139-147.
- [94] M. A. Brescia, V. Mazzilli, A. Sgaramella, S. Ghelli, F. P. Fanizzi, and A. Sacco,  $^1\text{H}$  NMR characterization of milk lipids: A comparison between cow and buffalo milk, *Journal of the American Oil Chemists Society*, 81 (2004) 431-436.
- [95] R. Lamanna, I. Piscioneri, V. Romanelli, and N. Sharma, A preliminary study of soft cheese degradation in different packaging conditions by  $^1\text{H}$  NMR, *Magnetic Resonance in Chemistry*, 46 (2008) 828-831.
- [96] E. Schievano, G. Pasini, G. Cozzi, and S. Mammi, Identification of the production chain of Asiago d'Allevio cheese by nuclear magnetic resonance spectroscopy and principal component analysis, *Journal of Agricultural and Food Chemistry*, 56 (2008) 7208-7214.
- [97] G. E. Pereira, J. P. Gaudillere, C. Van Leeuwen, G. Hilbert, O. Lavialle, M. Maucourt, C. Deborde, A. Moing, and D. Rolin,  $^1\text{H}$  NMR and chemometrics to characterize mature grape berries in four wine-growing areas in Bordeaux, France, *Journal of Agricultural and Food Chemistry*, 53 (2005) 6382-6389.
- [98] M. A. Brescia, I. J. Kosir, V. Caldarola, J. Kidric, and A. Sacco, Chemometric classification of Apulian and Slovenian wines using  $^1\text{H}$  NMR and ICP-OES together with HPICE data, *Journal of Agricultural and Food Chemistry*, 51 (2003) 21-26.
- [99] G. E. Pereira, J. P. Gaudillere, C. Van Leeuwen, G. Hilbert, M. Maucourt, C. Deborde, A. Moing, and D. Rolin,  $^1\text{H}$  NMR metabolite fingerprints of grape berry: Comparison of vintage and soil effects in Bordeaux grapevine growing areas, *Analytica Chimica Acta*, 563 (2006) 346-352.
- [100] C. Almeida, I. F. Duarte, A. Barros, J. Rodrigues, M. Spraul, and A. M. Gil, Composition of beer by  $^1\text{H}$  NMR spectroscopy: Effects of brewing site and date of production, *Journal of Agricultural and Food Chemistry*, 54 (2006) 700-706.
- [101] A. M. Gil and J. Rodrigues, High resolution NMR tools for the analysis of beer and wine, in: I. A. Farhat, P. S. Belton, and G. A. Webb (Eds.), *Magnetic Resonance in Food Science, From Molecules to Man*. The Royal Society of Chemistry, Cambridge, UK, 2007, pp. 96-104.

- [102] G. Del Campo, I. Berregi, R. Caracena, and J. I. Santos, Quantitative analysis of malic and citric acids in fruit juices using proton nuclear magnetic resonance spectroscopy, *Analytica Chimica Acta*, 556 (2006) 462-468.
- [103] M. Cuny, E. Vigneau, G. Le Gall, I. Colquhoun, M. Lees, and D. N. Rutledge, Fruit juice authentication by  $^1\text{H}$  NMR spectroscopy in combination with different chemometrics tools, *Analytical and Bioanalytical Chemistry*, 390 (2008) 419-427.
- [104] G. Le Gall, M. Puaud, and I. J. Colquhoun, Discrimination between orange juice and pulp wash by  $^1\text{H}$  nuclear magnetic resonance spectroscopy: Identification of marker compounds, *Journal of Agricultural and Food Chemistry*, 49 (2001) 580-588.
- [105] J. T. W. E. Vogels, L. Terwel, A. C. Tas, F. van den Berg, F. Dukel, and J. van der Greef, Detection of adulteration in orange juices by a new screening method using proton NMR spectroscopy in combination with pattern recognition techniques, *Journal of Agricultural and Food Chemistry*, 44 (1996) 175-180.
- [106] G. Del Campo, J. I. Santos, N. Iturriza, I. Berregi, and A. Munduate, Use of the  $^1\text{H}$  nuclear magnetic resonance spectra signals from polyphenols and acids for chemometric characterization of cider apple juices, *Journal of Agricultural and Food Chemistry*, 54 (2006) 3095-3100.
- [107] A. J. Charlton, W. H. H. Farrington, and P. Brereton, Application of  $^1\text{H}$  NMR and multivariate statistics for screening complex mixtures: Quality control and authenticity of instant coffee, *Journal of Agricultural and Food Chemistry*, 50 (2002) 3098-3103.
- [108] D. L. Garcia-Gonzalez, L. Mannina, M. D'Imperio, A. L. Segre, and R. Aparicio, Using  $^1\text{H}$  and  $^{13}\text{C}$  NMR techniques and artificial neural networks to detect the adulteration of olive oil with hazelnut oil, *European Food Research and Technology*, 219 (2004) 545-548.
- [109] K. Zur, A. Heier, K. W. Blaas, and C. Fauhl-Hassek, Authenticity control of pistachios based on  $^1\text{H}$ - and  $^{13}\text{C}$ -NMR spectroscopy and multivariate statistics, *European Food Research and Technology*, 227 (2008) 969-977.
- [110] C. E. Tyl, L. Brecker, and K. H. Wagner,  $^1\text{H}$  NMR spectroscopy as tool to follow changes in the fatty acids of fish oils, *European Journal of Lipid Science and Technology*, 110 (2008) 141-148.
- [111] S. Masoum, C. Malabat, M. Jalali-Heravi, C. Guillou, S. Rezzi, and D. N. Rutledge, Application of support vector machines to  $^1\text{H}$  NMR data of fish oils: methodology for the confirmation of wild and farmed salmon and their origins, *Analytical and Bioanalytical Chemistry*, 387 (2007) 1499-1510.

- [112] A. Caligiani, D. Acquotti, G. Palla, and V. Bocchi, Identification and quantification of the main organic components of vinegars by high resolution  $^1\text{H}$  NMR spectroscopy, *Analytica Chimica Acta*, 585 (2007) 110-119.
- [113] M. Cuny, E. Vigneau, M. Lees, and D. N. Rutledge, Dairy product authentication by  $^1\text{H}$  NMR spectroscopy in combination with different chemometric tools, in: I. A. Farhat, P. S. Belton, and G. A. Webb (Eds.), *Magnetic Resonance in Food Science, From Molecules to Man*. The Royal Society of Chemistry, Cambridge, UK, 2007, pp. 197-204.
- [114] A. K. Thybo, P. M. Szczypinski, A. H. Karlsson, S. Donstrup, H. S. Stodkilde-Jorgensen, and H. J. Andersen, Prediction of sensory texture quality attributes of cooked potatoes by NMR-imaging (MRI) of raw potatoes in combination with different image analysis methods, *Journal of Food Engineering*, 61 (2004) 91-100.
- [115] F. Savorani, F. Capozzi, S. B. Engelsen, M. T. Dell'Abate, and P. Sequi, Pomodoro di Pachino: An authentication study using  $^1\text{H}$  NMR and chemometrics – Protecting its P.G.I. European certification, in: P. S. Belton and G. A. Webb (Eds.), *Magnetic Resonance in Food Science, Challenges in a Changing World*. The Royal Society of Chemistry, Cambridge, UK, 2009. In press.
- [116] S. Y. S. Tu, Y. J. Choi, M. J. McCarthy, and K. L. McCarthy, Tomato quality evaluation by peak force and NMR spin-spin relaxation time, *Postharvest Biology and Technology*, 44 (2007) 157-164.
- [117] S. Tiziani, S. J. Schwartz, and Y. Vodovotz, Profiling of carotenoids in tomato juice by one- and two-dimensional NMR, *Journal of Agricultural and Food Chemistry*, 54 (2006) 6094-6100.
- [118] M. C. Todasca, S. Zarbock-Udrea, C. Deleanu, and S. Rosca, Rhubarb's (*Rheum rhabarbarum*) stalk study through  $^1\text{H}$  NMR spectroscopy method, *Revista de Chimie*, 57 (2006) 1019-1021.
- [119] O. Beckonert, H. C. Keun, T. M. D. Ebbels, J. Bundy, E. Holmes, J. C. Lindon, and J. K. Nicholson, Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts, *Nature Protocols*, 2 (2007) 2692-2703.
- [120] J. P. Renou, S. Benderbous, G. Bielicki, L. Foucat, and J. P. Donnat,  $^{23}\text{Na}$  Magnetic-Resonance-Imaging - Distribution of Brine in Muscle, *Magnetic Resonance Imaging*, 12 (1994) 131-137.
- [121] L. Foucat, J. P. Donnat, and J. P. Renou,  $^{23}\text{Na}$  and  $^{35}\text{Cl}$  NMR studies of the interactions of sodium and chloride ions with meat products, in: P. S. Belton, A. M. Gil, G. A. Webb, and D. N. Rutledge (Eds.), *Magnetic Resonance in Food Science, Latest Developments*. The Royal Society of Chemistry, Cambridge, UK, 2003, pp. 180-185.

- [122] C. Vestergaard, J. Risum, and J. Adler-Nissen,  $^{23}\text{Na}$  MRI quantification of sodium and water mobility in pork during brine curing, *Meat Science*, 69 (2005) 663-672.
- [123] U. Erikson, E. Veliyulin, T. E. Singstad, and M. Aursand, Salting and desalting of fresh and frozen-thawed cod (*Gadus morhua*) fillets: A comparative study using  $^{23}\text{Na}$  NMR,  $^{23}\text{Na}$  MRI, low-field  $^1\text{H}$  NMR, and physicochemical analytical methods, *Journal of Food Science*, 69 (2004) E107-E114.
- [124] N. Ishida, T. Kobayashi, H. Kano, S. Nagai, and H. Ogawa,  $^{23}\text{Na}$  NMR Imaging of Foods, *Agricultural and Biological Chemistry*, 55 (1991) 2195-2200.
- [125] T. R. Rosett, S. L. Kendregan, Y. Gao, S. J. Schmidt, and B. P. Klein, Thickening agents effects on sodium binding and other taste qualities of soup systems, *Journal of Food Science*, 61 (1996) 1099-1104.
- [126] A. Spyros and P. Dais, Application of  $^{31}\text{P}$  NMR spectroscopy in food analysis. 1. Quantitative determination of the mono- and diglyceride composition of olive oils, *Journal of Agricultural and Food Chemistry*, 48 (2000) 802-805.
- [127] C. P. Li, A. S. Salvador, H. R. Ibrahim, Y. Sugimoto, and T. Aoki, Phosphorylation of egg white proteins by dry-heating in the presence of phosphate, *Journal of Agricultural and Food Chemistry*, 51 (2003) 6808-6815.
- [128] M. M. Nielsen, N. Viereck, and S. B. Engelsen, Phytic acid degradation by phytase as viewed by  $^{31}\text{P}$  NMR and multivariate curve resolution, in: I. A. Farhat, P. S. Belton, and G. A. Webb (Eds.), *Magnetic Resonance in Food Science, From Molecules to Man*. The Royal Society of Chemistry, Cambridge, UK, 2007, pp. 214-222.
- [129] J. B. Lambert and E. P. Mazzola, Two-dimensional NMR, in: J. B. Lambert and E. P. Mazzola (Eds.), *Nuclear magnetic resonance spectroscopy, An introduction to principles, applications, and experimental methods*. Pearson Education Inc., New Jersey, USA, 2004, pp. 172-232.
- [130] G. J. Martin, M. L. Martin, and B. L. Zhang, Site-Specific Natural Isotope Fractionation of Hydrogen in Plant-Products Studied by Nuclear-Magnetic-Resonance, *Plant Cell and Environment*, 15 (1992) 1037-1050.
- [131] B. L. Zhang, S. Buddrus, M. Trierweiler, and G. J. Martin, Characterization of glycerol from different origins by  $^2\text{H}$ - and  $^{13}\text{C}$ -NMR studies of site-specific natural isotope fractionation, *Journal of Agricultural and Food Chemistry*, 46 (1998) 1374-1380.
- [132] A. Royer, N. Naulet, F. Mabon, M. Lees, and G. J. Martin, Stable isotope characterization of olive oils. II - Deuterium distribution in fatty acids studied by nuclear magnetic resonance, *Journal of the American Oil Chemists Society*, 76 (1999) 365-373.

- [133] M. Aursand, F. Mabon, and G. J. Martin, Characterization of farmed and wild salmon (*Salmo salar*) by a combined use of compositional and isotopic analyses, *Journal of the American Oil Chemists Society*, 77 (2000) 659-666.
- [134] M. Aursand, F. Mabon, and G. J. Martin, High-resolution  $^1\text{H}$  and  $^2\text{H}$  NMR spectroscopy of pure essential fatty acids for plants and animals, *Magnetic Resonance in Chemistry*, 35 (1997) S91-S100.
- [135] Z. Masud, C. Vallet, and G. J. Martin, Stable isotope characterization of milk components and whey ethanol, *Journal of Agricultural and Food Chemistry*, 47 (1999) 4693-4699.
- [136] D. Danho, N. Naulet, and G. J. Martin, Deuterium, Carbon and Nitrogen Isotopic Analysis of Natural and Synthetic Caffeines - Authentication of Coffees and Coffee Extracts, *Analytical and Bioanalytical Chemistry*, 20 (1992) 179-184.
- [137] G. Remaud, H. McManus, Y. L. Martin, and G. G. Martin, Application of SNIF-NMR in determining the authenticity of maple syrup, *Abstracts of Papers of the American Chemical Society*, 212 (1996) 23-AGFD.
- [138] E. F. Boffo, L. A. Tavares, A. G. Ferreira, M. M. C. Ferreira, and A. C. T. Tobias, Adulteration study in Brazilian honey by SNIF and  $^1\text{H}$  NMR, in: I. A. Farhat, P. S. Belton, and G. A. Webb (Eds.), *Magnetic Resonance in Food Science, From Molecules to Man*. The Royal Society of Chemistry, Cambridge, UK, 2007, pp. 105-113.
- [139] H. Martens, S. B. Engelsen, and F. van den Berg, Too much data -too little information!, *The Parliament Magazine*, (2003) 76.
- [140] H. Hotelling, Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, 24 (1933) 417-441.
- [141] R. Tauler, Multivariate curve resolution applied to second order data, *Chemometrics and Intelligent Laboratory Systems*, 30 (1995) 133-146.
- [142] R. Manne, On the Resolution Problem in Hyphenated Chromatography, *Chemometrics and Intelligent Laboratory Systems*, 27 (1995) 89-94.
- [143] R. Huo, C. Geurts, J. Brands, R. Wehrens, and L. M. C. Buydens, Real-life applications of the MULVADO software package for processing DOSY NMR data, *Magnetic Resonance in Chemistry*, 44 (2006) 110-117.
- [144] R. Bro, PARAFAC. Tutorial and applications, *Chemometrics and Intelligent Laboratory Systems*, 38 (1997) 149-171.
- [145] J. D. Carroll and J. J. Chang, Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition, *Psychometrika*, 35 (1970) 630-638.

- [146] R. A. Harshman, Foundation of the PARAFAC procedure: Models and conditions for an 'exploratory' multi-modal factor analysis, *UCLA Working Papers in Phonetics*, 16 (1970) 1-84.
- [147] M. Dyrby, M. Petersen, A. K. Whittaker, L. Lambert, L. Nørgaard, R. Bro, and S. B. Engelsen, Analysis of lipoproteins using 2D diffusion-edited NMR spectroscopy and multi-way chemometrics, *Analytica Chimica Acta*, 531 (2005) 209-216.
- [148] S. Wold, H. Martens, and H. Wold, The Multivariate Calibration-Problem in Chemistry Solved by the PLS Method, *Lecture Notes in Mathematics*, 973 (1983) 286-293.
- [149] L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, and S. B. Engelsen, Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy, *Applied Spectroscopy*, 54 (2000) 413-419.
- [150] R. Leardi and L. Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, *Journal of Chemometrics*, 18 (2004) 486-497.
- [151] J. van der Greef and A. K. Smilde, Symbiosis of chemometrics and metabolomics: past, present, and future, *Journal of Chemometrics*, 19 (2005) 376-386.
- [152] N. M. Faber and R. Rajko, How to avoid over-fitting in multivariate calibration - The conventional validation approach and an alternative, *Analytica Chimica Acta*, 595 (2007) 98-106.
- [153] D. I. Broadhurst and D. B. Kell, Statistical strategies for avoiding false discoveries in metabolomics and related experiments, *Metabolomics*, 2 (2006) 171-196.
- [154] M. Fujiwara, I. Ando, and K. Arifuku, Multivariate analysis for  $^1\text{H}$  NMR spectra of two hundred kinds of tea in the world, *Analytical Sciences*, 22 (2006) 1307-1314.
- [155] L. Nørgaard, R. Bro, F. Westad, and S. B. Engelsen, A modification of canonical variates analysis to handle highly collinear multivariate data, *Journal of Chemometrics*, 20 (2006) 425-435.
- [156] W. J. Krzanowski, *Principles of Multivariate Analysis: A User's Perspective*, Oxford Science Publication, New York 2000.
- [157] C. R. Rao, *Advanced Statistical Methods in Biometric Research*, Wiley, New York 1952.
- [158] J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. van Velzen, J. P. M. van Duijnhoven, and F. A. van Dorsten, Assessment of PLS-DA cross validation, *Metabolomics*, 4 (2008) 81-89.

- [159] H. C. Bertram, N. B. Kristensen, A. Malmendal, N. C. Nielsen, R. Bro, H. J. Andersen, and D. L. Harmon, A metabolomic investigation of splanchnic metabolism using  $^1\text{H}$  NMR spectroscopy of bovine blood plasma, *Analytica Chimica Acta*, 536 (2005) 1-6.
- [160] M. A. Constantinou, E. Papakonstantinou, M. Spraul, S. Sevastiadou, C. Costalos, M. A. Koupparis, K. Shulpis, A. Tsantili-Kakoulidou, and E. Mikros,  $^1\text{H}$  NMR-based metabonomics for the diagnosis of inborn errors of metabolism in urine, *Analytica Chimica Acta*, 542 (2005) 169-177.
- [161] H. L. Mao, M. Xu, B. Wang, H. M. Wang, X. M. Deng, and D. H. Lin, Evaluation of filtering effects of orthogonal signal correction on metabonomic analysis of healthy human serum  $^1\text{H}$  NMR spectra, *Acta Chimica Sinica*, 65 (2007) 152-158.
- [162] Z. Ramadan, D. Jacobs, M. Grigorov, and S. Kochhar, Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms, *Talanta*, 68 (2006) 1683-1691.
- [163] Z. Ramadan, P. F. Zhang, D. M. Jacobs, I. Tavazzi, and S. Kochhar, An NMR- and MS-based metabonomic investigation of saliva metabolic changes in feline odontoclastic resorptive lesions (FORL)-diseased cats, *Metabolomics*, 3 (2007) 113-119.
- [164] C. Stella, B. Beckwith-Hall, O. Cloarec, E. Holmes, J. C. Lindon, J. Powell, F. van der Ouderaa, S. Bingham, A. J. Cross, and J. K. Nicholson, Susceptibility of human metabolic phenotypes to dietary modulation, *Journal of Proteome Research*, 5 (2006) 2780-2788.
- [165] S. Wagner, K. Scholz, M. Sieber, M. Kellert, and W. Voelkel, Tools in metabonomics: An integrated validation approach for LC-MS metabolic profiling of mercapturic acids in human urine, *Analytical Chemistry*, 79 (2007) 2918-2926.
- [166] C. Wang, H. W. Kong, Y. F. Guan, J. Yang, J. R. Gu, S. L. Yang, and G. W. Xu, Plasma phospholipid metabolic profiling and biomarkers of type 2 diabetes mellitus based on high-performance liquid chromatography/electrospray mass spectrometry and multivariate statistical analysis, *Analytical Chemistry*, 77 (2005) 4108-4116.
- [167] M. van Doorn, J. Vogels, A. Tas, E. J. van Hoogdalem, J. Burggraaf, A. Cohen, and J. van der Greef, Evaluation of metabolite profiles as biomarkers for the pharmacological effects of thiazolidinediones in Type 2 diabetes mellitus patients and healthy volunteers, *British Journal of Clinical Pharmacology*, 63 (2007) 562-574.
- [168] A. D. Maher, S. F. M. Zirah, E. Holmes, and J. K. Nicholson, Experimental and analytical variation in human urine in  $^1\text{H}$  NMR spectroscopy-based metabolic phenotyping studies, *Analytical Chemistry*, 79 (2007) 5204-5211.

- [169] S. Kochhar, D. M. Jacobs, Z. Ramadan, F. Berruex, A. Fuerhoz, and L. B. Fay, Probing gender-specific metabolism differences in humans by nuclear magnetic resonance-based metabolomics, *Analytical Biochemistry*, 352 (2006) 274-281.
- [170] E. K. Kemsley, G. Le Gall, J. R. Dainty, A. D. Watson, L. J. Harvey, H. S. Tapp, and I. J. Colquhoun, Multivariate techniques and their application in nutrition: a metabolomics case study, *British Journal of Nutrition*, 98 (2007) 1-14.
- [171] C. A. Andersson, Direct orthogonalization, *Chemometrics and Intelligent Laboratory Systems*, 47 (1999) 51-63.
- [172] H. R. Tang, Y. L. Wang, J. K. Nicholson, and J. C. Lindon, Use of relaxation-edited one-dimensional and two dimensional nuclear magnetic resonance spectroscopy to improve detection of small metabolites in blood plasma, *Analytical Biochemistry*, 325 (2004) 260-272.
- [173] W. J. Yang, Y. W. Wang, Q. F. Zhou, and H. R. Tang, Analysis of human urine metabolites using SPE and NMR spectroscopy, *Science in China Series B-Chemistry*, 51 (2008) 218-225.
- [174] H. C. Keun, T. M. D. Ebbels, H. Antti, M. E. Bollard, O. Beckonert, G. Schlotterbeck, H. Senn, U. Niederhauser, E. Holmes, J. C. Lindon, and J. K. Nicholson, Analytical reproducibility in  $^1\text{H}$  NMR-based metabolomic urinalysis, *Chemical Research in Toxicology*, 15 (2002) 1380-1386.
- [175] FDA, United States Food and Drug Administration, Guidance for industry PAT -A framework for innovative pharmaceutical development, manufacturing and quality assurance, U. S. Department of Health and Human Services, Rockville, (2004).
- [176] M. A. Bernstein, Nuclear magnetic resonance: Stepping up to the PAT challenge, *Spectroscopy Europe*, 19 (2007).
- [177] B. P. Hills and K. M. Wright, Towards on-line NMR sensors, in: S. B. Engelsen, P. S. Belton, and H. J. Jakobsen (Eds.), *Magnetic Resonance in Food Science, The Multivariate Challenge*. The Royal Society of Chemistry, Cambridge, UK, 2005, pp. 175-185.
- [178] L. Shintu, S. Caldarelli, and B. M. Franke, Pre-selection of potential molecular markers for the geographic origin of dried beef by HR-MAS NMR spectroscopy, *Meat Science*, 76 (2007) 700-707.
- [179] A. Sacco, I. N. Bolsi, R. Massini, M. Spraul, E. Humpfer, and S. Ghelli, Preliminary investigation on the characterization of durum wheat flours coming from some areas of south Italy by means of  $^1\text{H}$  high-resolution magic angle spinning nuclear magnetic resonance, *Journal of Agricultural and Food Chemistry*, 46 (1998) 4242-4249.



- [180] A. M. Gil, I. F. Duarte, I. Delgadillo, I. J. Colquhoun, F. Casuscelli, E. Humpfer, and M. Spraul, Study of the compositional changes of mango during ripening by use of nuclear magnetic resonance spectroscopy, *Journal of Agricultural and Food Chemistry*, 48 (2000) 1524-1536.
- [181] L. Shintu and S. Caldarelli, High-resolution MAS NMR and chemometrics: Characterization of the ripening of Parmigiano Reggiano cheese, *Journal of Agricultural and Food Chemistry*, 53 (2005) 4026-4031.
- [182] R. Consonni and L. R. Cagliani, Ripening and geographical characterization of Parmigiano Reggiano cheese by  $^1\text{H}$  NMR spectroscopy, *Talanta*, 76 (2008) 200-205.
- [183] P. D. Jamieson, P. J. Stone, and M. A. Semenov, Towards modeling quality in wheat - from grain nitrogen concentration to protein composition, *Aspects of Applied Biology*, 64 (2001) 111-126.
- [184] J. R. Raven, L. Handely, and B. Wollenweber, Plant nutrition and water use efficiency, in: M. Bacon (Ed.), *Water Use Efficiency in Plant Biology*. Blackwell Science Publications, 2004, pp. 162-188.
- [185] H. F. Seefeldt, F. H. Larsen, N. Viereck, B. Wollenweber, and S. B. Engelsen, Bulk carbohydrate grain filling of barley beta-glucan mutants studied by  $^1\text{H}$  HR MAS NMR, *Cereal Chemistry*, 85 (2008) 571-577.
- [186] N. Viereck, H. Winning, H. F. Seefeldt, and S. B. Engelsen,  $^1\text{H}$  HR MAS NMR: Profiling metabolites in single cereal kernels (of wheat and barley), in: M. Guðjónsdóttir, P. S. Belton, and G. A. Webb (Eds.), *Magnetic Resonance in Food Science, Challenges in a Changing World*. The Royal Society of Chemistry, Cambridge, UK, 2009. In press.
- [187] C. Castro and C. Manetti, A multiway approach to analyze metabonomic data: a study of maize seeds development, *Analytical Biochemistry*, 371 (2007) 194-200.
- [188] J. J. Jansen, R. Bro, H. C. J. Hoefsloot, F. W. J. van den Berg, J. A. Westerhuis, and A. K. Smilde, PARAFASCA: ASCA combined with PARAFAC for the analysis of metabolic fingerprinting data, *Journal of Chemometrics*, 22 (2008) 114-121.
- [189] H. C. Keun, T. M. D. Ebbels, M. E. Bollard, O. Beckonert, H. Antti, E. Holmes, J. C. Lindon, and J. K. Nicholson, Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles, *Chemical Research in Toxicology*, 17 (2004) 579-587.
- [190] E. J. Saude, D. Adamko, B. H. Rowe, T. Marrie, and B. D. Sykes, Variation of metabolites in normal human urine, *Metabolomics*, 3 (2007) 439-451.

- [191] Q. Zhang, G. J. Wang, Y. Du, L. L. Zhu, and A. Jiye, GC/MS analysis of the rat urine for metabonomic research, *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences*, 854 (2007) 20-25.
- [192] M. R. Viant, C. Ludwig, S. Rhodes, U. L. Guenther, and D. Allaway, Validation of a urine metabolome fingerprint in dog for phenotypic classification, *Metabolomics*, 3 (2007) 453-463.
- [193] C. M. Slupsky, K. N. Rankin, J. Wagner, H. Fu, D. Chang, A. M. Weljie, E. J. Saude, B. Lix, D. J. Adamko, S. Shah, R. Greiner, B. D. Sykes, and T. J. Marrie, Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles, *Analytical Chemistry*, 79 (2007) 6995-7004.
- [194] T. Kind, V. Tolstikov, O. Fiehn, and R. H. Weiss, A comprehensive urinary metabolomic approach for identifying kidney cancer, *Analytical Biochemistry*, 363 (2007) 185-195.
- [195] M. Lauridsen, S. H. Hansen, J. W. Jaroszewski, and C. Cornett, Human urine as test material in  $^1\text{H}$  NMR-based metabonomics: Recommendations for sample preparation and storage, *Analytical Chemistry*, 79 (2007) 1181-1186.
- [196] M. C. Walsh, L. Brennan, J. P. G. Malthouse, H. M. Roche, and M. J. Gibney, Effect of acute dietary standardization on the urinary, plasma, and salivary metabolomic profiles of healthy humans, *American Journal of Clinical Nutrition*, 84 (2006) 531-539.
- [197] E. J. J. van Velzen, J. A. Westerhuis, J. P. M. Van Duynhoven, F. A. van Dorsten, H. C. J. Hoefsloot, D. M. Jacobs, S. Smit, R. Draijer, C. I. Kroner, and A. K. Smilde, Multilevel data analysis of a crossover designed human nutritional intervention study, *Journal of Proteome Research*, 7 (2008) 4483-4491.
- [198] F. A. van Dorsten, C. A. Daykin, T. P. J. Mulder, and J. P. M. Van Duynhoven, Metabonomics approach to determine metabolic differences between green tea and black tea consumption, *Journal of Agricultural and Food Chemistry*, 54 (2006) 6929-6938.
- [199] C. A. Daykin, J. P. M. Van Duynhoven, A. Groenewegen, M. Dachtler, J. M. M. Van Amelsvoort, and T. P. J. Mulder, Nuclear magnetic resonance spectroscopic based studies of the metabolism of black tea polyphenols in humans, *Journal of Agricultural and Food Chemistry*, 53 (2005) 1428-1434.
- [200] T. P. Mulder, A. G. Rietveld, and J. M. van Amelsvoort, Consumption of both black tea and green tea results in an increase in the excretion of hippuric acid into urine, *American Journal of Clinical Nutrition*, 81 (2005) 256S-260S.
- [201] K. S. Solanky, N. J. C. Bailey, E. Holmes, J. C. Lindon, A. L. Davis, T. P. J. Mulder, J. P. M. Van Duynhoven, and J. K. Nicholson, NMR-based metabonomic studies on the

biochemical effects of epicatechin in the rat, *Journal of Agricultural and Food Chemistry*, 51 (2003) 4139-4145.

- [202] K. S. Solanky, N. J. Bailey, B. M. Beckwith-Hall, S. Bingham, A. Davis, E. Holmes, J. K. Nicholson, and A. Cassidy, Biofluid <sup>1</sup>H NMR-based metabonomic techniques in nutrition research metabolic effects of dietary isoflavones in humans, *Journal of Nutritional Biochemistry*, 16 (2005) 236-244.
- [203] Y. L. Wang, H. R. Tang, J. K. Nicholson, P. J. Hylands, J. Sampson, and E. Holmes, A metabonomic strategy for the detection of the metabolic effects of chamomile (*Matricaria recutita* L.) ingestion, *Journal of Agricultural and Food Chemistry*, 53 (2005) 191-196.
- [204] S. Rezzi, Z. Ramadan, L. B. Fay, and S. Kochhar, Nutritional metabonomics: Applications and perspectives, *Journal of Proteome Research*, 6 (2007) 513-525.
- [205] J. K. Nicholson, E. Holmes, J. C. Lindon, and I. D. Wilson, The challenges of modeling mammalian biocomplexity, *Nature Biotechnology*, 22 (2004) 1268-1274.
- [206] M. A. van Baak, T. M. Larsen, S. A. Jebb, A. Kafatos, A. Pfeiffer, J. A. Martinez, S. Handjiev, M. Kunesova, A. Astrup, and W. H. M. Saris, Weight loss maintenance on ad libitum diets varying in protein content and glycemic index: First results of the DIOGENES highly-controlled shop-based intervention, [www. diogenes-eu. org](http://www.diogenes-eu.org), (2008).
- [207] M. Assfalg, I. Bertini, D. Colangiuli, C. Luchinat, H. Schafer, B. Schutz, and M. Spraul, Evidence of different metabolic phenotypes in humans, *Proceedings of the National Academy of Sciences of the United States of America*, 105 (2008) 1420-1424.
- [208] W. Willett, *Nutritional epidemiology*, Oxford University Press, Oxford, United Kingdom 1998.
- [209] N. M. McKeown, N. E. Day, A. A. Welch, S. A. Runswick, R. N. Luben, A. A. Mulligan, A. McTaggart, and S. A. Bingham, Use of biological markers to validate self-reported dietary intake in a random sample of the European Prospective Investigation into Cancer United Kingdom Norfolk cohort, *American Journal of Clinical Nutrition*, 74 (2001) 188-196.
- [210] R. J. Kaaks, Biochemical markers as additional measurements in studies of the accuracy of dietary questionnaire measurements: Conceptual issues, *American Journal of Clinical Nutrition*, 65 (1997) 1232S-1239S.
- [211] M. Harker, H. Coulson, I. Fairweather, D. Taylor, and C. A. Daykin, Study of metabolite composition of eccrine sweat from healthy male and female human subjects by <sup>1</sup>H NMR spectroscopy, *Metabolomics*, 2 (2006) 105-112.

- [212] C. Rolin, Pectin, in: R. L. Whistler and J. N. BeMiller (Eds.), *Industrial Gums: polysaccharides and their derivatives*. Academic Press, San Diego, 1993, pp. 257-293.
- [213] M. C. Ralet, E. Bonnin, and J. F. Thibault, Pectins, in: E. J. Vandamme, S. De Baets, and A. Steinbüchel (Eds.), *Polysaccharides II, Biopolymers*. WILEY-VCH Verlag GmbH, Weinheim, 2002.
- [214] M. Marounek, Z. Volek, A. Synytsya, and J. Copikova, Effect of pectin and amidated pectin on cholesterol homeostasis and cecal metabolism in rats fed a high-cholesterol diet, *Physiological Research*, 56 (2007) 433-442.
- [215] A. T. Hotchkiss, E. Olano-Martin, W. E. Grace, G. R. Gibson, and R. A. Rastall, Pectic oligosaccharides as prebiotics, *Oligosaccharides in Food and Agriculture*, 849 (2003) 54-62.
- [216] N. C. Carpita and D. M. Gibeaut, Structural Models of Primary Cell Walls in Flowering Plants Consistency of Molecular Structure with the Physical Properties of the Walls During Growth, *Plant Journal*, 3 (1993) 1-30.
- [217] B. L. Ridley, M. A. O'Neill, and D. A. Mohnen, Pectins: structure, biosynthesis, and oligogalacturonide-related signaling, *Phytochemistry*, 57 (2001) 929-967.
- [218] M. C. Ralet, M. J. Crepeau, H. C. Buchholt, and J. F. Thibault, Polyelectrolyte behaviour and calcium binding properties of sugar beet pectins differing in their degrees of methylation and acetylation, *Biochemical Engineering Journal*, 16 (2003) 191-201.
- [219] M. C. Ralet, V. Dronnet, H. C. Buchholt, and J. F. Thibault, Enzymatically and chemically de-esterified lime pectins: characterisation, polyelectrolyte behaviour and calcium binding properties, *Carbohydrate Research*, 336 (2001) 117-125.
- [220] I. Braccini, R. P. Grasso, and S. Perez, Conformational and configurational features of acidic polysaccharides and their interactions with calcium ions: a molecular modeling investigation, *Carbohydrate Research*, 317 (1999) 119-130.
- [221] H. Grasdalen, A. K. Andersen, and B. Larsen, NMR spectroscopy studies of the action pattern of tomato pectinesterase: Generation of block structure in pectin by a multiple-attack mechanism, *Carbohydrate Research*, 289 (1996) 105-114.
- [222] A. K. Andersen, B. Larsen, and H. Grasdalen, Sequential Structure by <sup>1</sup>H NMR As A Direct Assay for Pectinesterase Activity, *Carbohydrate Research*, 273 (1995) 93-98.
- [223] S. B. Engelsen and L. Nørgaard, Comparative vibrational spectroscopy for determination of quality parameters in amidated pectins as evaluated by chemometrics, *Carbohydrate Polymers*, 30 (1996) 9-24.

- [224] S. B. Engelsen, E. Mikkelsen, and L. Munck, New Approaches to Rapid Spectroscopic Evaluation of Properties in Pectic Polymers, *Progress in Colloid and Polymer Science*, (1998) 166-174.
- [225] H. Winning, N. Viereck, T. Salomonsen, J. Larsen, and S. B. Engelsen, Quantification of blockiness in pectins -A comparative study using vibrational spectroscopy and chemometrics, *Carbohydrate Research*, In press (2009).
- [226] P. L. Hubbard, P. J. Watkinson, L. K. Creamer, A. Gottwald, and P. T. Callaghan, Two-dimensional laplace inversion NMR technique applied to the molecular properties of water in dry-salted Mozzarella-type cheeses with various salt concentrations, *Magnetic Resonance in Food Science, The Multivariate Challenge*. The Royal Society of Chemistry, Cambridge, UK, 2005, pp. 225-232.
- [227] T. Salomonsen, H. M. Jensen, D. Stenbaek, and S. B. Engelsen, Chemometric prediction of alginate monomer composition: A comparative spectroscopic study using IR, Raman, NIR and NMR, *Carbohydrate Polymers*, 72 (2008) 730-739.



## Paper I

*Quantitative analysis of NMR spectra with chemometrics*

H. Winning, F.H. Larsen, R. Bro, S.B. Engelsen

Journal of Magnetic Resonance **190** (2008) 26–32.

## Quantitative analysis of NMR spectra with chemometrics

H. Winning<sup>\*</sup>, F.H. Larsen, R. Bro, S.B. Engelsen

*Quality and Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark*

Received 11 April 2007; revised 8 October 2007

Available online 14 October 2007

### Abstract

The number of applications of chemometrics to series of NMR spectra is rapidly increasing due to an emerging interest for quantitative NMR spectroscopy e.g. in the pharmaceutical and food industries.

This paper gives an analysis of advantages and limitations of applying the two most common chemometric procedures, Principal Component Analysis (PCA) and Multivariate Curve Resolution (MCR), to a designed set of 231 simple alcohol mixture (propanol, butanol and pentanol) <sup>1</sup>H 400 MHz spectra. The study clearly demonstrates that the major advantage of chemometrics is the visualisation of larger data structures which adds a new exploratory dimension to NMR research. While robustness and powerful data visualisation and exploration are the main qualities of the PCA method, the study demonstrates that the bilinear MCR method is an even more powerful method for resolving pure component NMR spectra from mixtures when certain conditions are met.

© 2007 Elsevier Inc. All rights reserved.

**Keywords:** Principal component analysis; Multivariate curve resolution; Experimental design; Signal overlap

### 1. Introduction

NMR is a unique and versatile spectroscopic method capable of measuring samples in the solid, liquid and gas phases. No other spectroscopic method contains equally detailed structural and dynamic information about chemical systems under investigation. However, a serious challenge in NMR spectroscopy lies between the technical capacity to generate data (such as in NMR metabonomics) and the human capacity to interpret and integrate these data [1]. In complex systems such as biofluids, a wide range of components (metabolites, acids, proteins, carbohydrates, etc.) are present with a majority of overlapping resonances distributed over several thousand data points [2]. This amount of data is difficult, if not impossible, to interpret.

The study of more complex systems, such as biofluids is characterised by many hidden relationships. To find these hidden relationships in complex data, experimental design, unsupervised data exploration and data mining techniques

are required. Chemometrics is a multivariate data analysis field using statistics to compute models for extracting chemical information from large two-dimensional multivariate data sets. Development of chemometric data models requires a minimum of assumptions and the relationships may be visualised by intuitive illustrations by the graphic computer interface.

We have chosen a ternary model design with three simple linear water soluble alcohols containing different amounts of hydrocarbons with highly overlapping resonances. Using this design we can explore subtle differences in the methylene peak—a simplified simulation of one of the major metabolomic applications of NMR, namely lipoprotein profiling of blood. Besides lipid and lipoprotein resonances, the 0.7–1.5 ppm chemical shift region in blood plasma is characterised by many overlapping signals from small organic species [3]. Spectral assignments in this region have been limited by the extensive chemical shift overlap and by the broadness of the signals. Similar spectral problems may be encountered in organic or pharmaceutical samples when identifying impurities that mimic the compound of interest.

<sup>\*</sup> Corresponding author.

E-mail address: [haw@life.ku.dk](mailto:haw@life.ku.dk) (H. Winning).



The first application of chemometrics to NMR spectra appeared in 1983 by Johnels et al. [4]. In the early nineties Gartland et al. [5] introduced PCA to classify proton NMR spectra of urine. The same group also introduced the research branch *metabonomics* defined as: “understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data” [6] or just “metabolic processes studied by NMR spectroscopy of biofluids” [1]. Impetus for the coupling of NMR spectroscopy with multivariate data analysis was clearly the terribly complex metabolic system in body fluids that gives rise to equally complex NMR spectra. Chemometrics were also applied early on for exploration of solid-state MAS NMR spectra [7]. Now chemometrics is rapidly gaining momentum in the analysis of NMR spectra [8–10] and this work aims to provide an understanding of some of the useful property of basic chemometric tools by their application to a designed set of alcohol mixture NMR spectra.

## 2. Materials and methods

Principal component analysis and multivariate curve resolution

Principal component analysis (PCA) [11] is the fundamental method in chemometrics. In PCA the data collected on a set of samples is resolved into principal components. The first principal component is defined by the spectral profile (loading) in the data which describes most of the variation, the second principal component is the profile describing the second most of the variation orthogonal to the first, and so on. Later components describe less variation and are more uncertain than the first components, because the systematic variation is primarily described in the first components. Deciding the right number of components is a most important issue and will be described for both PCA and MCR although for PCA, the choice is often less critical especially in exploratory studies because the first and most important component will not change as a function of the number of components chosen. The principal components are composed of so-called scores and loadings. Loadings contain information about the variables (chemical shifts) in the data set and the scores hold information on samples (concentrations) in the data set. For a given principal component, the loading vector is a spectral profile and the score for each sample is the amount of that particular loading in the sample in a least squares sense. Thus, the sum of loadings weighted by a certain sample's score values will provide an approximation of the spectrum of that sample. The similarity to Beers law is apparent as each measured spectrum is hence described by varying amounts of the same few underlying spectral loadings. The individual loadings, though, will mostly not resemble real chemical spectra due to orthogonality constraints of the scores and loadings, but the peaks in a loading are indicative of large spectral variation in the data. Thus, the loadings indicate which parts of the spectrum represent

the main variation amongst the samples. The scores, on the other hand, provide information about the extent to which the spectral information represented by the loadings are high or low for particular samples. Hence, the scores can be considered as concentrations of multivariate, so-called *latent*, variables.

As the number of spectral components in a data set is typically much lower than the number of chemical shifts, the whole data set can for the most part be represented by a few (typically much less than 10–20) components that still represent the full chemical variation in the data. The scores are often plotted against each other in a scatter plot giving a ‘map’ of all the samples in the score plot. Samples that are grouped in a score plot are spectrally similar with respect to the selected principal components. One of the strengths of PCA is to provide a quick unsupervised view of the samples and thereby to identify samples that exhibit deviating features (outliers) or discover trends and groups in the samples. Prior to PCA modelling, data are centred by subtracting from every chemical shift value the average value at that particular shift calculated across all samples.

An alternative method to PCA is decomposition of the data matrix by multivariate curve resolution (MCR) using alternating least-squares (ALS) [12] which has also been named molecular factor analysis (MFA) along with a number of other names [13–15]. Principal component analysis is mostly used for exploration and classification and cannot normally provide direct estimates of real chemical spectra and concentrations because the loadings and scores are constrained to be orthogonal. MCR-ALS on the contrary can offer resolution of the spectra into the ‘true’ underlying components, i.e. the pure spectra. Huo et al. have proved that multivariate curve resolution was able to provide unique pure spectra and pure decay profiles from DOSY NMR data [16].

An appealing property of MCR is that the solution often looks much more ‘chemical’ than a PCA solution, because imposed non-negativity constraints make the spectral and sample profiles be positive. This often leads to oversimplifying interpretations where the solution is assumed to be real estimates of chemical spectra and their relative concentrations. However, the MCR solutions are generally not unique, hence the solution can be assumed to be just *one* arbitrary solution out of an infinity of equally well-fitting possible nonnegative solutions. The problem is due to the so-called rotational ambiguity [17] and even though imposing non-negativity helps removing some ambiguity it is not enough to guarantee uniqueness in general. This can only be achieved if the data have certain characteristics such as selective variables where only some analytes are present or samples where some analytes are absent [17].

Hence, for any specific MCR solution, uniqueness must be assessed before the solution can be assumed to be providing estimates of real chemical analytes. Uniqueness can be assessed in different ways, but in this investigation the model was simply restarted several times using different

sets of random initial parameters and was verified to provide the same solution. There are other tools for multivariate analysis with a similar aim as MCR, such as direct exponential curve resolution algorithm (DECRA) [18] but these are not applicable to the type of matrix-data (2D) discussed here.

### 2.1. NMR data

The experimental design is a ternary design of mixtures of the linear alcohols: propanol, butanol and pentanol [13,19]. Each alcohol component (50 mM) has 21 concentration levels in increments of five from 0% to 100%. The samples were prepared from 495  $\mu$ l of the mixture and 55  $\mu$ l of D<sub>2</sub>O (with 5.8 mM of TSP-*d*<sub>4</sub> (per-deuterated 3-trimethylsilyl propionate sodium salt) (Fig. 1).

<sup>1</sup>H NMR spectra were recorded for each of the 231 mixtures. The spectra were acquired on a Bruker Avance Ultra Shield 400 spectrometer (Bruker Biospin GmbH, Rheinstetten, Germany) operating at 400.13 MHz using a broad band inverse detection probe head equipped with 5 mm (o.d.) NMR sample tubes. Data were accumulated at 298 K employing a pulse sequence using presaturation of the water resonance during the recycle period followed by a composite 90° pulse [20] with an acquisition time of 4 s, a recycle delay of 20 s, eight scans and a sweep width of 8278.15 Hz, resulting in 64k complex data points. All samples were individually tuned, matched and shimmed. Prior to Fourier transformation, each FID was apodised by Lorentzian line broadening of 0.30 Hz and the corresponding spectra were automatically phased and baseline corrected and referenced to TSP-*d*<sub>4</sub>. In order to secure quantitative measurements the receiver gain was set constant for all the samples.

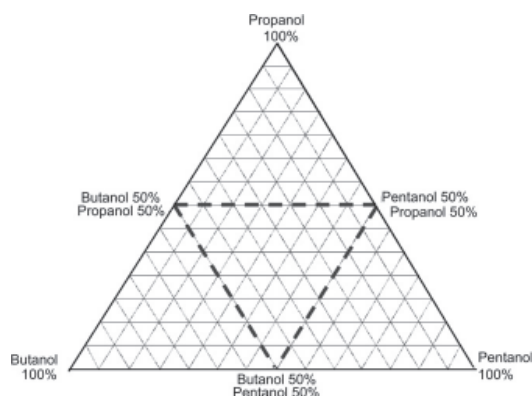


Fig. 1. Tri-axial experimental design of propanol, butanol and pentanol. Each alcohol component has 21 different levels in increments of 5 from 0% to 100%, 231 samples in total. The corners of the triangle represent 100% of the pure alcohol. The triangle with the dashed line shows the reduced experimental design of 66 samples.

Prior to the chemometric analysis the raw proton NMR spectra data matrix to be investigated had the dimensions (231 × 65,536) but was reduced to 14,000 data points (3.85–0.65 ppm) in order to remove the water signal and make the investigation more efficient. No further pre-processing or alignment of the data such as co-shifting and warping [21] proved necessary.

The NMR spectra of the 231 alcohol mixtures results in only four specific signals (Fig. 2). The spectrum of pure propanol yields a triplet at 0.90 ppm from the CH<sub>3</sub>, a quintet at 1.55 ppm from CH<sub>2</sub> and a triplet at 3.57 from the CH<sub>2</sub> next to the OH group. Similar assignments apply to butanol and pentanol, but they also contain aliphatic CH<sub>2</sub>s with chemical shift in the range 1.30–1.35 ppm. The spectra of pure propanol, butanol and pentanol are displayed in the bottom of Fig. 2.

### 3. Results

Using PCA the raw 231 <sup>1</sup>H NMR spectra was decomposed into principal components to describe the systematic variation in the spectra. The data is mean centred prior to PCA, which means that the mean spectrum is subtracted from the individual sample spectra. This simple pre-transformation provides spectra that show the deviation from the average spectrum. PCA results in an almost perfect recovery of the ternary experimental design by the two first PCs, as seen by the score plot in Fig. 3.

Fig. 3 shows the scores and the loadings (of component one and two) of the PCA model where the scores are coloured by the propanol content. The first two components together describe 98% of the variation in the spectra. The 2% of the variation that remains to be explained appears non-systematic, hence due to noise. The loadings of the first two components (PC#1 and PC#2) are displayed in the corresponding loading plot. The first loading describes the overall data structure of the NMR spectra which before mean centering of the spectra are similar to the average spectrum. Upon mean centering, the first loading will change to describe the main variation of all centred data. The fact that the scores are negative is due to the centering as well as the imposed orthogonality that also indirectly causes the loadings to be negative. This clearly illustrates that PCA does not provide estimates of real chemical analytes. However it is also clear that the scores are (linearly) related to the true concentrations (compare Fig. 3, left and Fig. 1), and it is also clear that the loadings reflect the underlying spectral variation.

Apparently only two principal components are necessary to describe all the variation in the spectra, but this is due to the principle of closure specific to these data, i.e. that the concentration of any chemical component in a sample is defined by the remaining two because they add up to 100%. The ternary experimental design is reflected in the score plot which reveals the direct proportional signal intensities with analyte concentrations. Had the data

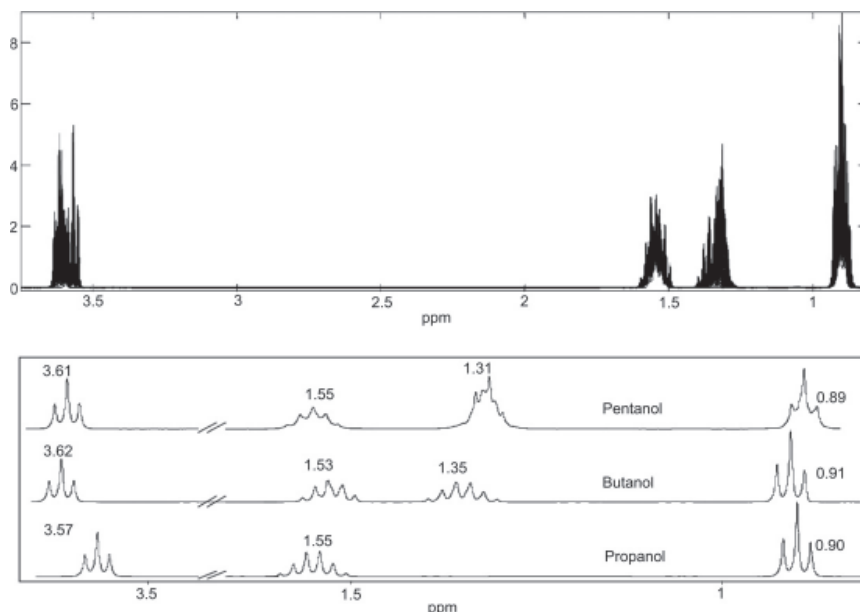


Fig. 2. (Top) NMR spectra of the 231 alcohol mixtures from 3.85 to 0.65 ppm. The NMR spectra of mixtures show highly overlapping signals. (Bottom) The <sup>1</sup>H NMR spectra of the pure alcohol samples of propanol, butanol and pentanol.

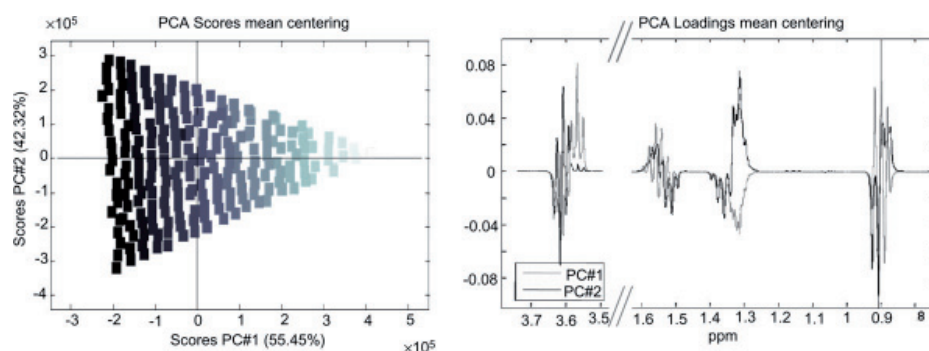


Fig. 3. Scores and loadings plot of the first two principal components from a PCA model calculated on mean-centred NMR spectra. For increased interpretability the score plot is coloured according to the propanol content. The first two principal components explain 97.8% of the variation.

not been centred, three components would be needed to describe the data [22].

In order to pursue the purpose of resolving overlapping resonances from CH<sub>2</sub> and CH<sub>3</sub> groups, the focus of the analysis is restricted to the methyl groups with chemical shifts around 0.9 ppm. As is obvious from Fig. 2, the resonances of the alcohols differ slightly in chemical shift as well as in line width. Our strategy is to perform the chemometric analysis on the restricted data set (0.85–0.95 ppm) representing a region with significant spectral overlap and compare it with results obtained on the full spectrum (3.85–0.65 ppm). The result in Fig. 4 is convincing; the

PCA model recovers the ternary experimental design based on spectra of the methyl groups alone. Still, as evidenced by the loading plot, PCA cannot provide real estimates of the pure analyte spectra and concentrations.

MCR is an alternative multivariate data analytical method that can potentially decompose a data set into pure spectra and concentration profiles. The number of components to be extracted can be assessed by looking at the explained variance as a function of number of components, similar to how the number of components is often determined in PCA. Furthermore, visual interpretation of the results is often used as a practical guide in assessing the

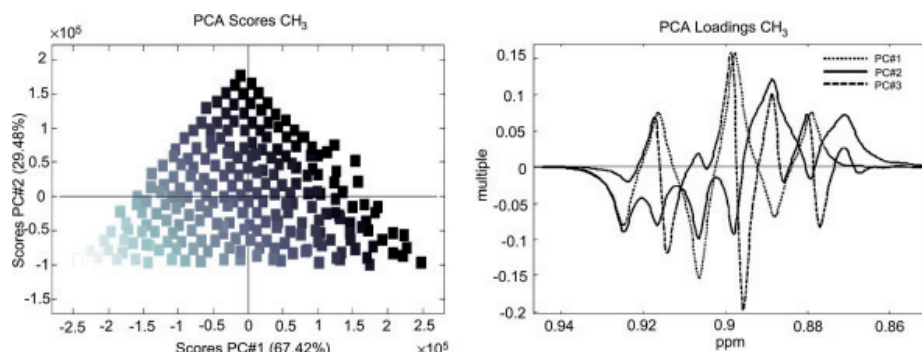


Fig. 4. Scores and loadings plot of the first two principal components from a PCA model calculated on mean-centred NMR spectra (0.85–0.95 ppm). For increased interpretability the score plot is coloured according to the propanol content. The first two principal components explain 96.9% of the variation.

validity of a given model. From Fig. 5 it is obvious that a model with three components is optimal, considering that the variance explained is over 99% and almost remains constant when using more than three components. This is also consistent with the fact that the samples are mixtures of three analytes. That 99% variance explained is adequate can be further assessed and validated by comparing with the intrinsic noise in the data (not shown).

The MCR model with three components is calculated without mean centering the data as is usual in MCR. Looking at a scatter plot of the scores from the MCR model, the triangle now shows perfect concentrations (Fig. 6). The slight non-ideality observed in the triangle, can be attributed to noise in the spectra and small uncertainties in the alcohol concentration. Non-negativity of estimated concentrations and spectra is imposed as part of the model.

The loadings from MCR shown in Fig. 6 resemble spectra of each of the pure alcohol compounds. By repeating the estimation of the MCR model many times from different random starting points, it is verified that the same fit and solution is obtained (results not shown). Hence, the

solution can be assumed to be unique. Like PCA, each sample has a score for each of the loadings. The score is simply the amount of the corresponding loading, and as the loadings can be considered as estimates of real spectra, then the scores are then (relative) estimates of the concentrations. These scores are compared with the 'true' value (i.e. the concentration of the three alcohols) by plotting them against each other, yielding three correlation coefficients higher than 0.99.

The results above are encouraging and imply that complex mixture NMR spectra can be separated mathematically into the underlying constituents. However, the main reason that the results are as good as they are is the presence of pure samples in the sample set. The presence of pure samples adds selectivity in the data. Selectivity means samples or variables for which only one analyte is present. This is one of the key requisites for obtaining uniqueness in MCR.

To demonstrate how well MCR can model more complex data, a model is calculated on a reduced experimental design of only 66 samples where no spectra of pure alcohols exist. The simplest samples in this reduced design consist of mixtures of at least two of the alcohols. The result is that the triangle of the experimental design is fully recovered and the three concentration profiles still yield correlations over 0.99 to the true concentrations.

However, the loading are not as perfectly resolved as in the full design, which is due to the overlap of the signals from the alcohols. This is particularly apparent in the spurious propanol peak in Fig. 7 at 0.8 ppm. Apparently the MCR model determined on this dataset is unique. Repeating the model estimations more than 1000 times from different starting points all lead in 74% of the cases to the same local solution which is spectrally correct (Fig. 7, left). However in 19% of the cases the global model, which explains almost the same variance but is spectrally incorrect, is the result (Fig. 7, right). This result represent the Achilles' heel of MCR when applied to unknown systems. The best model may not be the correct one in the physical sense.

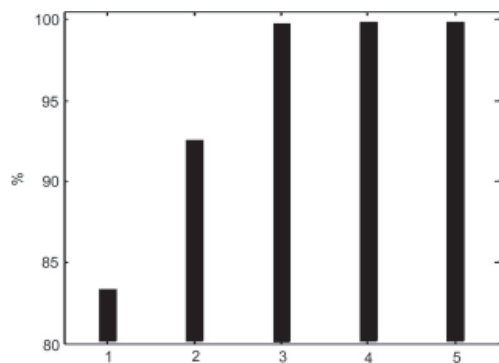


Fig. 5. The bar plot show the percentage explained variance as a function of number of components. Using three components, 99.7% of the variation is explained.

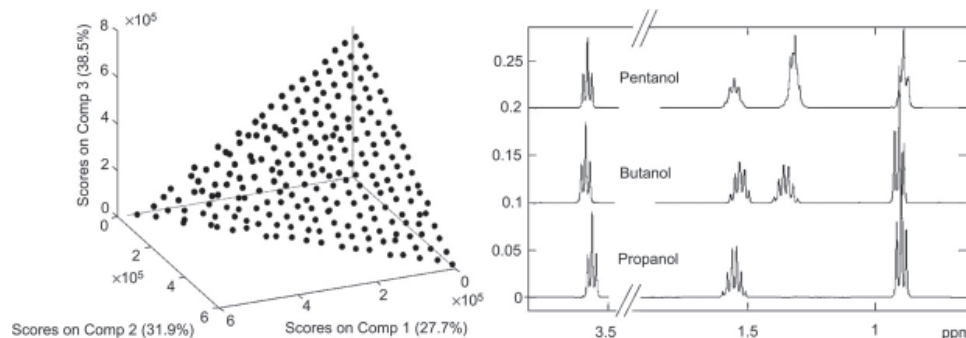


Fig. 6. Scores of the first three components (left) and loading plot (right) of the three components from the MCR model, obtained on the NMR spectra. The two components explain 98.1% of the variation. The score plot shows a perfectly equilateral triangle.

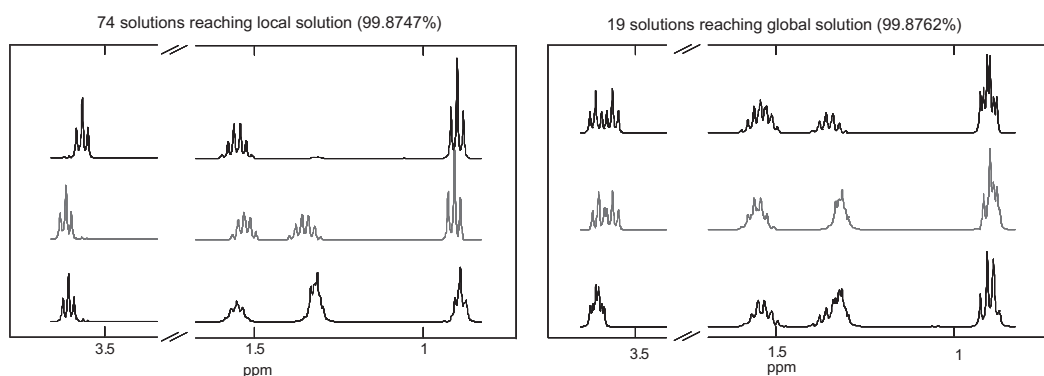


Fig. 7. Plot to the left show the loading of the three components from the local MCR model. This local solution is spectrally correct and is the result in 74% of the cases. The loading plot to the right is obtained from the global model which is the result in 19% of the cases. These loadings are mixtures of the loadings of the three pure alcohols.

#### 4. Conclusions

The main objective of this work was to show how principal component analysis and multivariate curve resolution can be useful in the investigation of highly overlapping data from NMR studies. While it has been shown that PCA can be used to provide a comprehensive overview of complex data with many variables, it was also shown that there are some limits on the usefulness of PCA. The MCR method was demonstrated to possess the powerful ability to separate mixtures into pure spectra and concentrations even for much reduced designs. By applying the basic chemometric methods to a well defined ternary experimental design of  $^1\text{H}$  NMR spectra the potential and characteristics of chemometric multivariate data analysis were demonstrated. It should be obvious that perhaps the greatest advantage of chemometrics is the simplicity by which even large data structures are analysed and visualised and thereby adding an exploratory dimension to modern NMR science. We have shown which results can be expected when applying quantitative chemometric methods

to multivariate high resolution NMR data and our future research will focus on how MCR can perform on more complex metabonomic data.

#### Acknowledgments

The Ministry of Food, Agriculture and Fisheries is greatly acknowledged for financial support to the ASYNCRON project. Gilda Kischinsky is acknowledged for proof reading.

#### References

- [1] J.C. Lindon, J.K. Nicholson, E. Holmes, J.R. Everett, Metabonomics: metabolic processes studied by NMR spectroscopy of biofluids, *Concepts Magn. Reson.* 12 (2000) 289–320.
- [2] J.L. Griffin, The potential of metabonomics in drug safety and toxicology, *Drug Discov. Today Technol.* 1 (2004) 285–293.
- [3] J.D. Bell, P.J. Sadler, A.F. Macleod, P.R. Turner, A. Laville,  $^1\text{H}$ -NMR studies of human-blood plasma assignment of resonances for lipoproteins, *FEBS Lett.* 219 (1987) 239–243.

- [4] D. Johnels, U. Edlund, H. Grahn, S. Hellberg, M. Sjostrom, S. Wold, S. Clementi, W.J. Dunn, Clustering of aryl C-13 nuclear magnetic-resonance substituent chemical-shifts—a multivariate data-analysis using principal components, *J. Chem. Soc., Perkin Trans. 2* (1983) 863–871.
- [5] K.P.R. Gartland, C.R. Beddell, J.C. Lindon, J.K. Nicholson, Application of pattern-recognition methods to the analysis and classification of toxicological data derived from proton nuclear-magnetic-resonance spectroscopy of urine, *Mol. Pharmacol.* 39 (1991) 629–642.
- [6] J.K. Nicholson, J.C. Lindon, E. Holmes, ‘Metabonomics’: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data, *Xenobiotica* 29 (1999) 1181–1189.
- [7] J.M. Koons, P.D. Ellis, Applicability of factor-analysis in solid-state NMR, *Anal. Chem.* 67 (1995) 4309–4315.
- [8] H.C. Keun, T.M.D. Ebbels, H. Antti, M.E. Bollard, O. Beckonert, E. Holmes, J.C. Lindon, J.K. Nicholson, Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling, *Anal. Chim. Acta* 490 (2003) 265–276.
- [9] S. Halouska, R. Powers, Negative impact of noise on the principal component analysis of NMR data, *J. Magn. Reson.* 178 (2006) 88–95.
- [10] R. Stoyanova, T.R. Brown, NMR spectral quantitation by principal component analysis, *NMR Biomed.* 14 (2001) 271–277.
- [11] H. Hotelling, Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.* 24 (1933) 417–441.
- [12] R. Tauler, Multivariate curve resolution applied to second order data, *Chemometrics Intel. Lab. Syst.* 30 (1995) 133–146.
- [13] M.K. Alam, T.M. Alam, Multivariate analysis and quantitation of O-17-nuclear magnetic resonance in primary alcohol mixtures, *Spectrochim. Acta A-Mol. Biomol. Spectrosc.* 56 (2000) 729–738.
- [14] C.D. Eads, C.M. Furnish, I. Noda, K.D. Juhlin, D.A. Cooper, S.W. Morrall, Molecular factor analysis applied to collections of NMR spectra, *Anal. Chem.* 76 (2004) 1982–1990.
- [15] T.M. Alam, M.K. Alam, Chemometric analysis of NMR spectroscopy data: a review, *Ann. Rep. NMR Spectrosc.* 54 (2005) 41–80.
- [16] R. Huo, C. Geurts, J. Brands, R. Wehrens, L.M.C. Buydens, Real-life applications of the MULVADO software package for processing DOSY NMR data, *Magn. Reson. Chem.* 44 (2006) 110–117.
- [17] R. Manne, On the resolution problem in hyphenated chromatography, *Chemometrics Intel. Lab. Syst.* 27 (1995) 89–94.
- [18] W. Windig, B. Antalek, L.J. Sorriero, S. Bijlsma, D.J. Louwerse, A.K. Smilde, Applications and new developments of the direct exponential curve resolution algorithm (DECRA). Examples of spectra and magnetic resonance images, *J. Chemometrics* 13 (1999) 95–110.
- [19] T. Næs, T. Isaksson, The chemometric space: Interpreting principal components in NIR spectroscopy, *NIR News* 3 (1992) 7.
- [20] A. Bax, A spatially selective composite 90-degrees radiofrequency pulse, *J. Magn. Reson.* 65 (1985) 142–145.
- [21] F.H. Larsen, F. Berg, S.B. Engelsen, An exploratory chemometric study of <sup>1</sup>H NMR spectra of table wines, *J. Chemometrics* 20 (2006) 198–208.
- [22] R. Bro, A.K. Smilde, Centering and scaling in component analysis, *J. Chemometrics* 17 (2003) 16–33.



## Paper II

*Exploring abiotic stress on asynchronous protein metabolism in single kernels of wheat studied by NMR spectroscopy and chemometrics*

H. Winning, N. Viereck, B. Wollenweber, F. H. Larsen, S. Jacobsen, I. Søndergaard, S. B. Engelsen,

Journal of Experimental Botany **1** (2009) 291-300





RESEARCH PAPER

## Exploring abiotic stress on asynchronous protein metabolism in single kernels of wheat studied by NMR spectroscopy and chemometrics

H. Winning<sup>1,\*</sup>, N. Viereck<sup>1</sup>, B. Wollenweber<sup>2</sup>, F. H. Larsen<sup>1</sup>, S. Jacobsen<sup>3</sup>, I. Søndergaard<sup>3</sup> and S. B. Engelsen<sup>1</sup>

<sup>1</sup> University of Copenhagen, Faculty of Life Sciences, Department of Food Science, Quality and Technology, Rolighedsvej 30, 1958 Frederiksberg C, Denmark

<sup>2</sup> University of Aarhus, Faculty of Agricultural Sciences, Department of Genetics and Biotechnology, Forsøgsvej 1, 4200 Slagelse, Denmark

<sup>3</sup> Technical University of Denmark, Enzyme and Protein Chemistry, Department of Systems Biology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

Received 19 August 2008; Revised 14 October 2008; Accepted 24 October 2008

### Abstract

Extreme climate events are being recognized as important factors in the effects on crop growth and yield. Increased climatic variability leads to more frequent extreme conditions which may result in crops being exposed to more than one extreme event within a growing season. The aim of this study was to examine the implications of different drought treatments on the protein fractions in grains of winter wheat using <sup>1</sup>H nuclear magnetic resonance spectroscopy followed by chemometric analysis. *Triticum aestivum* L. cv. Vinjett was studied in a semi-field experiment and subjected to drought episodes either at terminal spikelet, during grain-filling or at both stages. Principal component trajectories of the total protein content and the protein fractions of flour as well as the <sup>1</sup>H NMR spectra of single wheat kernels, wheat flour, and wheat methanol extracts were analysed to elucidate the metabolic development during grain-filling. The results from both the <sup>1</sup>H NMR spectra of methanol extracts and the <sup>1</sup>H HR-MAS NMR of single kernels showed that a single drought event during the generative stage had as strong an influence on protein metabolism as two consecutive events of drought. By contrast, a drought event at the vegetative growth stage had little effect on the parameters investigated. For the first time, <sup>1</sup>H HR-MAS NMR spectra of grains taken during grain-filling were analysed by an advanced multiway model. In addition to the results from the chemical protein analysis and the <sup>1</sup>H HR-MAS NMR spectra of single kernels indicating that protein metabolism is influenced by multiple drought events, the <sup>1</sup>H NMR spectra of the methanol extracts of flour from mature grains revealed that the amount of fumaric acid is particularly sensitive to water deficits.

**Key words:** Chemometrics, drought, fumaric acid, grain-filling, HR-MAS NMR, PARAFAC, PCA trajectory, single kernel, wheat.

### Introduction

Increased climatic variability leading to more frequent extreme conditions may result in crops being exposed to more than one extreme event in a single growing season. As with temperature, variability in drought can occur through variation in its timing, intensity, and duration (IPCC, 2007). Due to limited water resources, drought has become the single most limiting factor to crop production worldwide (Wollenweber *et al.*, 2003). In the last decade, severe

\* To whom correspondence should be addressed: E-mail: [haw@life.ku.dk](mailto:haw@life.ku.dk)

Abbreviations: DAA, days after anthesis; CT, control (no drought); TD, early drought (drought at double ridge); TA, late drought (drought at anthesis); T2, double drought (drought at double ridge and drought at anthesis); NMR, nuclear magnetic resonance; HR, high resolution; MAS, magic angle spinning; TSP-d4, perdeuterated 3-trimethylsilyl propionate sodium salt; COSY, correlation spectroscopy; TOCSY, total correlation spectroscopy; HSQC, heteronuclear single quantum coherence; ANOVA, analysis of variance; PC, principal component; PCA, principal component analysis; PARAFAC, parallel factor analysis.

© 2009 The Author(s).

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

droughts leading to significant yield losses have become a major problem in parts of Europe. The overall prediction is that climate change will affect yield quantity (higher grain yield due to higher dry matter), but potentially at the expense of yield quality.

Besides starch, proteins are the most important components of wheat grains governing end-use quality. However, the quantity of protein per grain is mainly under environmental control. The deposition of the various protein fractions takes place asynchronously, which means that both the amount and concentration of these components vary throughout the maturity period of the grains (Martre *et al.*, 2003). One consequence of this is that conditions shortening grain-filling, such as episodes of high temperature and/or water deficits, will affect the balance of protein fractions (Jamieson *et al.*, 2001). With these considerations in mind, it is important both to understand the environmental constraints on crop quality and to predict how quality will be influenced by the interaction of multiple extreme climatic conditions (Raven *et al.*, 2004).

To date, there are only very few studies on the effect of environmental variability on the quantitative variation in crop protein composition (Martre *et al.*, 2003; Triboi *et al.*, 2003; Wollenweber *et al.*, 2003). The developmental stage of crops experiencing stress events will determine the degree of possible damage experienced by the plant. It has recently been shown that extreme heat events at the vegetative growth stage of double ridge does not affect subsequent growth and development of wheat exposed to heat stress at anthesis as well, while the development of fertile grains is affected by high temperature episodes (Martre *et al.*, 2003; Triboi *et al.*, 2003; Wollenweber *et al.*, 2003).

Nuclear magnetic resonance (NMR) spectroscopy is a powerful non-invasive analytical technique for measuring multiple parameters of plant tissue *in vivo* (Krishnan *et al.*, 2005).  $^1\text{H}$  NMR can be used as a single analytical method to obtain information about the vast majority of metabolites in a plant system, since all proton-bearing compounds such as carbohydrates, amino acids, organic and fatty acids, and lipids can be simultaneously detected. However, a challenge in  $^1\text{H}$  NMR spectroscopy of plants is the solid nature of plant tissue. In solids, anisotropic interactions like homo- and heteronuclear dipole-dipole couplings as well as chemical-shielding anisotropy induce significant line broadening in the NMR spectra. These effects may be eliminated in  $^1\text{H}$  NMR spectra of soft materials (high mobility) by the use of high resolution magic angle spinning (HR-MAS) and spectral resolution similar to liquid-state NMR is obtained.  $^1\text{H}$  HR-MAS NMR spectroscopy has already been used in metabonomic studies of animal and human tissue without the need of chemical extraction or further sample preparation (Garrod *et al.*, 2001; Lindon *et al.*, 2001; Griffin *et al.*, 2003). Recently, carbohydrate grain-filling of barley mutants has been studied by  $^1\text{H}$  HR-MAS NMR (Seefeldt *et al.*, 2008).

The overwhelming information produced by spectroscopic screening of complex biological samples calls for multivariate data analysis such as chemometrics in order to extract systematic information. Such analysis requires

a minimum of assumptions and the relationships may be visualized by intuitive illustrations by means of the graphic computer interface (Winning *et al.*, 2008).

In the field of metabolomics, the analysis of metabolic changes in time is a fundamental aspect of understanding the biochemical response of an organism to an external perturbation (Lindon *et al.*, 2001). Thus, the aim of the current study was to investigate the implications of one or more drought events on protein quality in wheat grains using information from chemical protein analysis, liquid state  $^1\text{H}$  NMR of methanol extracts, and  $^1\text{H}$  HR-MAS NMR of single kernels and flours followed by unsupervised exploratory chemometric data analysis.

## Materials and methods

### Plant material

In a semi-field pot experiment, winter wheat (*Triticum aestivum* L. cv. Vinjett) was grown outdoors at the semi-field facility of the Faculty of Agricultural Sciences at Flakkebjerg (Slagelse, Denmark) in the growing season of 2005. Pots with both depth and diameter of 25 cm were equally filled with 4.2 kg 1:2:1 (by vol.) mixture of peat substrate, loamy soil, and sand. A dose of 5.25 g  $\text{K}_2\text{SO}_4$ , 3.5 g  $(\text{NH}_4)_2\text{SO}_4$ , 4.67 g  $\text{NH}_4\text{NO}_3$ , 1.9 g  $\text{CaSO}_4$ , 1.9 g  $\text{MgSO}_4$ , 0.4 g  $\text{MnSO}_4$ , 0.4 g  $\text{CuSO}_4$ , and 11.67 g  $\text{CaCO}_3$  per pot was also mixed in the soil. Spring wheat was sown at a rate of 15 seeds per pot and then thinned to five seedlings per pot at three-leaf stage.























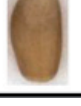

Water deficits were applied during two growth stages, namely terminal spikelet (end of spikelet initiation) and at anthesis by withholding irrigation. Spikes were harvested 10 d after anthesis (DAA), at four time points during the grain-filling period (17, 23, 31, and 43 DAA) and at harvest maturity (50 DAA), yielding six harvests in total (Fig. 1). The spikes were harvested and immediately frozen in liquid nitrogen and stored at  $-80^\circ\text{C}$ . Afterwards, the spikes were freeze-dried for 2 d.

The freeze-dried grains were milled (0.5 mm, Cyclotec 1093, Foss Tecator AB, Högenas, Sweden). The flour material was stored in sealed plastic bottles at  $4^\circ\text{C}$  until analysis. A flour sample unit consisted of the seeds from two spikes. A total of 48 flour samples were analysed, covering a total of six harvests including two replicates of all four treatments ( $6 \times 4 \times 2$ ).

See Fig. 1 for an overview of the experimental design. The four treatments constitute of the control treatment (CT) which has sufficient water supply throughout the period, the early drought treatment which is exposed to drought at double ridge (TD), the late drought treatment which is exposed to drought at anthesis (TA), and the fourth treatment which is exposed to drought at double ridge and at anthesis (T2).

### Methanol extract preparation

Samples were prepared using a protocol performed by Baker *et al.* (2006) which was a modified form of the

Harvest 2005	8 <sup>th</sup> July	15 <sup>th</sup> July	21 <sup>th</sup> July	29 <sup>th</sup> July	10 <sup>th</sup> August	17 <sup>th</sup> August
Days after anthesis	10	17	23	31	43	50
Control:	Contr 1	Contr 2	Contr 3	Contr 4	Contr 5	Contr 6
No drought						
Early drought:	TD 1	TD 2	TD 3	TD 4	TD 5	TD 6
Drought at double ridge						
Late drought:	TA 1	TA 2	TA 3	TA 4	TA 5	TA 6
Drought at anthesis						
Two drought periods:	T2 1	T2 2	T2 3	T2 4	T2 5	T2 6
Double ridge+ anthesis						

**Fig. 1.** The experimental design of the four different drought treatments: control (CT), early drought (TD), late drought (TA), and double drought (T2) together with the six harvests.

method described by Ward *et al.* (2003). Replicate aliquots of white flour (30 mg) were weighed into 1.5 ml Eppendorf tubes. D<sub>2</sub>O–CD<sub>3</sub>OD (1 ml, 80:20) containing 0.05% (w/v) TSP-d4 was added to each sample. The contents of the tube were mixed thoroughly and heated at 50 °C in a water bath for 10 min. The samples were then spun down in a micro centrifuge for 5 min; 800 µl of the supernatant was transferred to an Eppendorf tube and kept at 90 °C in a water bath for 2 min. The high-temperature (90 °C) step was incorporated to ensure that enzyme activity had stopped. The samples were then stored at 4 °C for 45 min prior to recentrifugation for 5 min (still at 4 °C); 700 µl of the supernatant was transferred to a 5 mm (o.d.) NMR tube. The residual CD<sub>2</sub>HOD multiplet in the region 3.36–3.32 ppm was excluded from all data sets.

#### <sup>1</sup>H HR-MAS NMR

<sup>1</sup>H HR-MAS NMR spectra were obtained using a Bruker AVANCE-400 (Bruker BioSpin, Rheinstetten, Germany) spectrometer, operating at a frequency of 400.13 MHz for protons equipped with a HR-MAS double channel probe using a 50 µl zirconia rotor (4.0 mm o.d.). Samples were prepared in the rotor using approximately 14 mg flour or one single kernel and 30 µl of D<sub>2</sub>O (with 5.8 mM of TSP-d4 (per-deuterated 3-trimethylsilyl propionate sodium salt). Data were accumulated at 298 K employing a pulse sequence using presaturation of the water resonance during

the 2 s recycle period followed by a composite 90 degree pulse (Bax, 1985) with an acquisition time of 2.045 s, 256 scans, and a spectral width of 8012.82 Hz, resulting in 16 k complex data points. A spin-rate of 7 kHz was used for all experiments. All samples were individually tuned and matched and the corresponding spectra were automatically phased and baseline-corrected and referenced to TSP-d4 at 0.0 ppm. Prior to Fourier transformation, each FID was apodized by Lorentzian line broadening of 0.3 Hz and zero filled to 64 k points. Flour and whole kernels were analysed approximately 2 h after preparation. All spectra were normalized relative to the TSP-d4 signal. The spectral range from 0.45–8.70 ppm was chosen, resulting in 27 000 data points. Bruker Topspin 1.3 (Bruker BioSpin 2005) was used for acquisition and processing of NMR data.

#### Liquid state <sup>1</sup>H NMR

The liquid state <sup>1</sup>H NMR spectra were obtained using a Bruker AVANCE-400 (Bruker BioSpin, Rheinstetten, Germany) spectrometer, operating at 400.13 MHz for protons using a broad band inverse detection probe head equipped with 5 mm (o.d.) NMR sample tubes. Data were accumulated at 298 K employing a pulse sequence using presaturation of the water resonance during the 2 s recycle period followed by a composite 90 degree pulse (Bax, 1985) with an acquisition time of 2.045 s, 256 scans, and a spectral width of 8012.82 Hz, resulting in 16 k complex data points.

All samples were individually tuned, matched, and shimmed. Prior to Fourier transformation, each FID was apodised by Lorentzian line broadening of 0.3 Hz and the corresponding spectra were automatically phased and baseline corrected and referenced to TSP-d4. In order to resolve the complex carbohydrate part of the spectra, a series of 1D and 2D experiments ( $^{13}\text{C}$ -HSQC, COSY, and TOCSY) were recorded at 18.8 T using a Bruker Avance 800 spectrometer operating at 799.92 MHz for protons and equipped with a 5 mm cryo probe.

#### Protein analysis

The protein content of flour ( $N \times 5.7$ ) was determined by Kjeldahl analysis (Kjeldahl, 1883) in duplicate. The protein fractions (albumins, globulins, and gliadins) were extracted from flour in triplicate according to Osborne (Ghirardo *et al.*, 2005), and quantified according to Popov *et al.* (1975).

#### Chemometric analysis and software

Multivariate data analysis in the form of principal component analysis (PCA) (Hotelling, 1933) was applied to obtain systematic variations from the measured spectra. PCA is the primary tool for investigation of large bilinear data structures for the study of trends, groupings, and outliers. By means of PCA it is possible to find the main variation in a two-dimensional data set by creating new linear combinations, PCs, from the underlying latent structures in the raw data. To study the grain-filling process through time, PCA score trajectories are constructed from NMR data to identify changes in the biochemical profile. This procedure has previously been applied for analysis of dynamic change during bread baking (Engelsen *et al.*, 2001) and of metabolic time response to toxic lesion (Keun *et al.*, 2004).

In exploratory studies investigating changes over time, time series are naturally arranged as a three-way data set. The first dimension represents the treatment, the second dimension represents the harvest time, and the third dimension represents the metabolic profile measured by NMR (Castro and Manetti, 2007). Parallel factor analysis (PARAFAC) (Bro, 1997; Carrol and Chang, 1970; Harshman, 1970) can be considered as a multiway extension of PCA able to handle three-way data. The PARAFAC model is based on the decomposing of the data into trilinear components in a similar way to the bilinear components extracted in PCA. When such higher-order data are available, the so-called second-order data advantage gives unique solutions and, for example, the pure analyte spectra will be found in a mixture (Harshman, 1970). In this study, data were arranged in a three-way cube with single kernels with the four treatments in mode 1, the six harvest times in mode 2, and NMR spectra in mode 3 constituted by 27 000 data points, giving a data cube with dimensions  $4 \times 6 \times 27\,000$ , and the PARAFAC model was calculated with non-negativity constraints and two factors. Data were mean-centred prior to the chemometric analysis.

The spectra were analysed using the chemometric software LatentIX 1.0 ([www.latentix.com](http://www.latentix.com), Latent5, Copenhagen,

Denmark), PLS Toolbox 4.11 (Eigenvector Research, Manson, Washington, USA), and MATLAB 2007a (The MathWorks, Inc., Natick, Massachusetts, USA).

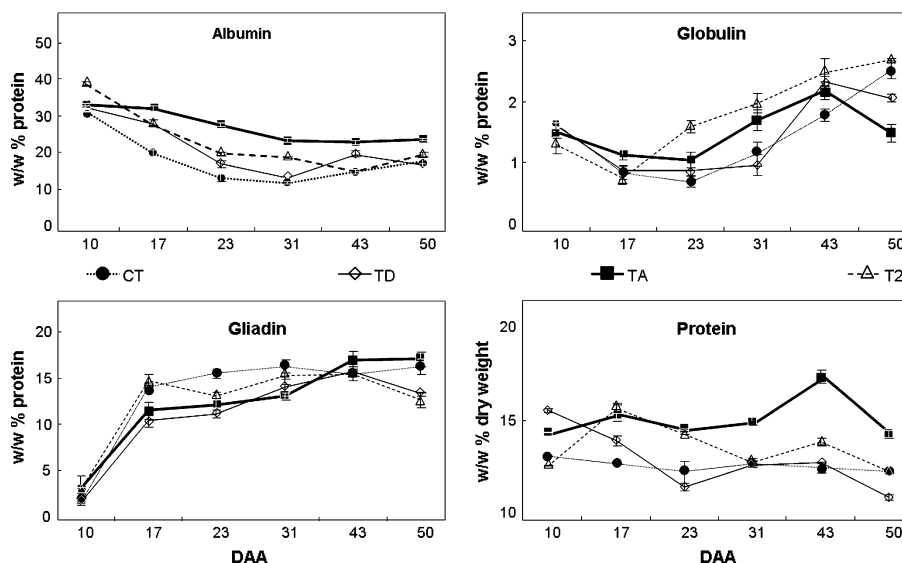
## Results and discussion

Drought-induced changes in the accumulation of protein and the distribution in the main fractions, albumins, globulins, and gliadins were examined for the wheat kernels during grain-filling. Figure 2 shows the results from the Popov and Kjeldahl analysis demonstrating the asynchronous protein metabolism during grain-filling and the influence of the drought exposure.

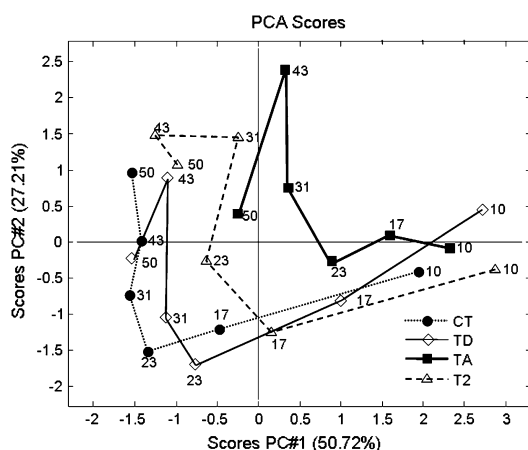
A common feature in all the drought treatments and in the control is the rapid increase in the gliadin fraction from 10 DAA to 17 DAA accompanied by reductions in the albumin and globulin fractions, an observation that is in accordance with the function of gliadins in the further development of the kernel (Shewry and Halford, 2002).

The time development of the albumin fraction in the control wheat (CT) deviates from the wheat exposed to drought, with the TA treatment most affected. The time development of the gliadin fractions was less influenced by the different drought treatments. The globulins showed a marked response to drought. The globulin level of the CT wheat decreased from 10 DAA to 23 DAA after which the level increased throughout the rest of the period of growth. However, when wheat suffered from drought, whether early or late (TA and TD), the globulins show the same 'pattern' from 10 DAA to 43 DAA, but decrease, in fact, from 43 DAA to 50 DAA. Surprisingly, after experiencing two drought periods, the globulin fraction showed the same development profile throughout grain-filling as the control group which had sufficient water, except that the increase had already started at 17 DAA. A one-factor analysis of variance (ANOVA) model on the total protein content showed significant difference between treatments of all six harvests. However, a Fisher's least significant difference test showed that CT and T2 are not different at harvest times of 17, 23, and 50 DAA and that no significant difference is observed between CT and TD at harvest times of 31 and 43 DAA. This simple analysis indicates that the late drought treatment induces the most significant changes in the total protein content compared with the control. The asynchronous development of the protein fractions is underlined by an almost complete lack of correlation between the protein fractions during development. However, the albumin and the gliadin fractions in CT wheat during development were correlated with a correlation coefficient ( $R^2$ ) of 0.88, but the correlation deteriorated significantly when exposed to drought.

In order to investigate changes in the protein profile of kernels affected by the different drought treatments, principal component trajectories have been calculated from the chemical data. These are presented in Fig. 3. Compared with the trajectory of CT, drought treatment apparently alters the trajectory pattern. The trajectories of the four



**Fig. 2.** Raw data from Popov and Kjeldahl analysis showing asynchronous protein metabolism in winter wheat as a function of DAA of the control group (CT), early drought (TD), late drought (TA), and twice drought (T2). Protein concentrations are measured in % dry weight kernel in duplicate. The albumin, gliadin, and globulin fractions are w/w % protein in triplicate: Error bars represent the deviation between replicates.

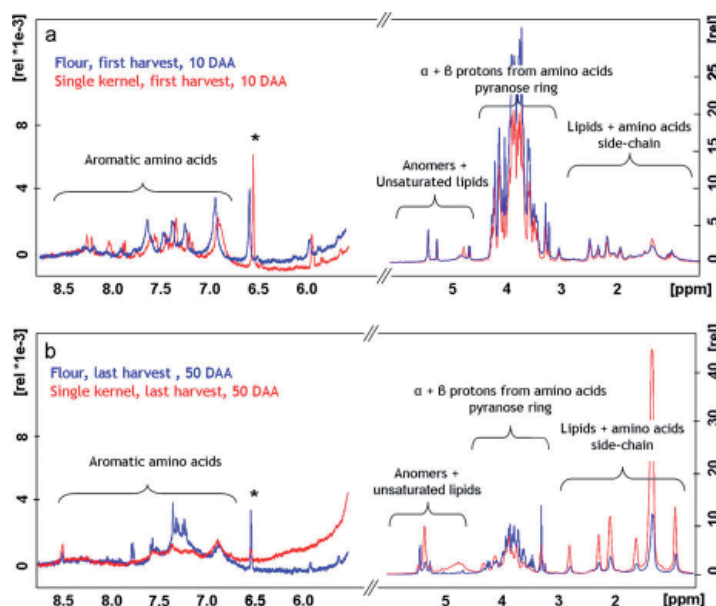


**Fig. 3.** PCA score plot of albumin, globulin, gliadin, and protein content during grain-filling showing different trajectories according to the drought treatment. PC1 and PC2 explain 51% and 27% of the variation, respectively.

treatments develop similarly progressing from 10–23 DAA along PC1, after which a bend along PC2 indicates that 23 DAA is a turning point for the protein synthesis of the CT, TD, and TA, whereas T2 deviates from this pattern by an increase in PC2 from 17–23 DAA. Moreover, a decrease in PC2 is observed from 43–50 DAA for the drought

treatment, but not for CT. An evaluation of the time development along PC1 shows that TD resembles CT the most and TA represents the most deviating treatment compared to CT. Altogether, the PCA revealed that late drought (TA) differs most from the control (CT) trajectory, indicating that the strongest effect on total protein content and the distribution in the protein fractions is after late drought. The PCA loadings (not shown) indicate that PC1 primarily concerns contributions from albumins and gliadins which are negatively correlated. PC2 primarily concerns contributions from globulins and total protein content (both positively correlated). In conclusion, the treatment which most resembles the control regarding protein development is the TD treatment, whereas the TA treatment was the most deviating.

In addition to the Popov and Kjeldahl protein analysis, wheat single kernels and wheat flour were analysed by  $^1\text{H}$  HR-MAS NMR. In Fig. 4, the spectra of flour and whole kernels from the control group at 10 DAA (Fig. 4a) and 50 DAA (Fig. 4b) are displayed. In the region 0–3 ppm, contributions arise from protons of the aliphatic side-chains of the amino acids and protons from the saturated parts of the lipids. The  $\alpha$ - and  $\beta$ -protons from amino acids resonate in the 3–5.5 ppm region, overlapping with the HOD signal and the carbohydrates. The signals observed at 6.7–8.5 ppm arise from aromatic protons. The most obvious difference between the early and late harvests is that the spectra of the early harvest are dominated by signals from small carbohydrates and free amino acids and the spectra of the late harvest are dominated by signals from lipids.

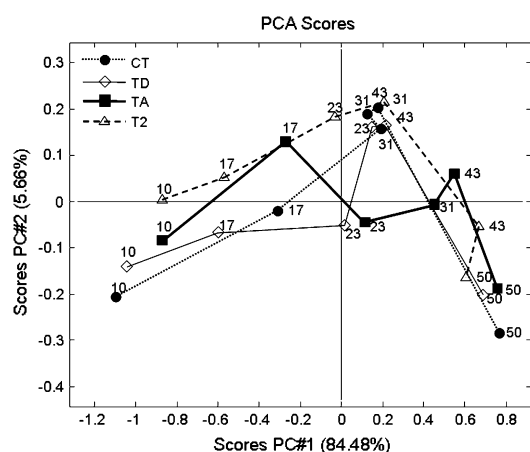


**Fig. 4.**  $^1\text{H}$  HR-MAS NMR spectra of wheat flour and single kernels of the control group from the first harvest, 10 DAA (a) and the last harvest, 50 DAA (b), respectively. \* Indicates fumaric acid. The region 5.5–9.0 ppm is vertically scaled by factor 20.

The NMR spectra of flour are better resolved in the 3–5.5 ppm region compared with the spectra of single kernels due to increased water accessibility of the molecules in the flour. In accordance with the continuous biosynthesis, amino acids, small carbohydrates etc. were observed in samples obtained at 10 DAA. Larger molecules (e.g. proteins or starch) are characterized by slower tumbling or a rigid structure which will increase the line width of the corresponding  $^1\text{H}$  resonances. Indeed, they can even be broadened beyond detection in the liquid-state  $^1\text{H}$  NMR spectrum. Therefore, the absence of resonances from the smaller molecules in the later stages of grain-filling implies that these are incorporated into larger molecules such as proteins, lipoproteins, glycolipids, or polysaccharides. The aromatic region is of particular interest with regard to proteins, since specific signals from aromatic amino acids (Phe, Trp, and Tyr) as well as His are located in this area without interfering signals from lipids and carbohydrates. Because it was possible to obtain information of the intact matrix of the wheat seed using single kernels, the  $^1\text{H}$  HR-MAS NMR measurements of single kernels are preferred in this work, compared to the results obtained in a destroyed sample matrix as flour. Due to low signal intensity of the aromatic region, full spectra will be analysed.

The overall differences in lipid, carbohydrate, and protein accumulation pattern are all combined in the PCA kernel development trajectory score plot (Fig. 5) which is based on the average NMR measurements of two single kernels. The PCA of the scaled  $^1\text{H}$  HR-MAS NMR spectra of the single

kernels confirms that periods of drought change the development of the wheat kernel through grain-filling. Compared to the trajectory of the protein data in Fig. 3, the NMR spectra of the single kernels are more variable at the first harvest. This is probably due to physical differences (size of kernels, spike, and position of the spike etc.) between the chosen single seeds. Apart from this, the  $^1\text{H}$  HR-MAS NMR spectra include information about carbohydrates, lipids, and proteins, whereas only proteins are included in Fig. 3. The PCA loadings (not shown) indicate that PC1 primarily describes aliphatic compounds, whereas PC2 primarily describes aromatic signals. The four trajectories have very different shapes according to treatment. However, the trajectory shape which most resembles the shape of the control is, in fact, the T2 trajectory which shows a parallel trajectory, although furthest from CT. The TD and the TA treatments follow different patterns in the score plot, with the TA treatment being the most deviating. The similarity of trajectories from T2 and CT indicates that treatment of wheat with two periods of drought does not significantly alter the final result. Compared to effects observed for TD and TA subjected to only one period of drought, this indicates that two drought periods induce better defence against drought compared to kernels treated with a single drought period. Apparently, the early drought treatment induces some drought immunity compared to drought later in the growth period. However, it was not possible to point to one compound expressing the drought treatment using the  $^1\text{H}$  HR-MAS NMR single kernel spectra.



**Fig. 5.** PCA trajectory score plots obtained from single kernel  $^1\text{H}$  HR-MAS NMR full spectra of the four different drought treatments (control (CT), early drought (TD), late drought (TA) and double drought (T2)). Spectra were block-scaled corresponding to the scaling used in Fig. 4. PC1 and PC2 explain 85% and 6% of the total NMR variation, respectively.

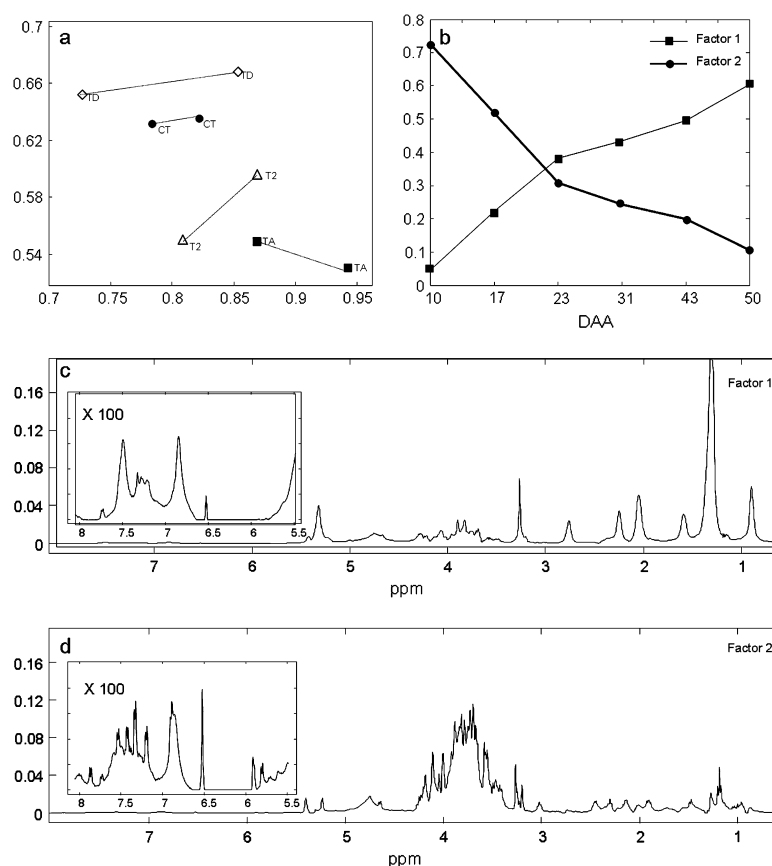
In order to investigate the measured  $^1\text{H}$  HR-MAS NMR spectra further and to utilize the fact that the time dimension allows the application of a trilinear model, PARAFAC was applied on the  $^1\text{H}$  HR-MAS NMR spectra of the single kernels (Fig. 6). The sample scores (Fig. 6a) of the four drought treatments were measured in duplicate showing the uncertainty of the replicates. The score plot reveals how the different drought treatments are clustered with the TD and CT placed close to each other and the TA and T2 placed together. The conclusion from this plot is that the  $^1\text{H}$  HR-MAS NMR spectra of wheat treated with late drought, again, is mostly affected compared to the CT  $^1\text{H}$  HR-MAS NMR spectra, whereas the TD treatment resembles the NMR profile of the control treatment. However, the result is not in total agreement with the results shown in Fig. 3 and Fig. 5. This is due to the modelling, since the three-way technique expresses only one concentration score value from each treatment and not a trajectory of the grain-filling pattern. The PARAFAC time loading (Fig. 6b) shows the time development of the two extracted components. The first time loading indicates how the NMR loading for Factor 1 increases during grain-filling. The Factor 1 NMR loading (Fig. 6c) resembles the profile of the fat fraction with characteristic signals from fatty acid methyl groups (0.9 ppm), methylene protons (1.3 ppm),  $\beta$ -carboxylic protons (1.6 ppm),  $\alpha$ -methyl protons (2.0),  $\alpha$ -carboxylic protons (2.3 ppm), aliphatic protons in conjugated chains (2.7 ppm), and protons on unsaturated carbons (5.3 ppm). The second time loading indicates how the NMR loading for Factor 2 decreases during grain-filling. The Factor 2 NMR loading (Fig. 6d) shows signals from protons in carbohydrates anomers and pyranose. The

small insert show the aromatic area which show small differences between Factor 1 and Factor 2. These have only a little influence on the model. These observations are consistent with observations from the  $^1\text{H}$  HR-MAS NMR spectra of the early and late harvests in Fig. 4.

In order to enhance the resolution of the spectra in a relative simple manner, methanol extracts were prepared from wheat flour. During extraction of flour with a water-methanol mixture, the water-soluble proteins (e.g. albumins) and some alcohol-soluble proteins will be extracted. The extract contains no lipids or starch, only small saccharides and some water-soluble lipoproteins. The proteins and enzymes denature in the heating step and there is no hydrolysis of proteins to amino acids. Small peptides and amino acids can be naturally present in the flour from the onset (Shewry *et al.*, 1984).

Figure 7 displays  $^1\text{H}$  NMR spectra of the liquid methanol extracts from the last harvest (50 DAA), showing the final result of the differences between the four drought treatments. Several carbohydrates were identified in the spectral region 3.5–5.5 ppm. The assignments were based on 2D experiments of the extracts, and include  $\alpha$ -glucose (5.20, 3.50, 3.70, 3.39, and 3.82 ppm, H1–H5, respectively),  $\beta$ -glucose (4.61, 3.21, 3.45, and 3.38 ppm, H1–H4, respectively), maltose [ $\alpha$ (1–4) 5.36, 3.55, 3.68, 3.40 ppm, H1–H4 and  $\beta$  4.61, 3.25, 3.74, 3.61, 3.58 ppm, H1–H5, respectively, and  $\alpha$  5.20, 3.54, 3.96 ppm, H1–H3, respectively], glucose-1-phosphate and glucose 1,6-diphosphate (5.41/5.40, 3.55/3.53, 3.74/3.74, 3.51/3.46, and 4.04/3.83 ppm, H1–H5, respectively), together with  $\alpha$ -galactose (4.98, 3.81, 3.88, 3.81, and 3.98 ppm, H1–H5, respectively). In the aliphatic region (0–3.5 ppm), the signals from the side-chain of amino acids including aspartic acid [(2.64 ( $H_\beta$ ), 2.80( $H_\beta$ ), and 3.87 ( $H_\alpha$ ) ppm, respectively) and alanine (1.48 ppm ( $H_\beta$ ) and 3.75 ppm ( $H_\alpha$ )), together with small acids such as malic acid [2.35 ( $H_\beta$ ), 2.66 ( $H_\beta$ ), and 4.28 ( $H_\alpha$ ) ppm, respectively] arise. However, the most significant differences between the four treatments were observed in the region 7.1–7.8 ppm, covering the aromatic amino acids. In this region, the intensity increases in the order: T2 < TA < TD < CT, indicating that the amount of water-soluble proteins is negatively affected by the drought treatment in this order. The opposite effect is observed for the sharp singlet at 6.51 ppm assigned to fumaric acid by confirmation by 2D NMR experiments. This assignment is in agreement with Ward *et al.* (2003) who measured methanol extracts of *Arabidopsis* and assigned the sharp singlet at 6.5 ppm to be fumaric acid, but without further augmentation (Ward *et al.*, 2003). This ordering of the drought treatments mentioned above is only observed for the last harvest, but the last harvest from the trajectory score plot of the  $^1\text{H}$  HR-MAS NMR spectra of whole kernels (Fig. 5, 50 DAA) confirms that T2 is the treatment furthest from the CT and the TD treatment is closest. This agreement is not surprising since the results obtained from methanol extracts includes water-soluble proteins only, which is about the same protein fraction as the water-mobilized proteins included in the  $^1\text{H}$  HR-MAS NMR spectra of whole single kernels.





**Fig. 6.** The PARAFAC results: samples scores (a) of the four drought treatments: control (CT), early drought (TD), late drought (TA), and double drought (T2) together with PARAFAC time loading (b) PARAFAC NMR loading of Factor 1 (c) and Factor 2 (d).

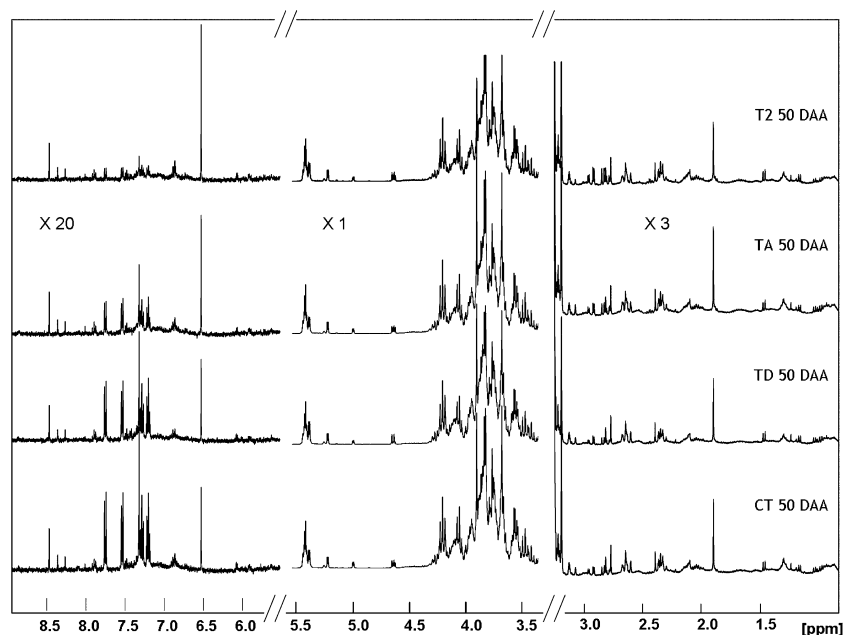
Turning the focus on the spectral region (6.5–7.8 ppm) in Fig. 7, the aromatic ring protons show decreasing intensity with respect to severity of drought treatment, and they are highly (negatively) correlated ( $R^2=0.94$ ) to the fumaric acid signal. This correlation between aromatic amino acids and a citrate cycle metabolite indicates that the effects of drought are almost equally severe on both.

The  $^1\text{H}$  HR-MAS NMR spectra of single wheat kernels in Fig. 8 show the fumaric acid signal in mature wheat kernels (50 DAA) coloured by increasing drought intensity. As apparent from the figure, the fumaric signal is visible, but the signal does not show increasing intensity for increasing drought treatment as the spectra of the methanol extracts did (Fig. 7). This could be due to the fact that the size, the spike, and the position of the spike differs amongst kernels. These differences are not present when measuring flour or methanol extracts of flour as the intravariation in wheat kernels are averaged out. However, the  $^1\text{H}$  HR-MAS NMR spectra of wheat flour did not show this interesting connection either. In wheatgrass metabolism, it was evident

that drought stress increased the extrusion from roots that contain higher concentration of fumaric and succinic acids (Henry *et al.*, 2007). Fumaric acid is involved in the citric acid cycle as succinate, is oxidized to fumarate by succinate dehydrogenase which is directly linked to the electron transport chain. The next step in the citric acid cycle is the hydration of fumarate to form L-malate, a step that needs water. We therefore speculate whether the lack of water during the growth of wheat, which apparently resulted in the accumulation of fumarate, has a negative effect on the citric acid cycle.

## Conclusions

In this paper, the effects of water deficiency during grain-filling in wheat have been investigated with emphasis on the asynchronous protein synthesis monitored by  $^1\text{H}$  NMR analysis of single kernels, flour, and methanol extracts. Protein fractions and methanol extraction were tested for



**Fig. 7.** Liquid-state  $^1\text{H}$  NMR spectra of methanol extractions of wheat harvested at 50 DAA showing the differences between the four drought treatments (control (CT), early drought (TD), late drought (TA), and double drought (T2)).

the wheat is drought-treated early in grain-filling. The  $^1\text{H}$  NMR spectra of methanol extracts of wheat flour samples provided better spectral resolution and enabling assignment of a sharp singlet at 6.51 ppm to fumaric acid. This metabolite, which is also found in the single kernel  $^1\text{H}$  HR-MAS NMR spectra, was found to be a potential marker for drought treatment in mature kernels.

The  $^1\text{H}$  HR-MAS NMR spectra of single wheat kernels show considerable differences between early and late harvest. For the first time, data from grain-filling has been analysed by a complex multiway model and the result showed an excellent overview of the data, where the main variation in the data is expressed in two spectral profiles.

**Fig. 8.**  $^1\text{H}$  HR-MAS NMR spectra of single wheat kernels from the last harvest coloured according to drought treatment.

their ability to differentiate between drought treatments, each method applying a different view on the protein depending on the solubility. Visualization of the data by PCA trajectories for the drought treatments yielded good contrast of the protein development during grain-filling. The results indicate that two periods of drought do not have as remarkable an influence as late drought. Apparently, some kind of resistance to drought is induced when

## Acknowledgements

The Ministry of Science, Technology, and Innovation is greatly acknowledged for financial support to the project; Exploring the asynchronous protein metabolism of wheat. We thank Betina Hansen, Ulla Andersen, and Jesper Pedersen for their help in setting up the experiment in the semi-field. Gilda Kischinovsky from the Faculty of Life Sciences, University of Copenhagen, is acknowledged for proofreading the manuscript. Bent O Petersen from Carlsberg Laboratory is greatly acknowledged for the recording of the 800 MHz NMR experiments and for the following assignment of the carbohydrate part of the NMR spectra. The 800 MHz NMR experiments were obtained on the Bruker

Avance 800 spectrometer of the Danish Instrument Center for NMR spectroscopy of Biological Macromolecules.

## References

- Baker JM, Hawkins ND, Ward JL, Lovegrove A, Napier JA, Shewry PR, Beale MH.** 2006. A metabolomic study of substantial equivalence of field-grown genetically modified wheat. *Plant Biotechnology Journal* **4**, 381–392.
- Bax A.** 1985. A spatially selective composite 90-degrees radio-frequency pulse. *Journal of Magnetic Resonance* **65**, 142–145.
- Bro R.** 1997. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems* **38**, 149–171.
- Carroll JD, Chang JJ.** 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition. *Psychometrika* **35**, 630–638.
- Castro C, Manetti C.** 2007. A multiway approach to analyse metabonomic data: a study of maize seeds development. *Analytical Biochemistry* **371**, 194–200.
- Engelsen SB, Jensen MK, Pedersen HT, Nørgaard L, Munck L.** 2001. NMR-baking and multivariate prediction of instrumental texture parameters in bread. *Journal of Cereal Science* **33**, 59–69.
- Garrod S, Humphreys E, Connor SC, Connelly JC, Spraul M, Nicholson JK, Holmes E.** 2001. High-resolution <sup>1</sup>H-NMR and magic angle spinning NMR spectroscopic investigation of the biochemical effects of 2-bromoethanamine in intact renal and hepatic tissue. *Magnetic Resonance in Medicine* **45**, 781–790.
- Ghirardo A, Sorensen HA, Petersen M, Jacobsen S, Sondergaard L.** 2005. Early prediction of wheat quality: analysis during grain development using mass spectrometry and multivariate data analysis. *Rapid Communications in Mass Spectrometry* **19**, 525–532.
- Griffin JL, Pole JCM, Nicholson JK, Carmichael PL.** 2003. Cellular environment of metabolites and a metabonomic study of tamoxifen in endometrial cells using gradient high resolution magic angle spinning H-1 NMR spectroscopy. *Biochimica et Biophysica Acta-General Subjects* **1619**, 151–158.
- Harshman RA.** 1970. Foundation of the PARAFAC procedure: models and conditions for an 'exploratory' multi-modal factor analysis. *UCLA Working Papers in Phonetics* **16**, 1–84.
- Henry A, Doucette W, Norton J, Bugbee B.** 2007. Changes in crested wheatgrass root exudation caused by flood, drought, and nutrient stress. *Journal of Environmental Quality* **36**, 904–912.
- Hottelling H.** 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24**, 417–441.
- IPCC.** 2007. *Impacts, adaptation and vulnerability* IPCC Fourth Assessment Report: Working Group II Report. The Intergovernmental Panel on Climate Change.
- Jamieson PD, Stone PJ, Semenov MA.** 2001. Towards modeling quality in wheat: from grain nitrogen concentration to protein composition. *Aspects of Applied Biology* **64**, 111–126.
- Keun HC, Ebbels TMD, Bollard ME, Beckonert O, Antti H, Holmes E, Lindon JC, Nicholson JK.** 2004. Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. *Chemical Research in Toxicology* **17**, 579–587.
- Kjeldahl J.** 1883. A new method for the determination of nitrogen in organic matter. *Zeitschrift für Analytische Chemie Fresenius* **22**, 366–382.
- Krishnan P, Kruger NJ, Ratcliffe RG.** 2005. Metabolite fingerprinting and profiling in plants using NMR. *Journal of Experimental Botany* **56**, 255–265.
- Lindon JC, Holmes E, Nicholson JK.** 2001. Pattern recognition methods and applications in biomedical magnetic resonance. *Progress in Nuclear Magnetic Resonance Spectroscopy* **39**, 1–40.
- Martre P, Porter JR, Jamieson PD, Triboi E.** 2003. Modeling grain nitrogen accumulation and protein composition to understand the sink/source regulations of nitrogen remobilization for wheat. *Plant Physiology* **133**, 1959–1967.
- Popov N, Schmitt M, Schulzeck S, Matthies H.** 1975. Reliable micromethod for determining protein-content in tissue material. *Acta Biologica et Medica Germanica* **34**, 1441–1446.
- Raven JR, Handley L, Wollenweber B.** 2004. Plant nutrition and water use efficiency. In: Bacon M, ed. *Water use efficiency in plant biology*. Oxford: Blackwell Science Publications, 162–188.
- Seefeldt HF, Larsen FH, Viereck N, Wollenweber B, Engelsen SB.** 2008. Bulk carbohydrate grain-filling of barley β-glucan mutants studied by <sup>1</sup>H HR MAS NMR. *Cereal Chemistry* **85**, 571–577.
- Shewry PR, Halford NG.** 2002. Cereal seed storage proteins: structures, properties and role in grain utilization. *Journal of Experimental Botany* **53**, 947–958.
- Shewry PR, Lafandra D, Salcedo G, Aragoncillo C, Garciaolmedo F, Lew EJL, Dietler MD, Kasarda DD.** 1984. N-terminal amino-acid-sequences of chloroform-methanol-soluble proteins and albumins from endosperms of wheat, barley and related species: homology with inhibitors of alpha-amylase and trypsin and with 2S storage globulins. *FEBS Letters* **175**, 359–363.
- Triboi E, Martre P, Triboi-Blondel AM.** 2003. Environmentally-induced changes in protein composition in developing grains of wheat are related to changes in total protein content. *Journal of Experimental Botany* **54**, 1731–1742.
- Ward JL, Harris C, Lewis J, Beale MH.** 2003. Assessment of <sup>1</sup>H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsis thaliana*. *Phytochemistry* **62**, 949–957.
- Winning H, Larsen FH, Bro R, Engelsen SB.** 2008. Quantitative analysis of NMR spectra with chemometrics. *Journal of Magnetic Resonance* **190**, 26–32.
- Wollenweber B, Porter JR, Schellberg J.** 2003. Lack of interaction between extreme high-temperature events at vegetative and reproductive growth stages in wheat. *Journal of Agronomy and Crop Science* **189**, 142–150.

## Paper III

*An exploratory in vivo metabonomic investigation reveals dimethylsulfone as a robust biomarker for onion intake*

H. Winning, E. Roldán-Marín, L.O. Dragsted, N. Viereck, M. Poulsen,  
C. Sánchez-Moreno, M.P. Cano, S.B. Engelsen

Submitted



# An exploratory *in vivo* metabonomic investigation reveals dimethylsulfone as a robust biomarker for onion intake

Hanne Wining<sup>a)</sup>, Eduvigis Roldán-Marín<sup>b) c)</sup>, Lars O. Dragsted<sup>b)</sup>, Nanna Viereck<sup>a)</sup>, Morten Poulsen<sup>d)</sup>, Concepción Sánchez-Moreno<sup>c)</sup>, M. Pila. Cano<sup>c)</sup>, Søren B. Engelsen<sup>a)</sup>

<sup>a)</sup>University of Copenhagen, Faculty of Life Sciences, Dept. of Food Science, Quality & Technology, Rolighedsvej 30, 1958 Frederiksberg C, Denmark

<sup>b)</sup>University of Copenhagen, Faculty of Life Sciences, Dept. of Human Nutrition, Centre for Advanced Food Studies, Rolighedsvej 30, 1958 Frederiksberg C, Denmark

<sup>c)</sup>Department of Plant Foods Science and Technology, Instituto del Frío, Consejo Superior de Investigaciones Científicas (CSIC), C/ José Antonio Novais, 10, Ciudad Universitaria, E-28040, Madrid, Spain

<sup>d)</sup> Danish Food Institute, Technical University of Denmark, Mørkhøj Bygade 19, 2860 Søborg, Denmark

Corresponding author: Hanne Wining, Tel +45 3335 3500, Fax, +453528 3245, email: [haw@life.ku.dk](mailto:haw@life.ku.dk)

## ABSTRACT

Onions contain several bioactive compounds and onion and onion by-products have been described to have potentially beneficial effects on health. This study is aimed to evaluate the *in vivo* metabolome following intake of onion by-products. Thirty-two rats were fed a diet containing an onion by-product or the derived onion by-product fractions; an ethanol extract and the residue resulting in three different onion diets. 24-hour urine was analyzed using proton nuclear magnetic resonance (<sup>1</sup>H NMR) spectroscopy in order to investigate the effects of onion intake on the rat metabolism. Principal component analysis (PCA) was not able to distinguish between urine from rats consuming normal feed and rats fed with an onion diet. Interval extended canonical variates analysis (iECVA) was applied which revealed two signals from an onion biomarker. The two onion biomarkers were identified as dimethylsulfone and 3-hydroxyphenylacetic acid. The same two spectral regions were subsequently revealed by interval partial least squares regression PLS (iPLS) to be perfect quantitative markers for onion intake. Quantitative calibration models using PLS were developed to predict percentage onion intake. The best calibration model yielded a root mean square error (RMSE) of cross-validation (CV) of 0.97 % (w/w) with only 1 latent variable (LV) and a squared correlation coefficient (R<sup>2</sup>) of 0.94. This indicates that urine from rat on the by-product diet, the extract diet, and residue diet all contains the same biomarkers. Therefore, it is concluded that dimethylsulfone and 3-hydroxyphenylacetic acid are robust biomarkers for onion intake. Being able to detect specific biomarkers is highly beneficial in control of nutritionally enhanced functional foods.

**KEYWORDS:** Metabonomics, NMR, rat, urine, iECVA, iPLS, biomarker, dimethylsulfone

## INTRODUCTION

Onions (*Allium cepa*) constitute a part of the daily diet for most of the world's population. Nutritionally, onion properties have been widely reported, indicating beneficial health effects. Most of the beneficial health effects have been related to the onion antioxidant, anticarcinogenic, antimutagenic, antiasthmatic, immunomodulatory, antimicrobial, prebiotic and cardiovascular protective properties.<sup>1-3</sup> The main bioactive compounds related to the onion beneficial health effects of onion include flavonols, particularly quercetin and quercetin glucosides<sup>4</sup>, soluble fibres, fructooligosaccharides and organosulfur compounds.<sup>5-9</sup> Organosulfur compounds have become subject of many investigations due to their potential chemopreventive and antioxidant effects.<sup>9-11</sup> For example, the *S*-methyl sulfoxide isolated from *Allium cepa* has been shown to have a lipid-lowering effect in cholesterol-fed rats.<sup>12</sup> The metabolism of onion is not yet fully understood, but cycloalliin, an organosulfur compound found in garlic and onion, initiates several biological activities and its metabolite, (3*R*,5*S*)-5-methyl-1,4-thiazane-3-carboxylic acid, has been found in urine after intravenous or oral administration to rats.<sup>13</sup> Boyle and co-workers (2000) found a significant decrease in the level of human urinary 8-hydroxy-2'-deoxyguanosine after ingestion of an onion meal.<sup>14</sup>

World wide, large amounts of onion disposal are produced from the onion production. Because onion is toxic for many animals, this waste product cannot be utilized in the general feeding industry. Onion waste can be stabilized as useful onion by-product which can act as an antioxidant or antibrowning agent. With respect to the health beneficial effect of onion, developing dietary supplements or nutritionally enhanced functional foods including onion could be highly beneficial. The concept of developing nutritionally enhanced functional foods requires the understanding of the mechanisms of prevention and protection if potential nutritional effects of an onion supplement should be utilized and the effect documented. The identification of biologically active molecules as potential biomarkers leads to a greater understanding of biochemical pathways and potentially allows objective quantification of onion intake in human studies.

High-resolution nuclear magnetic resonance (NMR) (most often proton (<sup>1</sup>H) NMR) has emerged as a powerful non-invasive technique for metabonomic studies due to its ability to simultaneously detect a large number of compounds in a rapid high-throughput manner that requires little sample manipulation.<sup>15</sup> <sup>1</sup>H NMR spectroscopy is widely used to study the metabolic variation in biofluids and its capabilities for metabonomics is well established.<sup>16</sup> It has become more and more common to combine investigations of complex NMR spectra with advanced multivariate data analysis such as chemometrics in order to extract systematic latent information from the complex biological NMR spectra. Such analysis requires a minimum of assumptions and the relationships may be visualized intuitively. Already in the early nineties, principal component analysis (PCA)<sup>17</sup> was introduced to classify <sup>1</sup>H NMR spectra of urine by the Nicholson group at Imperial College (UK).<sup>18</sup> This group also introduced the definition of metabonomics as: 'understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological <sup>1</sup>H NMR spectroscopic data'.<sup>19</sup> Subsequently, the term metabonomics has been broadened to 'metabolic processes studied by <sup>1</sup>H NMR spectroscopy of biofluids' and thus also include nutrition studies<sup>20</sup>.

Urine is often used as a biological fluid for metabonomic investigations due to the ease and non-invasive of collection of repeated samples, the variable metabolite composition, and the often higher metabolite concentrations achieved relative to blood plasma.<sup>21</sup> The fact that the urine profiles are generated and analyzed without *a priori* assumptions about the metabolic and physiological processes involved allows several hypotheses to be tested simultaneously, as well as new hypotheses to be generated from unexpected associations. In nutritional metabonomics, one often has to deal with large variability in the samples compared to the changes of interest induced by the nutritional intervention. The inter- and intra-individual metabolite variance within a normal population has to be evaluated qualitatively and quantitatively before conclusions can be made. In rodents, it has been determined that species, strain, genetics, sex, age, hormone concentrations, diurnal cycles, diet, temperature, stress and

gut microflora all contribute to the metabolic composition of the urine of the animals.<sup>22</sup> However, it is known that human volunteers in dietary metabonomic studies frequently do not report all their medication or food supplements<sup>23</sup> and it is therefore highly desirable to gain objective knowledge about the true diet of a test person. *In vivo* investigations using animals as model-system makes it possible to investigate biomarkers under controlled conditions.

A rodent study was recently conducted to evaluate possible health effects after feeding with an onion by-product and two derived onion fractions.<sup>24</sup> In this study, the effect of onion intake on antioxidant enzymes, DNA damage, and gut environment in healthy rats was investigated and it was found that the onion by-product and the onion sub-fractions have no *in vivo* genotoxicity, may support *in vivo* antioxidative defense and alters the functionality of the rat gut microbiota. The purpose of the work presented here is to investigate the effect of onion intake on urine composition of the same rats with explorative metabonomic analysis using <sup>1</sup>H NMR spectroscopy and chemometrics. Onion contains both soluble and insoluble compounds. Therefore, the onion product was fractionated into two fractions; an ethanol extract rich in fructooligosaccharides and the residue, the insoluble matrix. In this way, potential onion biomarkers in the onion by-product investigated can be either extracted into the extract or left in the residue.

## EXPERIMENTAL

### *Onion and rat study*

The onion product used to feed the studied animals, produced at Instituto del Frío (CSIC, Madrid, Spain) is a by-product produced from freeze-dried onion paste. The onion by-product was fractionated into an ethanol/water soluble extract which is 70 % (w/w) of the by-product and the rest, the dry residue, which is 30 % (w/w) of the by-product. The extraction and the rat study were carried out at the National Food Institute, Technical University of Denmark (Søborg, Denmark). The onion extraction procedure and the animal study are detailed described elsewhere.<sup>4</sup> Briefly, thirty-two rats were divided into four groups of 8 rats and fed four weeks either a control feed, a control feed supplemented with 10% of onion by-product, a control feed supplemented with 7% of onion extract or a control feed supplemented with 3% onion residue (Table 1). The control group was fed with an isocaloric diet, substituting onion sugars with sucrose and onion fiber with starch. The two onion fractions are supplemented to the feed in concentrations which match the concentration of the by-product. Due to the experimental design, the extract and the residue added should be similar to the by-product, provided that the extraction is complete.

Table 1 Composition of rat feed. For detailed composition of diet see Roldán-Marín et al. (2009)<sup>24</sup>

<i>g per kg feed</i>	<i>Control group</i>	<i>Onion by-product group</i>	<i>Onion extract group</i>	<i>Onion residual group</i>
Onion by-product	0	100g	0	0
Onion extract	0	0	70g	0
Onion residual	0	0	0	30g
Control feed	1000g	900g	930g	970g
<b>Total feed</b>	1000g	1000g	1000g	1000g



### *Urine samples*

Urine samples were collected for a period of 24 hours. 2 ml of 1 mM  $\text{NaN}_3$  were added to the urine sample test tubes which were kept at  $0^\circ\text{C}$ . The urine volume was recorded and samples were frozen in portions at  $-80^\circ\text{C}$  for further analysis. Prior the  $^1\text{H}$  NMR analysis, the thawed urine samples were prepared by centrifugation at 1600 rpm for 10 min, 340  $\mu\text{l}$  of the supernatant were transferred to NMR tubes followed by addition of 170  $\mu\text{l}$  of 100 mM phosphate buffer solution ( $\text{H}_2\text{O}$ ,  $\text{D}_2\text{O}$ , TSP-d4 (per-deuterated 3-trimethylsilyl propionate sodium salt),  $\text{NaN}_3$ , pH 7.4) to reduce the pH range of the samples. TSP-d4 was added to act as an internal chemical shift reference ( $\delta^1\text{H}$  0.0),  $\text{D}_2\text{O}$  was added to provide a lock signal for the NMR spectrometer and  $\text{NaN}_3$  was added as a preservative. The urine samples were prepared to run in a random order.

### *$^1\text{H}$ NMR measurements*

$^1\text{H}$  NMR spectra were recorded for the 32 urine samples. The spectra were acquired on a Bruker Avance Ultra Shield 400 spectrometer (Bruker Biospin GmbH, Rheinstetten, Germany) operating at 400.13 MHz using a broad band inverse probe head. Data were accumulated at 300 K employing a pulse sequence using pre-saturation of the water resonance during the recycle period followed by a 90 degree pulse<sup>25</sup> with an acquisition time of 2.04 s, a recycle delay of 5 s, 128 scans and a sweep width of 8012.82 Hz, resulting in 32 k complex data points. All samples were automatically tuned, matched and shimmed. Prior to Fourier transformation, each FID was apodized by Lorentzian line broadening of 1.0 Hz and zero-filled once and the corresponding spectra were manually phased and automatically baseline corrected. Receiver gain was automatically set. Prior to the chemometric analysis the raw proton NMR spectra data matrix to be investigated had the dimensions ( $32 \times 65536$ ) but was reduced to 32202 data points (10-0.2 ppm) excluding spectral areas with no signals. All spectra were aligned (rigid movement) in proportion to the TSP signal at 0.0 ppm. Furthermore, due to insufficient (unequal) depression of the water signal, the area from 5.00-4.50 ppm was removed. It also proved necessary to normalize spectra in proportion to the total sum of the spectra in order to remove the large concentration differences of the urine samples. Normalization of urinary metabolic data is best considered as a data transformation which minimizes inter-sample variation due to differences in gross urinary concentration between samples caused by volume and dry matter differences. Furthermore two 2D NMR experiments; total correlation spectroscopy (TOCSY) and heteronuclear single quantum coherence (HSQC) spectra were acquired on urine from a rat fed onion by-product supplement. These experiments were used for assignment of selected signals. The 2D NMR spectra were recorded using the Bruker pulse sequences; *mlewphpr* and *hsqcgpqh* (mixing time of 60 ms)<sup>25</sup>. The 2D NMR spectra were referenced to TSP-d4 at 0.0 ppm before data analysis. Besides the spectra of the rat urine,  $^1\text{H}$  NMR spectra of the three different onion fractions were obtained. 1 mg of each of the onion fractions were suspended in 1 ml  $\text{D}_2\text{O}$  solution added of 10% TSP-d4. Acquisitions parameters were similar the one used for the urine NMR spectra. The spectrum of dimethylsulfone was also measured with the same parameters.

### *Chemometric analysis and software*

Multivariate data analysis in the form of PCA and partial least squares regression (PLS)<sup>26</sup> was applied to obtain optimal quantitative and qualitative information from the measured spectra. PCA is the primary tool for investigation of large bilinear data structures for the study of trends, groupings and outliers. By means of PCA it is possible to find the main variation in a multidimensional data set by creating new linear combinations, principal components (PC's), from the underlying latent structures in the raw data. PLS is a multivariate calibration method by which two sets of data, X and y, are related by

means of regression. The purpose of PLS is to establish a linear model of latent variables (LV's), which enables the prediction of a reference value  $y$  (slow measurement) from the measured spectrum  $X$  (fast measurement). Furthermore, extended canonical variates analysis (ECVA)<sup>27</sup> models were applied for classification of feed groups. Canonical variates analysis (CVA)<sup>28,29</sup> is a method for estimation of directions in space that maximizes the differences between groups of samples. However, CVA cannot deal with highly collinear data such as spectroscopic data, where the number of variables is much larger than the number of samples. The ECVA method solves this problem by the use of PLS in the inner part of CVA and thereby allowing for the analysis of highly collinear data<sup>27</sup>. In order to improve the calibration models and to investigate the influential areas of the spectra, interval PLS (iPLS) and interval ECVA (iECVA) was employed.<sup>30</sup> iPLS is an extension of PLS which develops local PLS models on a number of subintervals of the full-spectrum region. The main advantage of iPLS is that it provides an overall picture of the relevant information in different spectral subdivisions, thereby facilitating interpretations and removing interferences from other regions. iECVA works similar to the iPLS model.

Scaling or other pre-transformations of NMR data can be often necessary before the data analysis in order to assure that all signals are influencing model. In this study, pareto-scaling was used as scaling method applied to the NMR data before the further data analysis. Pareto-scaling reduces the relative importance of large values, but keeps the data structure partially intact. Each variable is divided by the square root of the standard deviation of the column values.<sup>31</sup> All the calibration models were validated using cross-validation (CV) with five segments, leaving out one segment at a time from which the root mean square error of cross-validation (RMSECV) was calculated as a measure of the prediction error.

The spectra were analyzed using the chemometric software LatentIX 2.0 (www.latentix.com, Latent5, Copenhagen, Denmark), PLS Toolbox 4.11 (Eigenvector Research, Manson, Washington, USA), and MATLAB 7.6 2008a (The MathWorks, Inc., Natick, Massachusetts, USA). Regression (iPLS) and the iECVA model were performed in MATLAB using iToolbox and the ECVA Toolbox version 2.02, respectively (all available at [www.models.life.ku.dk](http://www.models.life.ku.dk)).

## RESULTS AND DISCUSSION

The <sup>1</sup>H NMR spectra of the three onion products to be fed to the rats are shown in Figure 1.

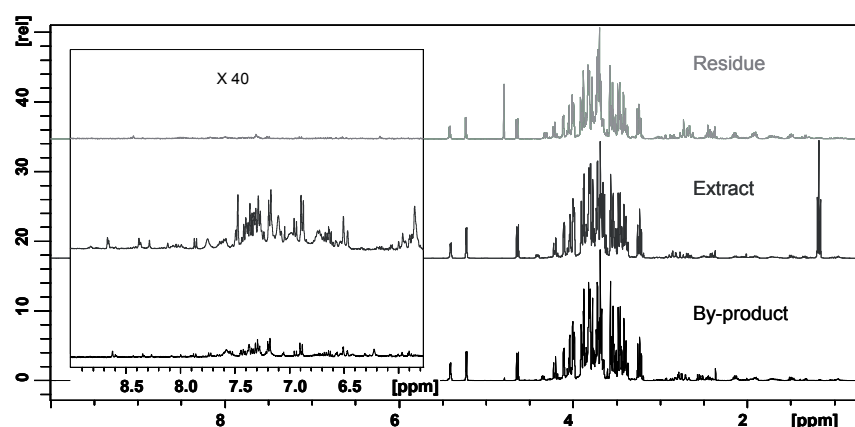
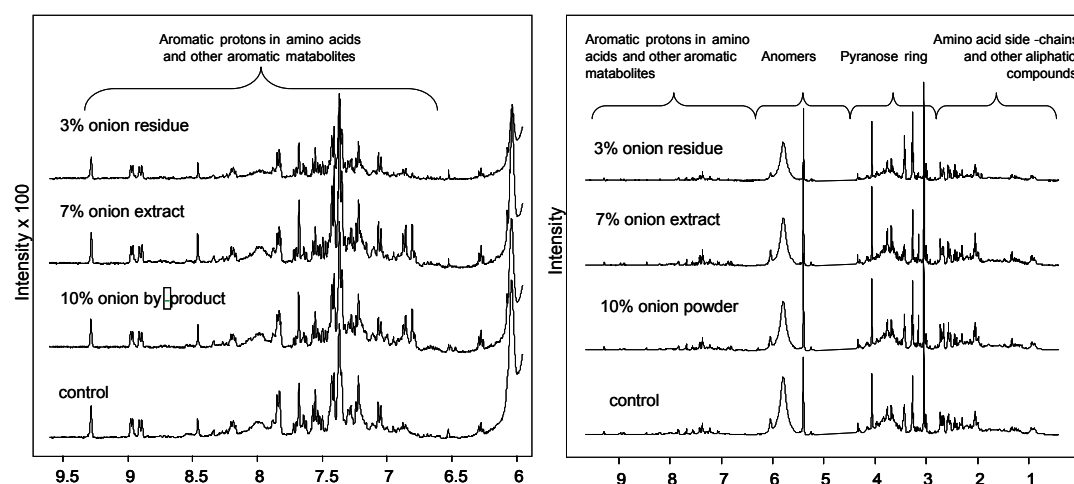


Figure 1 <sup>1</sup>H NMR spectra of onion by-product, onion extract and onion residue dissolved in D<sub>2</sub>O included 1mg/ml TPS-d4

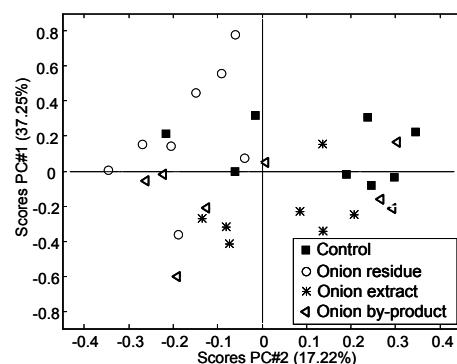
There are many similarities in the spectra of the three onion products, illustrating the complexity and the difference between the three onion fractions before they are metabolized by the rats. The spectra reveal diets high in fructans (3.5-6 ppm) with significant amount of aromatic (6-9 ppm) compounds. The extract differs from the residue by more intense signals in the aromatic region. Furthermore, the residue spectra have more intense signals in the high-field region of the spectrum compared to the two other fractions. The spectrum of the (ethanol) extract differs from the by-product and the residue by a triplet at 1.18 ppm which is assigned to the CH<sub>3</sub> signal from residual ethanol. From the spectra, it is difficult to assure that the by-product spectrum equals the residue plus the extract spectrum.

Figure 2 shows the average of <sup>1</sup>H NMR rat urine spectra of each of the feed groups. The spectra appear very similar despite the different feeding schemes. The <sup>1</sup>H NMR spectra of urine typically contain thousands of sharp lines from predominantly low molecular weight metabolites except for one broad band at 5.8 ppm from urea.<sup>32-41</sup> The spectra display a wide range of metabolites such as aromatics, aliphatic compounds, sugars, amino acids and other metabolites. However, from this global investigation of the raw data, no obvious difference in the urine profile of the three onion diets can be detected.



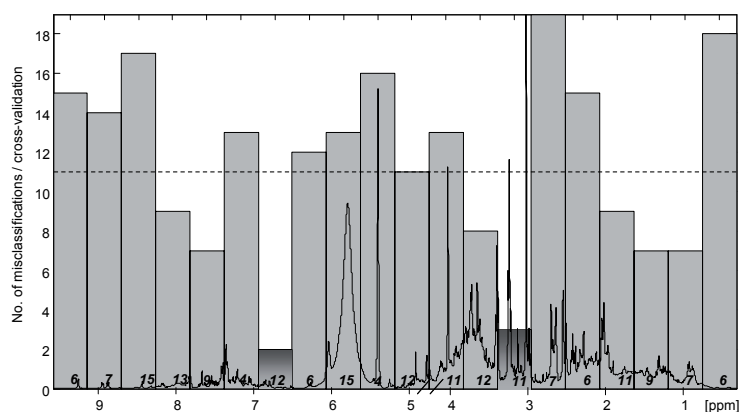
**Figure 2** <sup>1</sup>H NMR spectra of averaged rat urine from each of the four dietary groups. The water signal is removed and the aromatic region magnified by a factor 100.

In order to investigate possible metabolic differences between the different feeding schemes, a PCA model was established on the full NMR spectra of the 32 urine samples. However, the PCA model was not able to distinguish between the four different feeding groups or to group the samples in an onion and a control group (Figure 3). None of the four groups appear to differentiate significantly.



**Figure 3** PCA model (PC1 versus PC2) of pareto-scaled  $^1\text{H}$  NMR spectra of 32 urine samples of rats from the four different feeding schemes

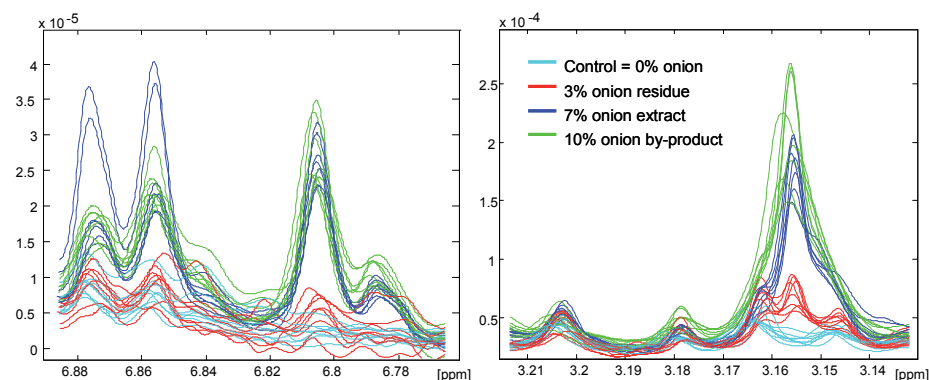
In order to scrutinize the spectra for signals able to distinguish between the control feed and the different onion fed groups (by-product, extract and residue), iECVA was carried out on the urine spectra (Figure 4) using 20 equally sized subintervals.



**Figure 4** iECVA plot  $^1\text{H}$  NMR spectra of urine from onion-fed rats indicating the two best intervals for lowest number of misclassifications. The dotted line is number of misclassifications (11 for 11 LV's) for global model and the italic numbers are optimal LV's in interval model

Indeed, two interesting intervals were found by iECVA which were able to improve the misclassifications rate significantly. The best interval, 6.50-6.95 ppm, was able to reduce the number of misclassifications from 11 to 2 (Figure 4). The interval includes signals from some of the aromatic compounds in the urine. The second best interval from 2.98-3.42 ppm was able to decrease the number of misclassifications to 3.

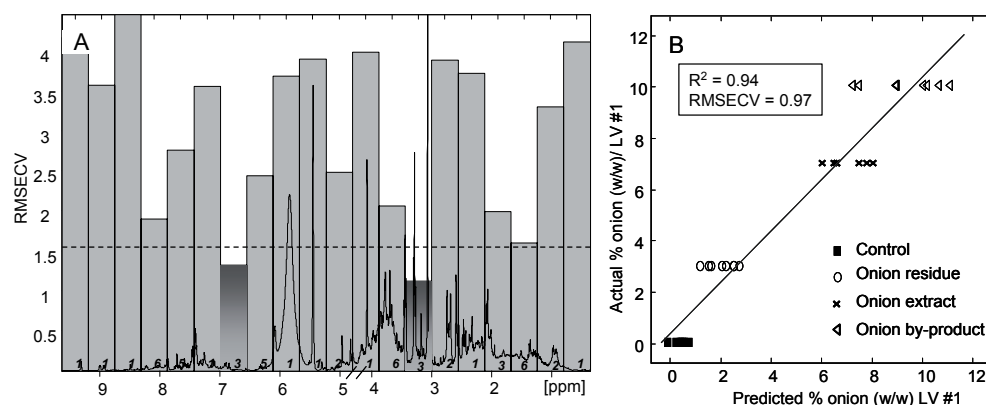
The signals in two intervals selected by iECVA are shown in Figure 5, colored according to feed group.



**Figure 5** Selected spectral regions from the iECVA model of the  $^1\text{H}$  NMR spectra of urine from onion-fed rats, revealing difference in signal intensity for each onion fraction

Figure 5 shows that the urine spectra with the highest signal intensity in the selected spectral region is the urine from rats fed with a diet which contained a 10 % supplement of onion by-product. In contrast, the urine from rats fed the control diet (without onion) shows no signal in this area. This indicates that the signal also contains quantitative information on onion dose. The signal with a chemical shift of 6.8 ppm shows (to a lesser extent) the same pattern. This signal matches the spectral profile of 3-hydroxyphenylacetic acid, when matched in an NMR spectral base (BBIOREFCODE) containing 535 compounds found in urine. The correlation between the signal at 3.15 ppm and the 3-hydroxyphenylacetic acid at 6.80 ppm is 0.94, which indicates that the 3-hydroxyphenylacetic acid and the compound which has signal at 3.15 ppm both are involved in the metabolism of onion. The correlation coefficient between the signal at 6.80 ppm and the doublet at 6.86 ppm is slightly lower: 0.89, which might indicate that other compounds also have a signal at this chemical shift.

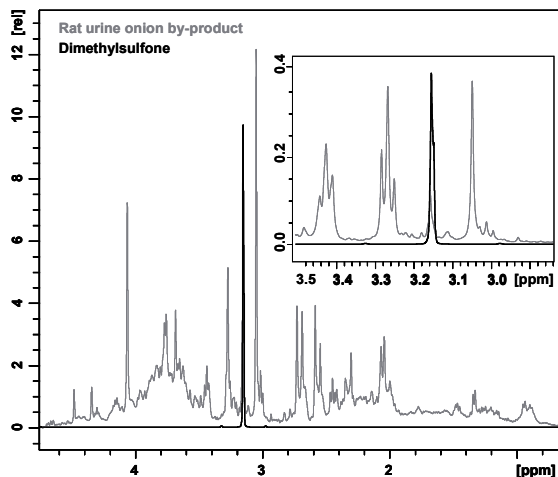
In order to investigate quantitative information regarding onion intake (*onion dose*) the  $^1\text{H}$  NMR spectra were analyzed using PLS relating to the  $^1\text{H}$  NMR spectra and the onion dose (0, 3, 7 and 10%). Variable selection using iPLS was applied to find which regions of the  $^1\text{H}$  NMR spectra of urine contain quantitative information about the onion biomarkers. The prediction error of the full-spectrum model was 1.56 % (w/w), as illustrated by the dashed line in Figure 6. Two intervals were found which were able to improve the prediction error significantly: 6.50-6.95 ppm and 2.98-3.42 ppm (marked in Figure 6). The optimal interval is 2.98-3.42 ppm which results in a prediction error of 1.12 % using only 3 LV's. Adding the interval around 6.80 ppm results in a further reduction of the error to 0.97 % (w/w) using only one LV. The actual versus predicted plot in Figure 6B shows a robust PLS model obtained on the two selected intervals, clearly revealing that the NMR urine spectra contain robust quantitative information about onion dose.



**Figure 6** (A) iPLS plot of the prediction of % (w/w) onion in feed obtained on  $^1\text{H}$  NMR spectra of urine from onion fed-rats. Dotted line is RMSECV (7 LV's) for global model. Italic numbers are optimal LV's in interval model. The two best intervals are highlighted. (B) Actual versus predicted plot of PLS model of onion dose and urine from onion-fed rats, performed on the highlighted intervals (A)

The two optimal spectra regions found by iPLS are exactly the same intervals as found by the iECVA. This indicates that it is the same quantitative information which is extracted by the ECVA and iPLS. Unfortunately, the experimental design used have dose and fraction confounded which makes it impossible to decide which effect is modeled, even if the  $^1\text{H}$  NMR spectra should hold information about both features. In theory, it should be possible to mathematically remove the information about dose and retain the information about fraction. This can be done by orthogonalization where the vector describing the dose response is withdrawn from the data.<sup>42</sup> However, it was tested in this study, but the orthogonalization approach led to a rather overfitted classification (results not shown).

In the interval from 2.98-3.42 ppm, one signal seems particular important. The signal has chemical shift of 3.15 ppm. Based on 2D experiments (TOCSY and HSQC) of the urine sample and NMR measurements of the pure compound, this signal was identified as the methyl protons ( $\text{CH}_3$ ) in dimethylsulfone (Figure 7).



**Figure 7**  $^1\text{H}$  NMR spectra of selected urine from rat on an onion by-product diet overlapped a spectrum of dimethylsulfone

Indeed, this symmetric compound has only one signal in the  $^1\text{H}$  NMR spectrum, and no cross-peak in TOCSY. From the HSQC experiment the chemical shift of  $^{13}\text{C}$  was found to be of 44 ppm. Dimethylsulfone is an oxidation product of dimethylsulfoxide (DMSO) and it is highly possible to find DMSO in urine as a result of an onion diet because onions contain many sulfoxides.<sup>43</sup> It has previously been shown that DMSO is metabolized to dimethylsulfone in humans and rats.<sup>44</sup> DMSO is a universal solvent and has the characteristic property that it is able to penetrate the skin. DMSO is an industrial approved solvent by the FDA for treatment of interstitial cystitis (bladder inflammation) and side effects include reports of onion odor breath. Several reports have suggested that DMSO may also be effective in inhibiting cholesterol-induced atherosclerosis in experimental animals.<sup>45-47</sup> Dimethylsulphone has been reported in human sweat,<sup>48</sup> in urine following asparagus consumption<sup>49,50</sup> and in cows milk from pasture-fed cows<sup>51</sup>. The compound therefore seems to originate from sulphur rich herbs and plant foods but there may be a genetic element involved in its formation in humans. Dimethylsulfone has recently been linked with the occurrence of skin cancer.<sup>52,53</sup> Gallagher and co-workers (2008) found that skin cancer patients showed significantly higher levels of dimethylsulfone in the skin measured by gas chromatography/mass spectrometry (GC-MS). An NMR study has also found detectable levels of dimethylsulfone normally present in the blood and cerebrospinal fluid, suggesting that it derives from dietary sources, intestinal bacterial metabolism, and the body's endogenous methanethiol metabolism.<sup>54</sup>

The good correlation between onion dose and the NMR spectra of urine show that the onion biomarker is present in all fractions and is equally distributed in the fractions of the by-product. Apparently the concentration of the biomarker is proportional with the onion dose intake independent of the fed onion fraction. That dimethylsulfone is present in all urine fractions may be due to the extraction procedure which not able to eliminate the biomarker from the by-product to the extract. The by-product consisted of intact cell walls which may be the reason why all dimethylsulfone was not removed from the residue. Another explanation could be that the compound is only partly soluble in ethanol. A third explanation may be that dimethylsulphone is a degradation product of more lipid-soluble organosulphur compounds from onion, and that these compounds were only weakly soluble in 60% ethanol.

## CONCLUSION

Two onion biomarkers were identified as dimethylsulfone and 3-hydroxyphenylacetic acid. Quantitative PLS models showed that the onion dose responded the quantitative information in the urine spectra primarily due to the biomarker dimethylsulfone. This indicated that urine from rats fed with the two fractions (extract and residue) of the onion by-product and rats fed with the onion by-product, all contain the biomarker and that the biomarker is present in all fractions and in the same concentration as doses. Therefore, it was possible to identify an objective biomarker for onion intake but not for the different onion products. Clearly the dimethylsulfone ends up in all fractions and is therefore a robust biomarker for onion intake. Being able to quantify the dietary intake can be very beneficial as control in diet intervention studies. The self-reported dietary intake in forms of food frequency questionnaires have been the dietary assessment method used most frequently in large-scale studies. This is primarily because it is easy to administer, it is less expensive than other dietary assessment methods, and it provides a rapid estimate of usual intake.<sup>55</sup> However, there exists a great problem in using food frequency questionnaires because self-reports of food intake are not accurate and sometime misleading. McKeown et al. (2001) showed that correlations between 24-hour urinary nitrogen excretion and dietary intake from the food frequency questionnaire were as low as 0.25.<sup>56</sup> Clearly a potential exist in using the onion biomarker in various nutrition studies. We are now in the process of carrying out a human study with an onion product in order to verify the usefulness of this marker and to whether genetic variability or variations in gut flora might affect its usefulness in humans.

## ACKNOWLEDGMENT

This study was supported by funding from the Danish Ministry of Food, Agriculture and Fisheries (NuBI, 3304-FVFP-060696-01) and 'Urinary biomarkers for eating pattern using NMR spectroscopy and chemometrics' (3304-FVFP-060706-01) and the Spanish Ministry of Science and Technology (AGL2003-09138-C04-01). In addition, the authors acknowledge the financial support of the Spanish Ministry of Education and Science (200670I081 and Program Consolider-Ingenio 2010, FUN-C-FOOD, CSD2007-00063). The Spanish Ministry of Education and Science is acknowledged for a Predoctoral fellowship. Finally, Dr. Manfred Spraul, Director NMR-Application and worldwide Coordinator Hyphenation, Bruker BioSpin, Germany is greatly acknowledged for help with assignment of the  $^1\text{H}$  NMR spectra.

## REFERENCES

1. Griffiths, G.; Trueman, L.; Crowther, T.; Thomas, B.; Smith, B. Onions - A global benefit to health. *Phytotherapy Research* **2002**, *16* (7), 603-615.
2. Gabler, N. K.; Osrowska, E.; Imsic, M.; Eagling, D. R.; Jois, M.; Tatham, B. G.; Dunshea, F. R. Dietary onion intake as part of a typical high fat diet improves indices of cardiovascular health using the mixed sex pig model. *Plant Foods for Human Nutrition* **2006**, *61* (4), 179-185.
3. Corzo-Martínez, M.; Corzo, N.; Villamiel, M. Biological properties of onions and garlic. *Trends in Food Science & Technology* **2007**, *18*, 609-625.
4. Roldán-Marín, E.; Sánchez-Moreno, C.; Lloría, R.; de Ancos, B.; Cano, M. P. Onion high pressure processing: Flavonol content and antioxidant activity. *LWT - Food Science and Technology* **2009**, *42*, 835-841.
5. Price, K. R.; Rhodes, M. J. C. Analysis of the major flavonol glycosides present in four varieties of onion (*Allium cepa*) and changes in composition resulting from autolysis. *Journal of the Science of Food and Agriculture* **1997**, *74* (3), 331-339.
6. Boyle, S. P.; Dobson, V. L.; Duthie, S. J.; Kyle, J. A. M.; Collins, A. R. Absorption and DNA protective effects of flavonoid glycosides from an onion meal. *European Journal of Nutrition* **2000**, *39* (5), 213-223.
7. Femia, A. P.; Caderni, G.; Ianni, M.; Salvadori, M.; Schijlen, E.; Collins, G.; Bovy, A.; Dolara, P. Effect of diets fortified with tomatoes or onions with variable quercetin-glycoside content on azoxymethane-induced aberrant crypt foci in the colon of rats. *European Journal of Nutrition* **2003**, *42* (6), 346-352.
8. Taché, S.; Ladam, A.; Corpet, D. E. Chemoprevention of aberrant crypt foci in the colon of rats by dietary onion. *European Journal of Cancer* **2007**, *43* (2), 454-458.
9. Nemeth, K.; Piskula, M. K. Food content, processing, absorption and metabolism of onion flavonoids. *Critical Reviews in Food Science and Nutrition* **2007**, *47* (4), 397-409.
10. Lanzotti, V. The analysis of onion and garlic. *Journal of Chromatography A* **2006**, *1112* (1-2), 3-22.
11. Stan, S. D.; Kar, S.; Stoner, G. D.; Singh, S. V. Bioactive Food Components and Cancer Risk Reduction. *Journal of cellular biochemistry* **2008**, *104* (1), 339-356.
12. Kumari, K.; Augusti, K. T. Lipid lowering effect of S-methyl cysteine sulfoxide from *Allium cepa* Linn in high cholesterol diet fed rats. *Journal of Ethnopharmacology* **2007**, *109* (3), 367-371.
13. Ichikawa, M.; Mizuno, I.; Yoshida, J.; Ide, N.; Ushijima, M.; Koderá, Y.; Hayama, M.; Ono, K. Pharmacokinetics of cycloalliin, an organosulfur compound found in garlic and onion, in rats. *Journal of Agricultural and Food Chemistry* **2006**, *54*, 9811-9819.



14. Boyle, S. P.; Dobson, V. L.; Duthie, S. J.; Kyle, J. A. M.; Collins, A. R. Absorption and DNA protective effects of flavonoid glycosides from an onion meal. *European Journal of Nutrition* **2000**, 39 (5), 213-223.
15. Saude, E. J.; Adamko, D. J.; Rowe, B. H.; Marrie, T. J.; Sykes, B. D. Variation of metabolites in normal human urine. *Metabolomics* **2007**.
16. Lindon, J. C.; Holmes, E.; Nicholson, J. K. Pattern recognition methods and applications in biomedical magnetic resonance. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2001**, 39 (1), 1-40.
17. Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **1933**, 24, 417-441.
18. Gartland, K. P. R.; Beddell, C. R.; Lindon, J. C.; Nicholson, J. K. Application of Pattern-Recognition Methods to the Analysis and Classification of Toxicological Data Derived from Proton Nuclear-Magnetic-Resonance Spectroscopy of Urine. *Molecular Pharmacology* **1991**, 39 (5), 629-642.
19. Nicholson, J. K.; Lindon, J. C.; Holmes, E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **1999**, 29 (11), 1181-1189.
20. Lindon, J. C.; Nicholson, J. K.; Holmes, E.; Everett, J. R. Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids. *Concepts in Magnetic Resonance* **2000**, 12 (5), 289-320.
21. Forshed, J.; Schuppe-Koistinen, I.; Jacobsson, S. P. Peak alignment of NMR signals by means of a genetic algorithm. *Analytica Chimica Acta* **2003**, 487 (2), 189-199.
22. Slupsky, C. M.; Rankin, K. N.; Wagner, J.; Fu, H.; Chang, D.; Weljie, A. M.; Saude, E. J.; Lix, B.; Adamko, D. J.; Shah, S.; Greiner, R.; Sykes, B. D.; Marrie, T. J. Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. *Analytical Chemistry* **2007**, 79 (18), 6995-7004.
23. Harker, M.; Coulson, H.; Fairweather, I.; Taylor, D.; Daykin, C. A. Study of metabolite composition of eccrine sweat from healthy male and female human subjects by <sup>1</sup>H NMR spectroscopy. *Metabolomics* **2006**, 2 (3), 105-112.
24. Roldán-Marín, E.; Krath, B. N.; Poulsen, M.; Binderup, M. L.; Hansen, M.; Langkilde, S.; Cano, M. P.; Sánchez-Moreno, C.; Dragsted, L. O. Effect of an onion by-product and two derived onion sub-fractions on bioactivity and safety markers in healthy rats. *British Journal of Nutrition* **2009**, Submitted November 2008.
25. Bax, A. A spatially selective composite 90-degrees radiofrequency pulse. *J. Magn. Reson.* **1985**, 65 (1), 142-145.
26. Wold, S.; Martens, H.; Wold, H. The Multivariate Calibration-Problem in Chemistry Solved by the PLS Method. *Lecture Notes in Mathematics* **1983**, 973, 286-293.
27. Nørgaard, L.; Bro, R.; Westad, F.; Engelsen, S. B. A modification of canonical variates analysis to handle highly collinear multivariate data. *Journal of Chemometrics* **2006**, 20 (8-10), 425-435.
28. Krzanowski, W. J. *Principles of Multivariate Analysis: A User's Perspective*; Revised edition ed.; Oxford Science Publication: New York, 2000.
29. Rao, C. R. *Advanced Statistical Methods in Biometric Research*; Wiley: New York, 1952.
30. Nørgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B. Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Appl. Spectrosc.* **2000**, 54 (3), 413-419.
31. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. Introduction to multi- and megavariable data analysis using projection methods (PCA & PLS). In *Scaling*, Umetrics: 1999; pp 213-225.
32. Anthony, M. L.; Gartland, K. P. R.; Beddell, C. R.; Lindon, J. C.; Nicholson, J. K. Cephaloridine-Induced Nephrotoxicity in the Fischer-344 Rat - Proton NMR Spectroscopic Studies of Urine and Plasma in Relation

to Conventional Clinical Chemical and Histopathological Assessments of Nephronal Damage. *Archives of Toxicology* **1992**, 66 (8), 525-537.

33. Beckonert, O.; Keun, H. C.; Ebbels, T. M. D.; Bundy, J.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols* **2007**, 2 (11), 2692-2703.
34. Constantinou, M. A.; Papakonstantinou, E.; Spraul, M.; Sevestiadou, S.; Costalos, C.; Koupparis, M. A.; Shulpis, K.; Tsantili-Kakoulidou, A.; Mikros, E. <sup>1</sup>H NMR-based metabonomics for the diagnosis of inborn errors of metabolism in urine. *Analytica Chimica Acta* **2005**, 542 (2), 169-177.
35. Ebbels, T. M. D.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Evaluation of metabolic variation in normal rat strains from a statistical analysis of H-1 NMR spectra of urine. *Journal of Pharmaceutical and Biomedical Analysis* **2004**, 36 (4), 823-833.
36. Holmes, E.; Nicholson, J. K.; Nicholls, A. W.; Lindon, J. C.; Connor, S. C.; Polley, S.; Connelly, J. The identification of novel biomarkers of renal toxicity using automatic data reduction techniques and PCA of proton NMR spectra of urine. *Chemometrics and Intelligent Laboratory Systems* **1998**, 44 (1-2), 245-255.
37. Law, W. S.; Huang, P. Y.; Ong, E. S.; Ong, C. N.; Li, S. F. Y.; Pasikanti, K. K.; Chan, E. C. Y. Metabonomics investigation of human urine after ingestion of green tea with gas chromatography/mass spectrometry, liquid chromatography/mass spectrometry and <sup>1</sup>H NMR spectroscopy. *Rapid Communications in Mass Spectrometry* **2008**, 22 (16), 2436-2446.
38. Wei, L.; Liao, P. Q.; Wu, H. F.; Li, X. J.; Pei, F. K.; Li, W. S.; Wu, Y. J. Toxicological effects of cinnabar in rats by NMR-based metabolic profiling of urine and serum. *Toxicology and Applied Pharmacology* **2008**, 227 (3), 417-429.
39. Holmes, E.; Bonner, F. W.; Sweatman, B. C.; Lindon, J. C.; Beddell, C. R.; Rahr, E.; Nicholson, J. K. Nuclear-Magnetic-Resonance Spectroscopy and Pattern-Recognition Analysis of the Biochemical Processes Associated with the Progression of and Recovery from Nephrotoxic Lesions in the Rat Induced by Mercury(II) Chloride and 2-Bromoethanamine. *Molecular Pharmacology* **1992**, 42 (5), 922-930.
40. Holmes, E.; Foxall, P. J. D.; Nicholson, J. K.; Neild, G. H.; Brown, S. M.; Beddell, C. R.; Sweatman, B. C.; Rahr, E.; Lindon, J. C.; Spraul, M.; Neidig, P. Automatic Data Reduction and Pattern-Recognition Methods for Analysis of <sup>1</sup>H Nuclear-Magnetic-Resonance Spectra of Human Urine from Normal and Pathological States. *Analytical Biochemistry* **1994**, 220 (2), 284-296.
41. Stanley, E. G.; Bailey, N. J. C.; Bollard, M. E.; Haselden, J. N.; Waterfield, C. J.; Holmes, E.; Nicholson, J. K. Sexual dimorphism in urinary metabolite profiles of Han Wistar rats revealed by nuclear-magnetic-resonance-based metabonomics. *Analytical Biochemistry* **2005**, 343 (2), 195-202.
42. Andersson, C. A. Direct orthogonalization. *Chemometrics and Intelligent Laboratory Systems* **1999**, 47 (1), 51-63.
43. Kumari, K.; Augusti, K. T. Lipid lowering effect of S-methyl cysteine sulfoxide from *Allium cepa* Linn in high cholesterol diet fed rats. *Journal of Ethnopharmacology* **2007**, 109 (3), 367-371.
44. Hucker, H. B.; Ahmad, P. M.; Miller, E. A.; Brobyn, R. Metabolism of Dimethyl Sulphoxide to Dimethyl Sulphone in Rat and Man. *Nature* **1966**, 209 (5023), 619-621.
45. Debons, A. F.; Fani, K.; Jimenez, F. A.; Maayan, M. L. Inhibition of Cholesterol-Induced Atherosclerosis in Rabbits by Dimethylsulfoxide. *Journal of Pharmacology and Experimental Therapeutics* **1987**, 243 (2), 745-757.
46. Layman, D. L.; Alam, S. S.; Newcomb, K. C. Suppression of Atherosclerosis in Cholesterolemic Rabbits by Dimethylsulfoxide. *Annals of the New York Academy of Sciences* **1983**, 411 (JUN), 336-339.
47. Kedar, I.; Sohar, E. Dimethylsulfoxide in Experimental Atherosclerosis of the Rabbit. *Israel Journal of Medical Sciences* **1981**, 17 (4), 289-291.
48. Cork, A.; Park, K. C. Identification of electrophysiologically-active compounds for the malaria mosquito, *Anopheles gambiae*, in human sweat extracts. *Medical and Veterinary Entomology* **1996**, 10 (3), 269-276.

49. Waring, R. H.; Mitchell, S. C.; Fenwick, G. R. The Chemical Nature of the Urinary Odor Produced by Man After Asparagus Ingestion. *Xenobiotica* **1987**, *17* (11), 1363-1371.
50. Mitchell, S. C.; Waring, R. H.; Land, D.; Thorpe, W. V. Odorous Urine Following Asparagus Ingestion in Man. *Experientia* **1987**, *43* (4), 382-383.
51. Toso, B.; Procida, G.; Stefanon, B. Determination of volatile compounds in cows' milk using headspace GC-MS. *Journal of Dairy Research* **2002**, *69* (4), 569-577.
52. Gallagher, M.; Wysocki, J.; Leyden, J. J.; Spielman, A. I.; Sun, X.; Preti, G. Analyses of volatile organic compounds from human skin. *British Journal of Dermatology* **2008**, *159* (4), 780-791.
53. Gallagher, M.; Preti, G.; Fakharzadeh, S.; Wysocki, J.; Kwak, J.; Miller, C. J.; Schmults, C. D.; Spielman, A. I.; Sun, X. Detecting skin cancer using volatile biomarkers. The 236th ACS National Meeting, Philadelphia, PA, August 17-21, 2008 AGFD227. 2008.  
Ref Type: Abstract
54. Engelke, U. F. H.; Tangerman, A.; Willemsen, M. A. A. P.; Moskau, D.; Loss, S.; Mudd, S. H.; Wevers, R. A. Dimethyl sulfone in human cerebrospinal fluid and blood plasma confirmed by one-dimensional H-1 and two-dimensional H-1-C-13 NMR. *NMR in Biomedicine* **2005**, *18* (5), 331-336.
55. Willett, W. *Nutritional epidemiology*; 2 ed.; Oxford University Press: Oxford, United Kingdom, 1998.
56. McKeown, N. M.; Day, N. E.; Welch, A. A.; Runswick, S. A.; Luben, R. N.; Mulligan, A. A.; McTaggart, A.; Bingham, S. A. Use of biological markers to validate self-reported dietary intake in a random sample of the European Prospective Investigation into Cancer United Kingdom Norfolk cohort. *American Journal of Clinical Nutrition* **2001**, *74* (2), 188-196.

## Paper IV

*Quantification of the degree of blockiness in pectins using  $^1\text{H}$  NMR spectroscopy and chemometrics*

H. Wining, N Viereck, L. Nørgaard, J. Larsen, S.B. Engelsen,

Food Hydrocolloids **21** (2007) 256–266.





ELSEVIER

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

ScienceDirect

Food Hydrocolloids 21 (2007) 256–266

FOOD  
HYDROCOLLOIDS[www.elsevier.com/locate/foodhyd](http://www.elsevier.com/locate/foodhyd)

## Quantification of the degree of blockiness in pectins using $^1\text{H}$ NMR spectroscopy and chemometrics

Hanne Winning<sup>a,\*</sup>, Nanna Viereck<sup>a</sup>, Lars Nørgaard<sup>a</sup>, Jan Larsen<sup>b</sup>, Søren B. Engelsen<sup>a</sup><sup>a</sup>Centre for Advance Food Studies, Quality and Technology, The Royal Veterinary and Agricultural University, Rolighedsvej 30, 1958 Frederiksberg C, Denmark<sup>b</sup>CP Kelco, Ved Banen 16, 4623 Lille Skensved, Denmark

Received 26 September 2005; accepted 29 March 2006

### Abstract

The gelling properties of pectins are known not only to be closely related to the degree of esterification (DE), but also to the distribution of the ester groups. In this study we have examined an experimentally designed series of pectins originating from the same mother pectin and deesterified using combinations of two different enzymatic mechanisms. The DE and distribution patterns of methyl ester groups have been analyzed using high-resolution (HR)  $^1\text{H}$  nuclear magnetic resonance (NMR) spectroscopy on pectin solutions. Quantitative calibration models using partial least squares (PLS) regression were developed with the ability to predict DE as well as the specific enzyme treatment, expressed as amount of ester groups removed with random and block enzyme, respectively. NMR spectroscopy was able to distinguish between enzyme treatments in simple classification by principal component analysis (PCA). This was due to the spatial structure of pectin together with the methyl ester distribution. Nuclear Overhauser effect spectroscopy (NOESY) experiments confirmed all the general assignments with the expected nuclear Overhauser effect (NOE) correlations. Degree of random deesterification (R) was better predicted than the degree of block deesterification (B). The calibration models for prediction of R obtained on extended inverted signal correction (EISC) processed data gave a root mean square error (RMSE) of cross-validation (CV) of 2%p with 4 PLS components (latent variables, LV) and a correlation coefficient ( $r$ ) of 0.98. Spectral variable selection using interval PLS (iPLS) was shown to be valuable, as all the calibration models were improved.

© 2006 Elsevier Ltd. All rights reserved.

**Keywords:** Pectin; Noesy; NMR; Chemometrics; Esterification; Blockiness; Block-enzyme; Random-enzyme

### 1. Introduction

Pectic molecules include primary homogalacturonan (HGA) and rhamnogalacturonan I (RG I). HGA is composed of unbranched  $\alpha$ -1,4-linked galacturonic acid (GalA) residues, whereas RG I is composed of a backbone of repeating  $\alpha$ -1,2-L-rhamnose(Rhap)- $\alpha$ -1,4-D-GalA disaccharide units. Side-chains, mainly consisting of arabinan and/or galactan, are attached to the RG I backbone at the C-4 position of the Rhap residues. These side-chains are called 'hairy regions' (Carpita & Gibeaut, 1993; Ridley, O'Neill, & Mohnen, 2001). The GalA residues in the HGA and RG I backbone may be methyl esterified and/or O-acetylated (Carpita & Gibeaut, 1993). The structure of pectin is important for the plant cell-wall strength and flexibility. Because of its excellent gelling, thickening, and

**Abbreviations:** GalA, galacturonic acid; HGA, homogalacturonan; DE, degree of esterification; DP, degree of polymerization; RG-I, Rhamnogalacturonan-I; Rhap, rhamnopyranosyl unit; PME, pectin methyl esterase; B, block deesterification; R, random deesterification; NMR, nuclear magnetic resonance; HR, high resolution; NOE, nuclear Overhauser enhancement; FID, free induction decay; FT, Fourier transformed; TSP, perdeuterated 3-(trimethylsilyl) propionate sodium salt; PC, principal component; PCA, principal component analysis; CV, cross validation; LV, latent variables; RMSECV, root mean square error of cross validation; PLS, partial least squares; iPLS, interval partial least squares; MSC, Multiplicative scatter correction; EISC, Extended inverted signal correction; %p, percent points; ppm, parts per million; Å, Ångström; Nd, not determined; IPA, Isopropanol

\*Corresponding author.

E-mail address: [hie@kvl.dk](mailto:hie@kvl.dk) (H. Winning).

stabilizing properties, the polymer is extensively utilized in the food industry (Ralet, Bonnin, & Thibault, 2002; Rolin, 1993).

The degree and pattern of methyl esterification has important commercial implications and has been the subject of many studies because of its effect on the rheological and gel-forming properties of pectins (Rolin, 1993). Daas, Meyer-Hansen, Schols, De Ruiter, & Voragen (1999) introduced the term ‘degree of blockiness’ as the total amount of free GalA residues expressed as the percentage of the total number of free GalA residues percent in the pectin. Previous investigations of the methyl ester distribution of pectins include  $^1\text{H}$  nuclear magnetic resonance (NMR) spectroscopy combined with enzymatic degradation, (Andersen, Larsen, & Grasdalen, 1995; Grasdalen, Andersen, & Larsen, 1996) and enzymatic degradation combined with mass spectrometry or high performance chromatography (Daas et al., 1999; Daas, Voragen, & Schols, 2000; Daas, Voragen, & Schols, 2001; Korner, Limberg, Christensen, Mikkelsen, & Roepstorff, 1999; Limberg et al. 2000b; Limberg et al. 2000a; Ralet, Dronnet, Buchholt, & Thibault, 2001; Ralet & Thibault, 2002; Willats, McCartney, Mackie, & Knox, 2001). These studies all concluded that alkaline deesterification and fungus-pectin methyl esterase (PME) resulted in random distribution, while plant-PME resulted in blockwise distribution of methyl ester groups. Elaborate studies made by Limberg et al. (2000a) suggested that deesterification using plant-PME is a process which selectively

introduces block structure of adjacent free GalA units, whereas fungus-PME and base treatment lead to two different forms of homogenous methyl esterification patterns. This finding led to the hypothesis that random deesterification is more ordered than first assumed (Limberg et al., 2000b). The terms ‘random’ and ‘blocky’ methyl ester distribution are merely interpretations and there is no precise definition of how many contiguous free GalA units constitute a block. In addition, random distribution can be completely random or almost systematic (Fig. 1).

The functional traits of the methyl ester pattern have been investigated in calcium-binding studies (Ralet, Crepeau, Buchholt, & Thibault, 2003; Ralet et al., 2001; Willats et al., 2001). These studies showed that the calcium sensitivity of high-ester pectins increased in the presence of inhomogeneous methyl ester distribution. Regions of free GalA groups form multiple calcium bridges, which create a domain of strong, intermolecular association between the HGA chains, resulting in increased viscosity. Multiple regions of predominant-free GalA residues in the presence of an excess of calcium ions or other divalent cations in water causes pectin gelling or precipitation (Ralet et al., 2001; Ralet et al., 2003). For a gel to be formed and not to be brittle, features such as the helical conformation of the pectic polymer are important. Braccini, Grasso, & Perez (1999) calculated the difference in energy between the 2- and the 3-fold conformation for poly-GalA, indicating that both types of helical conformation energetically are almost equally favorable (Braccini et al., 1999).

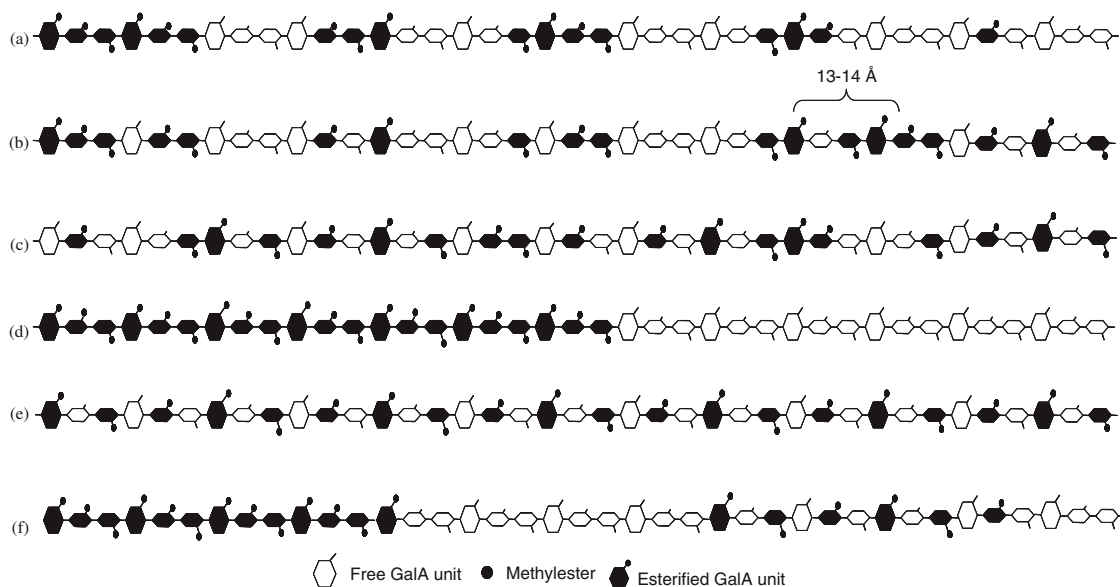


Fig. 1. Schematic example of pGalA chains. HGA with methyl groups of (a) short blockwise distribution, (b) partial blockwise with short blocks, (c) unsystematically random distribution, (d) large blocks, (e) systematically random distribution and (f) block and systematically random distribution of free GalA units. The pGalA chains are in a 3-fold conformation and the distance between two methyl ester groups are 13–14 Ångström (Å). The DE is 50% in the examples.

Thus, depending on the environment (neighboring ions, solvent, etc.), one or the other form may be favored.

Because the distribution of the free ester groups is critical for the detailed functionality of pectins, development of a fast and reliable method able to assess the ester distribution of native or semi-refined pectins is desirable. The purpose and challenge of this study was to develop an NMR method to examine the methyl ester distribution along the pectic polymer. High-resolution (HR)  $^1\text{H}$  NMR spectroscopy has previously proven not only to be a valuable tool in the determination of degree of esterification (DE) in pectins, (Andersen et al., 1995; Grasdalen et al., 1996; Rosenbohm, Lundt, Christensen, & Young, 2003), but has furthermore shown potential to analyze the sequence of free GalA and the distribution of the methyl esters (Andersen et al., 1995; Grasdalen et al., 1996). The NMR signals relating to specific protons related to block-polymer information are bound to be more subtle which is why it is necessary to apply data analytical methods such as partial least squares (PLS) (Wold, Martens, & Wold, 1983) in order to extract the desired information. For this investigation, 31 pectin samples were produced by degradation of the same mother pectin by two different PME's, a block esterase and a random esterase. The aim was to predict the amount of methyl esters, removed with a random or block esterase.

## 2. Materials and methods

### 2.1. Experimental design and samples

The high-methoxy pectin samples were designed by CP Kelco (Lille Skensved, Denmark) from a native (mother) citrus pectin (DE~72.3%) which had been remethylated to DE~93.8%. The remethylated pectin was produced from a 5°C suspension of pectin in dry methanol and added

thionyl chloride in excess over a period of 5 days. The product was separated by filtration and washed with methanol and dried. This remethylated pectin was then deesterified with two commercial pectin modifying enzymes, a random methyl esterase, Rheozym (now Novoshape<sup>®</sup>) (Novozymes, Denmark) and a block methyl esterase derived from papaya fruit. The commercial Rheozyme is cloned from *Aspergillus aculeatus* and is believed to have the same deesterification pattern as *Aspergillus niger* which is a multichain mechanism which removes the methyl groups random (CP Kelco, 2006) (van Alebeek, van Scherpenzeel, Beldman, Schols, & Voragen, 2003). The deesterification pattern of the papaya block esterase is suggested as a single chain mechanism similar to the mechanism described for other esterases from e.g. orange (Kohn, Furda, & Kopec, 1968; Kohn, Markovic, & Machova, 1983). The pectin samples were designed from a three-dimensional experimental design, illustrated in the triaxial diagram (Fig. 2). One pectin series with blockwise methyl ester distribution and one pectin series with random methyl ester distribution were produced. This yielded two groups of five samples with varying DE from about 90% in 10% steps down to 50% DE (Block: B90, B80, B70, B60, and B50 and Random: R90, R80, R70, R60 and R50). A third series was made from the block series which was subsequently deesterified with random enzyme. The fourth series was made from the random series which was subsequently deesterified with block enzyme. Thus, the total sample set consisted of four groups with different enzyme treatments; one pure random, one pure block, one first random followed by block and one first block followed by random (Fig. 2). Only one batch of each sample was produced in amounts of approximately 700 mg.

The reference DE was determined by titration and the deesterification with block and random enzyme was

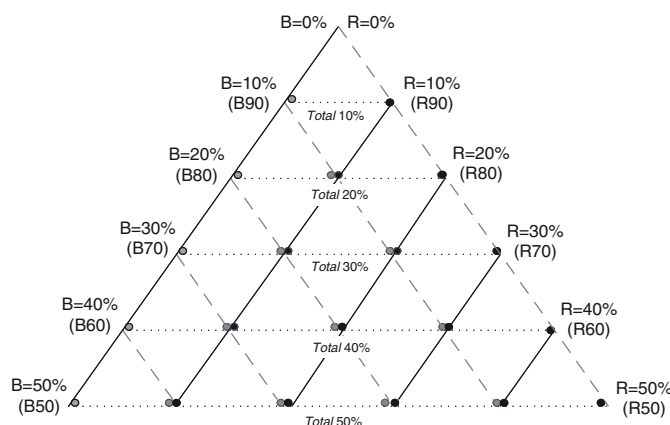


Fig. 2. Schematic model of the triaxial experimental design. Reading downwards (left): deesterification with block enzyme from DE~100 to DE 90, 80, 70, 60 and 50. Reading downwards (right): deesterification with random enzyme from DE~100 to DE 90, 80, 70, 60 and 50. The horizontal lines indicate the total deesterification.



thereby calculated. The designed pectins had a molecular weight of 10,000–15,000 g/mol, determined by a triple detector size exclusion chromatography module (Viscotek), which corresponds to a degree of polymerization (DP) of approximately 50 GalA units. There was a remarkable drop in molecular weight (from 180 to 13 kDa), as a result of the remethylation which suggests a removal of the hairy regions during the remethylation process. The measured DEs are shown in Tables 1 and 2, together with the precise deesterification with block and random enzyme. As the mechanisms of the two enzymes are not fully elucidated, the amounts of ester groups removed with random or block esterase, were used as reference values in the calibration models. The DE is calculated in percent of the number of esterified GalA units compared to the total number of GalA units and is denoted %point (%p), which is also the unit for block and random deesterification. Furthermore, it is important to limit the number of latent variables (LVs) to obtain parsimonious models and to avoid over-fitting when test-set validation is not possible. This is especially important when the sample set is small, as in the present study. All the calibration models were validated using full cross-validation (CV) (Wold, 1978).

The pH was measured in 1% (w/w) solutions of all samples with Baker-pHIX (pH 0.0–6.0) (J.T. Baker, Phillipsburg, USA). The total sample set consisted of 31 designed pectins (sample R90-B50 failed in the production). Furthermore, two identical remethylated pectins, the mother pectin and a pectate sample (DE~0%) were included in the study. These samples were used to span the maximum variation in the calculated calibration models. The samples order was randomized prior to measurements.

## 2.2. NMR spectroscopy

<sup>1</sup>H NMR spectra were recorded on pectin solutions (0.5% w/w), but due to the small sample amounts only single determinations were performed. The spectra were acquired at 9.4 T on a Bruker Avance Ultra Shield 400 spectrometer (Bruker Biospin GmbH, Rheinstetten, Germany) operating at 400.13 MHz at 303 K using a broad band inverse probe head. The pectin samples were prepared as 0.5% (w/w) solutions in D<sub>2</sub>O and added 10% (v/v) perdeuterated 3-(trimethylsilyl) propionate sodium salt (TSP) as chemical shift reference ( $\delta = 0.0$  ppm). The spectra were accumulated with a 30° pulse, an acquisition time of 1.98 s, a recycle time (relaxation delay) of 8 s, 128 scans and a sweep width of 8278.15 Hz, resulting in 32 K complex data points. A pulse experiment with water suppression (composite pulses) was chosen to avoid overlap of pectin signals and the water signal. A spectral window of 3–5.5 ppm was analyzed. All assignments were made on raw spectra.

Nuclear Overhauser effect spectroscopy (NOESY) spectra were recorded on five selected samples, B50, B90, R50, B50 and B70-R50, with mixing time of either 300 or 350 ms and a relaxation delay of 4.5 s. The size of the NOESY spectra was 2048 × 256 and they were recorded at 303 K in the phase-sensitive mode (TPP). The number of scans per  $t_1$  increment was 32 and the spectral width was 728.438 Hz (1.82 ppm).

## 2.3. Chemometric analysis

Multivariate data analysis was applied to the NMR spectra in order to extract optimal quantitative and

Table 1  
Reference deesterification determinations with random enzyme treatment of design pectins

Target deesterification code	1. Deesterification random enzyme DE	2. Deesterification block enzyme DE	Random enzyme treated %p	Block enzyme treated %p
R-90	93.8 → 88.9		4.9	0
R-90-B80		88.9 → 78.3		10.6
R-90-B70		88.9 → 70.9		18.0
R-90-B60		88.9 → 64.4		24.5
R-90-B50		nd <sup>a</sup>	—	—
R-80	93.8 → 80.0		13.8	0
R-80-B70		80.0 → 72.2		7.8
R-80-B60		80.0 → 66.2		13.8
R-80-B50		80.0 → 56.2		23.8
R-70	93.8 → 71.9		21.9	0
R-70-B60		71.9 → 62.2		9.7
R-70-B50		71.9 → 50.3		21.6
R-60	93.8 → 64.9		28.9	0
R-60-B50		64.9 → 51.7		13.2
R-50	93.8 → 58.8		35.0	0
R-50-B50		58.8 → 52		6.8

<sup>a</sup>nd: not determined.

Table 2  
Reference deesterification determinations with block enzyme treatment of design pectins

Target deesterification code	1. Deesterification block enzyme DE	2. Deesterification random enzyme DE	Block enzyme treated DE	Random enzyme treated DE
B-90	93.8 → 89.4		4.4	0
B-90-R80		89.4 → 81.1		8.3
B-90-R70		89.4 → 73.8		15.6
B-90-R60		89.4 → 66.6		22.8
B-90-R50		89.4 → 57.5		31.9
B-80	93.8 → 79.7		13.9	0
B-80-R70		79.7 → 69.5		10.2
B-80-R60		79.7 → 59.4		20.3
B-80-R50		79.7 → 49.1		30.6
B-70	93.8 → 71.2		22.6	0
B-70-R60		71.2 → 60.2		11.0
B-70-R50		71.2 → 47.3		23.9
B-60	93.8 → 60.3		33.5	0
B-60-R50		60.3 → 50.8		9.5
B-50	93.8 → 57.0		36.8	0
B-50-R50		57.0 → 50.6		6.4

qualitative information. PLS regression models were developed for DE, block deesterification (B) and random deesterification (R). Subsequent PLS regression models were made with interval PLS (iPLS) (Nørgaard et al., 2000). All models were full CV unless otherwise stated. Different transformation techniques (none, scaling after TSP signal, shifting by covariance optimization (co-shifting) (Tomasi, van den Berg, & Andersson, 2004), warping (Tomasi et al., 2004), multiplicative scatter correction (MSC) (Geladi, Macdougall, & Martens, 1985) and extended inverted signal correction (EISC) (Martens, Nielsen, & Engelsen, 2003)) were compared by analyzing the effect on the calibration (PLS) models as well as by visually evaluating the spectra. Models were tested in order to find the best model with the lowest cross-validated prediction error. Principal component analysis (PCA) (Wold, Esbensen, & Geladi, 1987) models were calculated for classification of enzyme treatment.

#### 2.4. Software

The multivariate data analysis was carried out using Unscrambler 8.0 (CAMO, Inc., Trondheim, Norway) and MATLAB 6.5 (The MathWorks, Inc., Natick, Massachusetts, USA). MATLAB codes for the preprocessing techniques (co-shifting, warping and EISC) and for iPLS can be found at [www.models.kvl.dk](http://www.models.kvl.dk).

### 3. Results and discussion

The  $^1\text{H}$  NMR spectra of the samples that were not part of the designed pectins, i.e. the mother pectin, pectate and remethylated pectin, proved to be significantly different.

The pH values of these samples were also different, as the pH for the designed pectins and mother pectin were measured to be 3.0–4.0 and for the pectate and the remethylated pectin was 2.0 and 4.5, respectively. In order to give an overview of the samples, PCA models were calculated. A PCA score plot of the NMR spectra revealed that the mother pectin was an outlier. This was confirmed by inspection of the spectrum of the mother pectin, revealing a number of extra signals which were not observed in the spectra of the designed pectins. These signals could be rhamnose or arabinose and galactose remnants from hairy regions; however, by comparing spectra of these compounds, this could not be confirmed. The designed pectins were deesterified with enzymes and it is assumed that the enzyme treatment cleaves the hairy regions from the pectin polymer. For this reason, the mother pectin, the remethylated pectin and the pectate were removed from the calibration set before modeling.

#### 3.1. NMR assignment

After preprocessing the data, but before modeling, the spectral region 0–3.3 ppm was removed due to of the absence of pectin signals. The excluded region included one large doublet at 1.14 ppm (1.15 and 1.13 ppm) assigned to isopropanol (IPA), used in pectin production to precipitate pectin from the aqueous solution. This implies that the signal at 4.0 ppm contains a significant contribution from IPA (Fig. 3). Fig. 3 displays the mean  $^1\text{H}$  NMR spectrum of the pectin data. The spectra of the pectin samples contain a sharp signal at 3.78 ppm which is the signal from the protons in the methoxy groups of the esterified pectin.

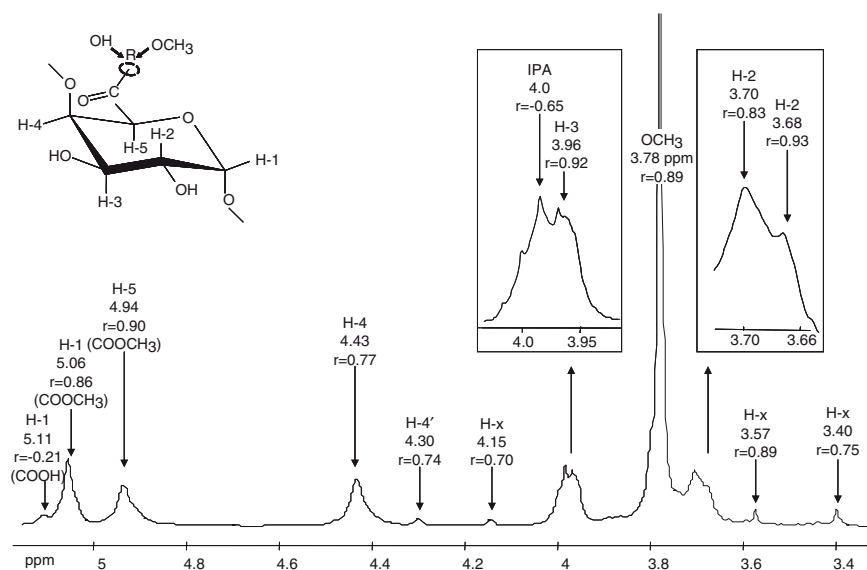


Fig. 3. Mean  $^1\text{H}$  NMR spectrum (untreated) of the pectin samples. Arrows show the assignment of the signals and the correlations coefficients to DE.

The direct correlation coefficients to DE indicated in Fig. 3 are calculated on raw NMR spectra. As expected, the intensity of the methoxy signal at 3.78 ppm showed good direct correlation to DE (0.89), but the higher field signal at 3.68 ppm ('shoulder') yielded a very good correlation (0.93) to DE as well. Furthermore, the correlation between the methoxy signal and the shoulder was 0.93. Rosenbohm et al. (2003) found that the H-2 and H-3 protons absorb around 3.7 and 4.0 ppm, respectively. For this reason, the signal at 3.70 ppm was assigned to H-2 and the signal at 3.96 ppm to H-3. This was also confirmed by Ló et al. (1994) who found the H-2 signals in heptagalacturonide around 3.77 ppm (Ló, Hahn, & Halbeek, 1994).

The good correlation between DE and the methoxy signal to the H-2 signal can be explained by large NOE effects between the methoxy group and H-2. The 3-fold helical turn of the pectic backbone has the consequence in the spectra that the difference in signal intensity is due to an NOE effect in the polymer. In this case, the protons in the GalA unit, through space connectivity, are able to sense whether or not an ester group is spatially located in close proximity. As a result, the esterification of a given pectin unit will affect the signal intensity of the protons spatial located just above or beneath the pectin unit. Furthermore, the positions of the signals in the spectra will be dependent on whether the pectin polymer is random or block deesterified. Furthermore, the shielding of the protons and hence their chemical shift values will be affected by the functional group positioned above or beneath it. Similarly, the H-3 signal at 3.96 ppm also showed a good correlation

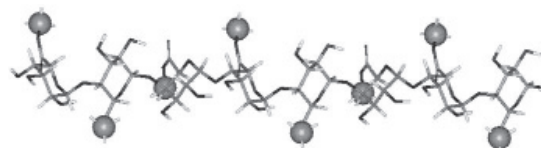


Fig. 4. Computer model of a 3-fold helical turn of an eight monomer GalA units of the pectic backbone with enlarged methylester groups.

(0.92) to DE, which is most likely also due to the NOE effect. A computer model of a 3-fold helical turn of an eight monomer GalA unit (fully esterified) of the pectic backbone supports this NOE hypothesis (Fig. 4). The distances between the methoxy group at one unit ( $n$ ) and the H-1, H-3, H-4 and the H-5 of the two neighbor GalA units ( $n-1$  and  $n+1$ ) seem very short. This structure is only approximate, as the position of the methoxy group is quite flexible.

In order to investigate the spatial structure, NOESY experiments were carried out. The NOESY spectra confirmed the assignments, as the expected NOE correlations appeared in the spectrum (Fig. 5). Due to spectral overlap, the NOE correlation between the methoxy group and H-2 could not be unequivocally confirmed in the NOESY spectrum.

Rosenbohm et al. (2003) also stated that the H-4 protons were absorbing at around 4.4 ppm as two signals depending on DE, which match the findings of this study. The connectivity in the NOESY spectrum in Fig. 5 at 4.30 ppm

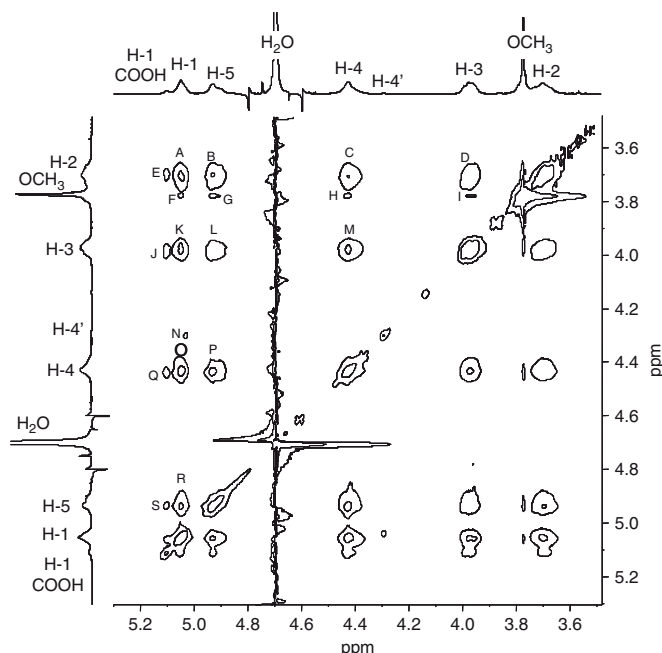


Fig. 5. Part of NOESY spectrum of pectin sample B70-R50 (DE~50). The letters represent the NOE correlations between the atom nuclei in the GalA residue. A: H1-H2, B: H5-H2, C: H4-H2, D: H3-H2, E: H1'-H2, F: H1-OCH<sub>3</sub>, G: H5-OCH<sub>3</sub>, H: H4-OCH<sub>3</sub>, I: H3-OCH<sub>3</sub>, J: H1'-H3, K: H1-H3, L: H5-H3, M: H4-H3, N: H1-H4', O: H1-H4, P: H5-H4, Q: H1'-H4, R: H1-H5, S: H1'-H5.

was suggested identified as H-4' (end residue) which is spatially connected to H-1 in the neighboring residue (Fig. 4). Furthermore, the H-4' has a weak connectivity to H-2 together with the methoxy protons placed at the same residue, (seen when lowering the level of the correlation plot, not shown). Similar correlations are clearly observed between H-4 and H-1 and between H-4 and H-2.

According to Andersen et al. (1995), a block-type distribution in enzyme-treated samples is indicated by stronger lines in the spectra corresponding to contiguous arrangements of esterified and deesterified units. Two esterified GalA units should have stronger H-1 lines at 4.98 ppm, four esterified units at 5.03 and two free GalA units at 5.10 ppm. The positions of the signals are very similar to the signals found in this study at 5.06 and 5.11 ppm together with a small shoulder on the H-5 signal at 4.96 ppm (only visible in the raw spectra). These assignments are in perfect agreement with the work of Rosenbohm et al. (2003). They found that the H-5 protons adjacent to free GalA units appeared as two signals at 4.9 and 4.6 ppm, whereas the signals for H-5 protons adjacent to ester groups had shifted downhill to about 5.0 ppm.

The two signals to the very high-field region of the selected area (3.57 and 3.40 ppm) could not be assigned in the present study. There are no literature reports on pectic signals in this region and there were no NOE effects to

support coherence with the remaining protons in the pectins. Similarly, the signal at 4.15 ppm could not be assigned. As previously mentioned, spectra of monosaccharide compounds which could be present in hairy regions (galactose, rhamnose and arabinose) were tested in order to assign these signals, but without any results.

### 3.2. Prediction models

The primary objective of this study was to explore the possibility of predicting the methyl ester distribution in new pectin samples by their NMR spectra, since the spectra contain regions with information about the degree of blockiness. When developing PLS calibration models, normally the primary aim is to obtain as small validated prediction errors, root mean square error of cross validation (RMSECV) as possible. Different transformation techniques (co-shifting, warping, MSC and EISC) were tested in order to improve the PLS models. However, none of the tested techniques were able to improve the PLS models for the prediction of DE and B. Only preprocessing of data with EISC improved the prediction of R considerably.

The PLS model obtained from raw NMR spectra yielded good prediction of DE using only 3 LVs with a correlation, *r*, of 0.97 and a RMSECV of 3.0 %. Even for PLS models

Table 3  
Results of cross-validated prediction models using raw and EISC treated NMR spectra

Range		DE~47–94	B~0–35	R~0–37
Raw	LV	3	8	8
	RMSECV	3.0	3.8	2.7
	<i>r</i>	0.97	0.94	0.97
EISC	LV	3	9	4
	RMSECV	5.0	3.7	2.0
	<i>r</i>	0.91	0.95	0.98

based on spectra without the methoxy signal, the correlation remained high,  $r = 0.96$ , and RMSECV of 3.3 %p using the same number of LVs. These results indicate that there are several signals that covariate with DE in agreement with the correlation coefficients to DE indicated in Fig. 3. In the prediction of the distribution of free GalA units expressed as B and R values, the best PLS model for the prediction of R was attained by the EISC treated data using 4 LVs, yielding a prediction error of 2.0 %p and a correlation coefficient of 0.98. The prediction of R was generally better than the prediction of B (Table 3).

### 3.3. Variable selection using iPLS

In order to improve the calibration models and to investigate the influential areas of the spectra, iPLS were employed. iPLS is an extension of PLS which develops local PLS models on a number of subintervals of the full spectrum region. The main advantage of iPLS is that it provides an overall picture of the relevant information in different spectral subdivisions, thereby removing interferences from other regions (Nørgaard et al., 2000).

Variable selection using iPLS was applied in this investigation primarily to visualize the areas of the spectra which include information about the methyl ester pattern. The spectrum was divided into 20 equally sized subintervals and PLS models were made from each interval. While no single interval was found by iPLS which could improve the prediction of DE or B, iPLS extracted NMR spectra proved to be a most precise tool in the determination of R. The influential interval for the prediction of R found by iPLS is displayed in Fig. 6. iPLS models calculated on only one single interval (3.7–3.6 ppm) resulted in lower prediction errors (1.8 %p) than the global model (2.0 %p) (Table 3). The spectral region from 3.9 to 3.6 ppm (inserted as an enlargement in Fig. 6) had generally low prediction errors. For this reason, this area was investigated for elucidating further information about the specific signals contributing to the model. The dot-and-dash line in Fig. 6 is the spectrum for the pectin with the lowest R value and the full-drawn line corresponds to the pectin with the highest R value. In the inserted enlargement, the chemical shift of the signal assigned to H-2 is shifted depending on the ester

distribution. For this reason, the most important area of the spectra with regard to the random deesterification is the H-2 signal. This is due to the spatially close position (Fig. 4) of the ester group to H-1 and the signals of the H-2 protons are thereby affected when the neighboring GalA residues are esterified.

### 3.4. Classification by enzyme treatment

To investigate the regions of importance, with regard to grouping by enzyme treatment, discriminant iPLS was carried out (Fig. 7). This yielded three areas of interest: interval 1–3 (5.2–4.9 ppm), interval 10 (4.3–4.2 ppm) and interval 17–18 (3.7–3.5 ppm). These intervals corresponded to H-1 (free and esterified), H-5, H-4' (end residue), H-2 and to the signal H-x at 5.57 ppm which we have not been able to assign. PCA models were carried out in order to classify by enzyme treatment. In Fig. 7, the small plot shows the PCA score plot. The grouping by deesterification treatment ( $\blacktriangle \sim R$ ,  $\triangle \sim B-R$ ,  $\blacksquare \sim B$ ,  $\square \sim R-B$ ) was enhanced when the PCA models were only calculated on intervals found significant by the discriminant iPLS models. The score plot demonstrates that the samples are grouped by the last enzyme treatment, as the R and B-R samples and the B and R-B samples are placed together. A PCA calculated on the 3.5–3.7 ppm (H-2) region confirmed the presence of variance in this region, indicating spectral differentiation between the different enzyme treatments. In addition, the signal of H-1 at 5.11 ppm is of interest, since the signal intensity increased with increasing random deesterification, with sample R50 being the most intense (Fig. 6). However, the H-1 signal could, as the H-2 region, only differentiate between the pure block and the pure random samples with tendencies towards division of the rest. This indicates that the H-2 and H-1 area is of great importance in the prediction of the ester group decoration of the pectin molecule. It is thus evident that the H-1 signal strongly covariates with random deesterification, but not with the blockiness. This could indicate that it is the degree of randomness that shows these characteristic spectral shapes. There were no intervals found by iPLS able to provide a clear grouping of all four sample set groups.

## 4. Conclusion

We have investigated whether NMR has the potential to characterize the distribution of methyl ester and carboxyl groups in enzyme-treated pectin. Quantitative models determining the pectic composition of untreated powders were developed using  $^1\text{H}$  NMR spectra of pectic solutions. The experimental design, which consisted of pectin samples treated by block deesterification, random deesterification, block-random deesterification, and random-block deesterification, could be reflected in PCA score plots. Very good calibration models were obtained, with a correlation of 0.97 for the prediction of DE using only 3 LVs with a prediction error of 3.0 %p. Furthermore, calibration

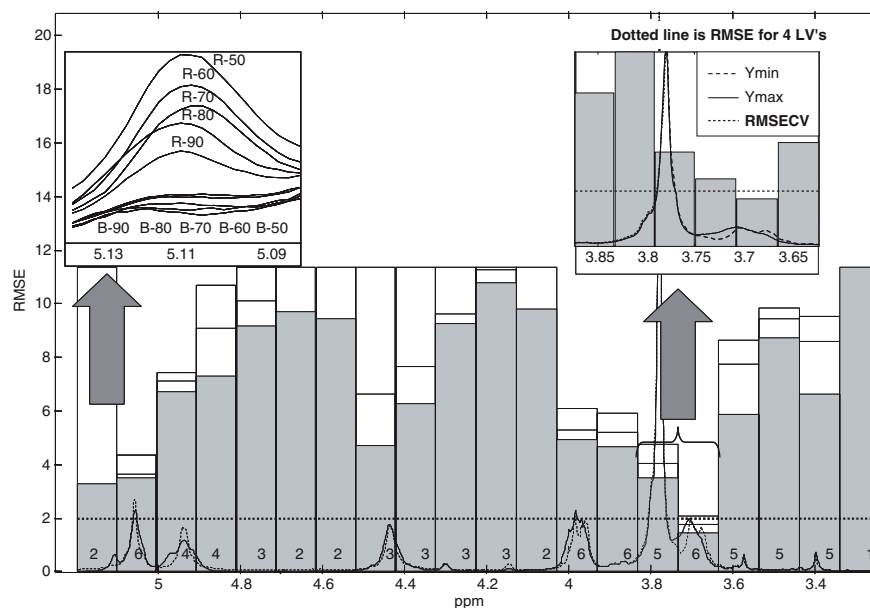


Fig. 6. iPLS plot of prediction of R by NMR. The dotted line represents RMSECV using 5 LVs for the global model. The italic numbers are optimal LVs in an interval model. The inserted small plots show an enlargement iPLS plot of a selected area from 3.9–3.6 ppm (right) and an enlargement of the area 5.12–5.08 ppm (left).

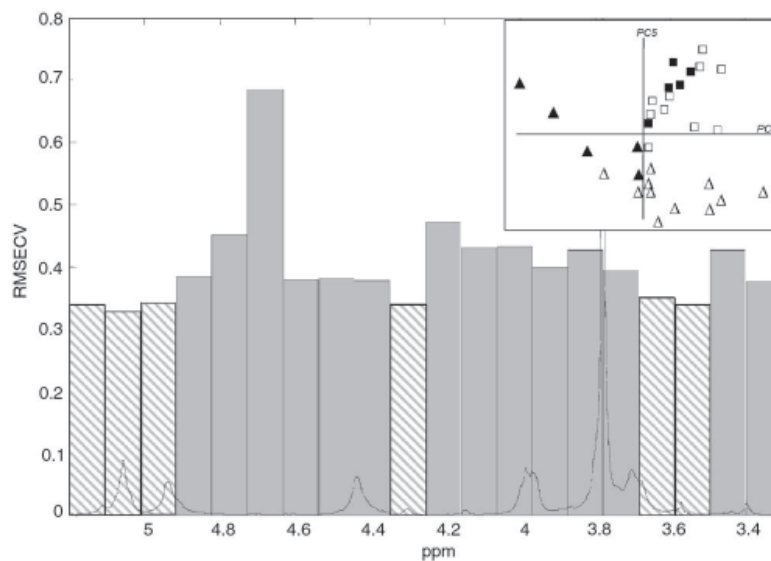


Fig. 7. iPLS plot of PLS model, discriminated by enzyme treatment and the corresponding PCA score plot (PC2 versus PC5) of the four deesterification treatments ▲~R, △~B-R, ■~B, □~R-B obtained on the hatched intervals. PC2 explains 18% of the variation, while PC5 explains 4%. The RMSECV is 0.43 for the global model using 5 LVs.

models for prediction of R using EISC preprocessed spectra gave very small RMSECV of 2 %p with 4 LVs and an  $r$  of 0.98. Using the variable selection method, iPLS,

spectral regions were found, where the models for prediction of random deesterification could be significantly improved.

The results show a clear tendency towards classifying the pectin by enzyme treatment. Unfortunately, the R-B samples were difficult to distinguish from the B samples. This was probably due to the block enzymes removing the random methyl ester pattern in the second deesterification. Specific signals were found in the NMR spectra which corresponded to the deesterification by random enzyme. For this reason, the prediction of R calibration models showed markedly higher correlation and lower prediction errors than for prediction of B. This result indicates that the random enzyme caused specific signals due to the spatially close proximity of the free- and esterified GalA unit. On the contrary, the block enzyme seemed to leave no specific marks in the  $^1\text{H}$  NMR spectrum, which confirms the hypothesis that block esterase removes the ester groups until it is stopped by hairy regions or outer influences. In order to confirm these conclusions, a test set is necessary to validate the developed calibration models.

A nearly complete assignment of the  $^1\text{H}$  NMR spectrum of pectin solutions was carried out. Some very interesting correlations were shown between the methoxy signal and the H-2 and H-3 signals which were found to be due to the NOE effects from the spatial structure.

## References

- Andersen, A. K., Larsen, B., & Grasdalén, H. (1995). Sequential structure by H-1-NMR as a direct assay for pectinesterase activity. *Carbohydrate Research*, 273(1), 93–98.
- Braccini, I., Grasso, R. P., & Perez, S. (1999). Conformational and configurational features of acidic polysaccharides and their interactions with calcium ions: a molecular modeling investigation. *Carbohydrate Research*, 317(1–4), 119–130.
- Carpita, N. C., & Gibeaut, D. M. (1993). Structural models of primary cell walls in flowering plants consistency of molecular structure with the physical properties of the walls during growth. *Plant Journal*, 3(1), 1–30.
- Daas, P. J. H., Meyer-Hansen, K., Schols, H. A., De Ruiter, G. A., & Voragen, A. G. J. (1999). Investigation of the non-esterified galacturonic acid distribution in pectin with endopolygalacturonase. *Carbohydrate Research*, 318(1–4), 135–145.
- Daas, P. J. H., Voragen, A. G. J., & Schols, H. A. (2000). Investigation of the galacturonic acid distribution of pectin with enzymes part 2—Characterization of non-esterified galacturonic acid sequences in pectin with endopolygalacturonase. *Carbohydrate Research*, 326(2), 120–129.
- Daas, P. J. H., Voragen, A. G. J., & Schols, H. A. (2001). Study of the methyl ester distribution in pectin with endo-polygalacturonase and high-performance size-exclusion chromatography. *Biopolymers*, 58(2), 195–203.
- Geladi, P., Macdougall, D., & Martens, H. (1985). Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy*, 39(3), 491–500.
- Grasdalén, H., Andersen, A. K., & Larsen, B. (1996). NMR spectroscopy studies of the action pattern of tomato pectinesterase: Generation of block structure in pectin by a multiple-attack mechanism. *Carbohydrate Research*, 289, 105–114.
- Kohn, R., Furda, I., & Kopec, Z. (1968). Distribution of free carboxyl groups in pectin molecule after treatment with pectin esterase. *Collection of Czechoslovak Chemical Communications*, 33(1), 264–269.
- Kohn, R., Markovic, O., & Machova, E. (1983). Deesterification mode of pectin by pectin esterases of *Aspergillus-Foetidus*, Tomatoes and Alfalfa. *Collection of Czechoslovak Chemical Communications*, 48(3), 790–797.
- Korner, R., Limberg, G., Christensen, T. M. I. E., Mikkelsen, J. D., & Roepstorff, P. (1999). Sequencing of partially methyl-esterified oligogalacturonates by tandem mass spectrometry and its use to determine pectinase specificities. *Analytical Chemistry*, 71(7), 1421–1427.
- Limberg, G., Korner, R., Buchholt, H. C., Christensen, T. M. I. E., Roepstorff, P., & Mikkelsen, J. D. (2000a). Analysis of pectin structure part 3—Quantification of the amount of galacturonic acid residues in block sequences in pectin homogalacturonan by enzymatic fingerprinting with exo- and endo-polygalacturonase II from *Aspergillus niger*. *Carbohydrate Research*, 327(3), 321–332.
- Limberg, G., Korner, R., Buchholt, H. C., Christensen, T. M. I. E., Roepstorff, P., & Mikkelsen, J. D. (2000b). Analysis of pectin structure part 1—Analysis of different de-esterification mechanisms for pectin by enzymatic fingerprinting using endopectin lyase and endopolygalacturonase II from *A-niger*. *Carbohydrate Research*, 327(3), 293–307.
- Ló, V.-M., Hahn, M. G., & Halbeck, H. (1994). Preparation, purification, and structural characterization of linear oligogalacturonides. An FAB-mass spectrometric and NMR spectroscopic study. *Carbohydrate Research*, 255, 271–284.
- Martens, H., Nielsen, J. P., & Engelsen, S. B. (2003). Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. *Analytical Chemistry*, 75(3), 394–404.
- Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., & Engelsen, S. B. (2000). Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy*, 54(3), 413–419.
- Ralet, M. C., Bonnin, E., & Thibault, J. F. (2002). Pectins. In E. J. Vandamme, S. De Baets, & A. Steinbüchel (Eds.), *Polysaccharides II*. Weinheim: Wiley-VCH Verlag GmbH.
- Ralet, M. C., Crepeau, M. J., Buchholt, H. C., & Thibault, J. F. (2003). Polyelectrolyte behaviour and calcium binding properties of sugar beet pectins differing in their degrees of methylation and acetylation. *Biochemical Engineering Journal*, 16(2), 191–201.
- Ralet, M. C., Dronnet, V., Buchholt, H. C., & Thibault, J. F. (2001). Enzymatically and chemically de-esterified lime pectins: Characterisation, polyelectrolyte behaviour and calcium binding properties. *Carbohydrate Research*, 336(2), 117–125.
- Ralet, M. C., & Thibault, J. F. (2002). Interchain heterogeneity of enzymatically deesterified lime pectins. *Biomacromolecules*, 3(5), 917–925.
- Ridley, B. L., O'Neill, M. A., & Mohnen, D. A. (2001). Pectins: structure, biosynthesis, and oligogalacturonide-related signaling. *Phytochemistry*, 57(6), 929–967.
- Rolin, C. (1993). Pectin. In R. L. Whistler, & J. N. BeMiller (Eds.), *Industrial Gums: Polysaccharides and their derivatives* (3rd ed., pp. 257–293). San Diego: Academic Press.
- Rosenbohm, C., Lundt, I., Christensen, T. M. I. E., & Young, N. W. G. (2003). Chemically methylated and reduced pectins: Preparation, characterisation by H-1 NMR spectroscopy, enzymatic degradation, and gelling properties. *Carbohydrate Research*, 338(7), 637–649.
- Tomasi, G., van den Berg, F., & Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics*, 18(5), 231–241.
- van Alebeek, G. J. W. M., van Scherpenzeel, K., Beldman, G., Schols, H. A., & Voragen, A. G. J. (2003). Partially esterified oligogalacturonides are the preferred substrates for pectin methylesterase of *Aspergillus niger*. *Biochemical Journal*, 372, 211–218.
- Willats, W. G. T., McCartney, L., Mackie, W., & Knox, J. P. (2001). Pectin: cell biology and prospects for functional analysis. *Plant Molecular Biology*, 47(1–2), 9–27.
- Willats, W. G. T., Orfila, C., Limberg, G., Buchholt, H. C., van Alebeek, G. J. G. M., Voragen, A. G. J., Marcus, S. E., Christensen, T. M. I. E., Mikkelsen, J. D., Murray, B. S., & Knox, J. P. (2001). Modulation of

- the degree and pattern of methyl-esterification of pectic homogalacturonan in plant cell walls—Implications for pectin methyl esterase action, matrix properties, and cell adhesion. *Journal of Biological Chemistry*, 276(22), 19404–19413.
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), 37–52.
- Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration-problem in chemistry solved by the PLS method. *Lecture Notes in Mathematics*, 973, 286–293.
- Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20, 397–405.