Spectroscopic detection of macromolecular interactions focusing on protein-protein interactions in food

Bruun, Susanne Wrang; Jacobsen, Susanne; Søndergaard, Ib

Protein- lipid- carbohydrate interactions and importance to food quality

Spectroscopic detection of macromolecular interactions

Susanne Wrang Bruun

Ph.D. Thesis

February 2006

BioCentrum-DTU
Biochemistry and Nutrition Group (BNG)
Technical University of Denmark
Søltofts Plads, building 224
2800 Kgs. Lyngby
Denmark

List of Publications included in this thesis

Paper I. Gottlieb, D.M., Schultz, J., Bruun, S.W., Jacobsen, S., Søndergaard, I. (2004). Multivariate approaches in plant science. Phytochemistry. 65, 1531-1548.

Paper II Bruun, S.W., Kohler, A., Adt, I., Sockalingum, G.D., Manfait, M., Martens, H. Correcting ATR-FTIR spectra for water vapour and carbon dioxide. *To be submitted to Applied Spectroscopy*.

Paper III Martens, H., Bruun, S.W., Kohler, A. Correction for temperature- and salt- effects in water in FTIR biospectroscopy by EMSC. S*ubmitted to Journal of Chemometrics*.

Paper IV. Bruun, S.W., Holm, J., Hansen, S.I., Jacobsen, S. Application of NIR and FTIR in characterisation of ligand-induced conformation changes in folate binding protein purified from bovine milk. Influence of buffer type and pH. S*ubmitted to Applied Spectroscopy*.

Abstract

It is essential to improve the knowledge of macromolecular interactions in foods, as the food quality is highly influenced by these interactions. Thus, new methods for detection and characterisation of macromolecule interactions are seeked. In this thesis, near-infrared spectroscopy (NIR) is examined as a tool for this type of analysis, as NIR has the advantages of being non-destructive, fast, flexible and applicable on a wide range of sample types. The attention in this work is to protein interactions and conformations.

A part of the work dealt with measurements of aqueous protein solutions, whereas most of the previous studies have used proteins in the dry state to show the sensitivity of NIR to protein secondary structures. The preliminary experiments showed NIR to be sensitive to β-sheet and α-helix contents also for measurements of the dilute protein solutions (10 mg/ml). The structure sensitive reference method mid-infrared spectroscopy (MIR) was applied to confirm this. However, further studies, involving more samples, are necessary in order to survey the possibility to obtain quantitative structure information from NIR spectra. The experiments gave indication that changes in amino acid side chain interactions and their micro-environments influence the spectra and thus indicated that conformational changes other than secondary structure changes may be detected by NIR as well. This was seen in studies of a protein, which adopted monomer or polymer forms depending on the buffer type (Paper IV).

Studies of protein conformations and interactions in the gluten-water system were done with the purpose of evaluating the ability of NIR to give information of more complex systems, for which several different constituents as well as light scattering influence the spectra (Paper I). The gluten system is useful for demonstrating the capacity of NIR in structure-functionality studies, as there is a well-known relation between protein structures and the functionality of gluten. The spectral changes were partly interpretable by means of the MIR reference measurements, but a full explanation of the NIR variations will require other types of measurements for monitoring the protein hydration, the hydrophobic interactions etc. Even though NIR was shown insensitive to the intermolecular β-sheet, which is of importance in the network structure of hydrated gluten, the experiments indicated that NIR could give information on the gluten functionality due to its sensitivity towards amino acid side chain hydrations and interactions. Additional experiments involving gluten and other model systems would be necessary in order to confirm this hypothesis and to show how general these results are.

Water is a major ingredient in many foods and has a great influence on the obtained spectra. Most times, the variations in the water spectrum are irrelevant and their dominance in the NIR spectra of food proteins can impair the analysis. An empirical model was in this work shown capable of removing these variations from MIR spectra (Paper III), and the same method is considered suitable for correction of NIR-spectra. A similar method has been employed for removal of the water vapour and CO_2 absorptions from MIR spectra, as a way to improve the spectroscopic analysis (Paper II).

Resumé

Protein-lipid-kulhydrat interaktioner og betydning heraf for levnedsmidlers kvalitet.

Det er ønskeligt at øge kendskabet til de interaktioner, der finder sted mellem makromolekylerne i fødevarer, idet disse har stor betydning for fødevarernes kvalitet. Der søges derfor nye metoder til at detektere og karakterisere makromolekyle-interaktioner i en kompleks prøve. I denne afhandling undersøges om nær-infrarød spektroskopi (NIR) er anvendeligt til dette formål. NIR er en ikke-destruktiv, hurtig og fleksibel metode, der kan anvendes til måling på mange typer prøver. Der fokuseres i rapporten på studiet af protein-interaktioner og -konformationer.

En del af arbejdet har omhandlet måling af proteiner i vandig opløsning, mens de fleste tidligere studier, som har påvist NIRs følsomhed overfor proteiners sekundære struktur, har omhandlet proteiner i tør tilstand. De indledende forsøg viste, at NIR-spektrene afhænger af de relative indhold af β-sheet og α-helix, også ved måling på relativt tynde proteinopløsninger (10 mg/ml). Dette blev bekræftet ved hjælp af den struktur-følsomme reference metode midt-infrarød spektroskopi (MIR). Anvendeligheden af NIR til kvantitativ struktur-analyse skal dog undersøges yderligere med brug af et større datamateriale. Forsøgene indikerede, at NIR ikke kun er sensitiv overfor den sekundære struktur men også kan detektere konformationsændringer, der involverer ændringer i aminosyresidekædernes interaktioner og deres 'mikro-miljø'. Dette blev set i studiet af et protein, som antog monomer hhv. polymer form i forskellige buffertyper (Artikel IV).

Undersøgelser af protein-konformationer og -interaktioner i gluten-vand systemet blev benyttet til at evaluere NIRs anvendelighed i mere sammensatte systemer, hvor forskellige komponenter influerer på spektrene, og hvor lysspredning giver yderligere kompleksitet (Artikel I). For gluten-systemet er der en kendt relation mellem proteinernes struktur og glutenets funktionalitet. Forsøgene demonstrerer derfor også NIRs anvendelighed i struktur-funktions-studier. Der blev påvist kvalitative ændringer i gluten proteinernes NIR-spektrum ved modificering af proteinernes strukturer og interaktioner både via hydrering og ved påvirkning med forskellige salte. Tolkning af de spektrale ændringer var delvis mulig ud fra MIR reference-målinger, men for en fuld forklaring af NIR-spektrenes variationer kræves andre typer målinger til at følge hydreringen, hydrofobe interaktioner mv. Forsøgene indikerede, at selvom NIR ikke kan detektere intermolekylær β-sheet, som indgår i netværksstrukturen af den hydrerede gluten matrix, kan NIR alligevel give information om glutenets funktionalitet pga. en følsomhed overfor aminosyresidekædernes hydrering og interaktioner. Yderligere forsøg med gluten samt andre model systemer er nødvendige for at bekræfte ovenstående samt generaliserbarheden af resultaterne.

Vand indgår som en vigtig ingrediens i mange fødevarer og har stor indflydelse på de målte spektre. Variationer i vandspektret er dog ofte irrelevante og kan dominere i NIR-analyser af fødevareproteiner og hæmme analysen af disse. En empirisk model til at fjerne disse variationer er blevet afprøvet med gode resultater for MIR-spektre, og samme metode menes anvendelig til korrigering af NIR-spektre (Artikel III).

Ligeledes er en lignende metode benyttet til fjernelse af vanddamp og CO_2 absorptioner fra MIR spektre med henblik på at fremme den spektroskopiske analyse (Artikel II).

Preface and Acknowledgement

This thesis is based on the work of my Ph.D. study, carried out from September 2002 until February 2006 in the Biochemistry and Nutrition Group (BNG) at BioCentrum-DTU, Technical University of Denmark, with supervision from Susanne Jacobsen (supervisor) and Ib Søndergaard (co-supervisor), both employed at BioCentrum-DTU. The three years work has been funded by a Ph.D.-grant from Technical University of Denmark.

The study has included two-month stay at Matforsk (Norwegian Food Research Institute), Ås, Norway.

The work aims to contribute to the improvement of biomacromolecule analysis by means of near-infrared spectroscopy and to explore new applications of this method -in particular in the field of food. The thesis is primarily concerned about the detection of intermolecular interactions, as these have considerable influence on the food quality. The speed, easiness and the non-destructive/non-invasive aspects of near-infrared spectroscopy make it suitable as an *on line* technique for quality control in the food production.

I am grateful to the following:

- My supervisors, Susanne Jacobsen and Ib Søndergaard, for helpful discussions and good advices, and for creating a positive and enthusiastic working atmosphere.

- My colleagues in the Protein Group at BNG. A particular thank to Xuxin Lai and Yiwu Zheng. It was a great pleasure to share office and work in the NIR field with you, and it wouldn't have been the same without you. Thanks to Ljiljana Nesic and Marianne K. Petersen for excellent company.

- People at Matforsk (Norwegian Food Research Institute) for their help and for making me feel at home during my two-month stay in Ås. A special thank to Harald Martens for his supervision at Matforsk, his good, inspiring and innovative advices and for presenting me to his chemometric and preprocessing methods. Also a particular thank to Achim Kohler for introducing me to infrared spectroscopy and new analysis methods.

- Jan Holm for giving me insight into the research of folate binding proteins and for supply of the folate binding protein. Thanks for very good cooperation.

- ALK-Abelló (Hørsholm, Denmark) for funding two month of the work.

- Lisbeth T. Hansen (Royal Veterinary & Agricultural University, Denmark), Elisabeth Fjærvoll Olsen, (Norwegian University of Life Sciences), for introducing me to the different spectroscopic instruments.

- All others who helped and gave me good advices. Marco van der Weert and Eva Horn Møller (The Danish University of Pharmaceutical Sciences) for their help in the infrared spectroscopic fields. In addition: Bradford W. Seabourn, Nikolaus Wellner and Abe Hideyuki.

- All people at BNG for making this a nice place to work.

Susanne Wrang Bruun

Kgs. Lyngby, February 2006

Abbreviations

ATR	Attenuated total reflection
BLG	β-lactoglobulin
BSA	Bovine serum albumin
CAS	Casein
CCA	Canonical correlation analysis
CV	Cross validation
CD	Circular dichroism
2DCOS	Two-dimensional correlation spectroscopy
EMSC	Extended multiplicative scatter correction
FBP	Folate binding protein
FIR	Far-infrared
FTIR	Fourier transform infrared spectroscopy
GMP	Glutenin macropolymer
HMW-GS	High molecular weight glutenin subunit
IR	Infrared (spectroscopy)
LMW	Low molecular weight glutenin subunit
Lys	Lysozyme
MIR	Mid-infrared (spectroscopy)
MSC	Multiplicative scatter correction
MT	Mixing time
MVA	Multivariate data analysis
NIR	Near-infrared (spectroscopy)
NMR	Nuclear magnetic resonance
OPA	Outer product analysis
OVA	Ovalbumin
PC	Principal component

PCA	Principal component analysis
PLSR	Partial least squares regression
RMSE	Root mean square error
SDS	Sodium dodecyl sulphate
SNR	Signal to noise ratio
TDC	Transition dipole coupling
TDM	Transition dipole moment
WEP	Water extractable pentosans
WUS	Water unextractable solids

Table of contents

APPENDIX LIST

APPENDIX I	Paper I: Multivariate approaches in plant science.
APPENDIX II	Paper II: Correcting ATR-FTIR spectra for water vapour and carbon dioxide.
APPENDIX III	Paper IV: Application of NIR and FTIR in characterisation of ligand-induced conformation changes in folate binding protein purified from bovine milk.
APPENDIX IV-1	NIR assignments to lipid and starch.
APPENDIX IV-2	Preprocessing method for NIR in Experiment IV.
APPENDIX V	Additional figures to Experiment V.

Chapter 1: Introduction and background

Quality control has become a highly important topic in the food industry. The assessment of product safety, nutritional value, eating quality, etc. requires the monitoring of quality throughout the production line. Therefore, the *on line* and *in line* spectroscopic techniques combined with appropriate data analysis methods may become increasingly important tools in food production and research [Bro, 2002].

1.1. Macromolecular interactions and importance to food quality

Food is basically composed of molecular and colloid dispersions of biopolymers and their complexes. The major biopolymers in foodstuff: protein and starch, are fundamental to the structure, rheology, and other physical properties of foods, as well as their taste and sensory perception. Water and lipids, binding to other components or acting as solvents, are important factors as well. As the complexed molecules obtain new properties compared to the individual, the functional properties of foods reflect the physico-chemical properties of both the complexed and the individual macromolecules. In this way, interactions between macromolecules contribute to the diversity of food structures [Tolstoguzov, 1996, 2003].

The interactions in foods are mostly of unspecific nature and are developed during the food processing (pumping, centrifugation, heating, extrusion etc.), as some unfolding of the macromolecules is a prerequisite for the interactions. The chemical and enzymatic reactions lead to altered intra- and intermolecular interactions that, at the macromolecular level, are reflected in conformation changes, complexation and aggregation. The interactions may end up in formation of colloids and three-dimensional networks, which are central to the development of macroscopic structures such as suspensions, foams, emulsions and gels etc. [Tolstoguzov, 1996].

Complexation may take place within a type of biopolymer (e.g. protein-protein) or between different types of biopolymers. The protein-polysaccharide complexation is essential in the formation of the complex gels such as alginate-gelatine, alginate-casein gels etc., and the protein-polysaccharide complexes are often applied as emulsion stabilisers due to their high surface activity and ability to form thick gel-like layers. For example, they are added in ice-cream mixes [Tolstoguzov, 1996].

Macromolecule complexation may also lead to phase separation caused by charge-neutralisation and the formation of insoluble compounds. These are easily formed by the unordered proteins (e.g. gelatine, casein and denatured grain storage proteins), which are not topologically restricted, and these proteins have special functionalities in foods [Tolstoguzov, 2003]. At high macromolecule concentration, also the incompatibility between proteins and polysaccharides can lead to phase separation and depletion flocculation. These phenomena influence the gel-strengths and the stability of oil-water emulsions and for example contribute to the structures of doughs and ice-cream mixes [Tolstoguzov, 1996, 2003].

As introduced, the macromolecule complexation as well as macromolecule compatibility and cosolubility are factors that govern the outcome of mixing and processing the various biopolymers together. Thus, the monitoring of complex-formation during food processing could contribute to an increased control of the processes that determine food quality.

1.2. Scope of the project

This project has the goal of evaluating the performance of near-infrared spectroscopy (NIR) as a method for detection and characterisation of biomacromolecule structures and interactions. The sensitivity of NIR to complex-formation between macromolecules is investigated with the purpose of bringing a new dimension to the control and improvement of food qualities. The recognition of NIR as a tool for gaining information on macromolecule interactions could bring about new applications of NIR in many fields. In the food industry, monitoring of protein-protein interactions during cheese ripening and dough mixing are possible applications. In this thesis, the focus is on protein structure and protein-protein interactions, so as to limit the study. Therefore, the work only constitutes a small part of this rather new research area.

The central techniques in this work: NIR and mid-infrared spectroscopy (MIR) and methods for their analysis are introduced in this chapter with the purpose of giving a basic understanding of their capabilities. MIR is applied to facilitate the interpretation of the NIR spectra. First some basic theory considering the origin of absorption bands in the NIR and MIR regions is given, where after mechanisms that underlie the sensitivity of the absorptions to intermolecular interactions are described.

1.3. Introduction to infrared spectroscopy

The possibility of rapid, accurate and non-destructive identification and quantification of biochemical components in many types of samples by NIR has made the technique an important method in the fields of biotechnology, food, human health etc. NIR has since 1970s been widely applied for quality control and compositional analysis in the agro-food industry e.g. for protein quantification in wheat grains and is now widely used for screening in wheat breeding programs [Halverson, 1988; Osborne, 2000]. The NIR spectrum has traditionally been considered containing much less structural information than the MIR spectrum, due to the weakness of the absorption bands and their low resolution. However, NIR provides also physical sample information and has found more widespread utility than MIR or Raman spectroscopy in certain analyses (e.g. of cereal compositions) [Barton, 1996]. In recent years NIR has also become accepted as a method for quality control and process monitoring in the pharmaceutical industry and is considered a promising tool in the medical area [Reich, 2005, Murayama 1998, Blanco, 2002; Heise, 2002a]. Most of the applications of NIR rely on the calibration against a reference method and use NIR as an empirical method.

The weak NIR absorptions are related to the more intense MIR absorptions. The MIR spectra show narrower and more isolated bands than the NIR spectra, and while NIR has been used predominantly for gaining compositional information, MIR has often been applied in structural analyses of biomacromolecules [Jackson, 1995]. The sensitivity of MIR towards aggregation and misfolding of proteins causes this technique to have potential for discriminating the healthy and diseased tissue occurring in Alzheimer's- and Kreutzfeldt Jakob's diseases and diabetes etc. [Pizzi 1995]. MIR has also shown capacity to measure starch crystallinity in the bread staling process, lipid phases and protein gel-formation [Sevenou, 2002; Allain, 1999; Lewis, 2000].

Some reviews of the theory, history, applications and practical aspects of NIR are given in: Osborne (2000), Pasquini (2003), and Blanco et al (2002) as well as in Paper I included in this thesis (Appendix I-1). The use and advances of MIR in biospectroscopy and human health is outlined in: Shaw (1999).

The vibrational transitions in NIR and MIR

Some basic theory is given in this section. For a more thorough study is referred to: Schrader (1995), Bokobza (2002) and Ciurczak (2001).

Electromagnetic radiation has the ability to interact with matter and exchange energy, thereby giving rise to an absorption spectrum. In the vibrational spectroscopies, to which belong NIR and MIR, the interactions result in the transfer of radiation energy to the mechanical energy associated predominantly with the vibration of atoms in chemical bonds (stretching and bending vibrations). Thus, NIR and MIR are concerned with the infrared (IR) part of the electromagnetic spectrum, as the vibrational transition energies are concurrent with the energy in this frequency/wavelength region. The NIR region is found closest to the region of visible light and includes the wavelength range from ~780 nm to 2500 nm (wavenumbers: 1280-4000 cm^{-1}). The MIR region spans the higher wavelength range from 2500 to 15,000 nm (wavenumbers: 4000-660 cm^{-1}). The far infrared (FIR) is the range from 15,000 nm to 100,000 nm (wavenumbers: 660-10 cm^{-1}). The FIR is associated with the rotations and translations of atoms in the gaseous state and will not be described here.

The vibrational energy transitions are explained from the harmonic and anharmonic oscillator models, which both consider the simple case of a diatomic molecule. Only the anharmonic oscillator model is valid for actual molecules.

Harmonic oscillator: In the harmonic oscillator approximation, the diatomic molecule is depicted as two masses (m and M) connected by a weak spring. Only a stretching vibration is taking place in this system. The mechanical model leads to a vibrational frequency ν that depends on the stiffness of the spring (*f*) and the reduced mass μ (=m+M):

(Eq. 1.1) $$\nu = \frac{1}{2\pi}\sqrt{\frac{f}{\mu}}$$

The vibrational energy of the harmonic oscillator increases if some photon energy is transferred to the molecule to increase the vibrational amplitude (excitation). However, the molecular oscillator, vibrating at frequency ν_m, can only obtain some discreet energy levels, which from quantum mechanics are found as:

(Eq. 1.2) $$E_v = h\nu_m\left(V + \frac{1}{2}\right) \text{ V =1, 2, 3..}$$

In Eq. 1.2, E_v is the energy associated with the quantum level V. A transition from one quantum level to the next level requires the absorption of the energy: $\Delta E = h\nu_m$, and this energy quantum can be supplied by electromagnetic radiation of the same frequency. Most observed transitions are from the ground level E_0, as most molecules at ambient temperatures exist in this state. The fundamental 0→1 transitions require energies that match the energy of photons in the MIR region, whereas the higher overtone transitions (0→2, 0→3 etc.) require the energies of photons in the NIR-region. In the harmonic oscillator, the overtones are not allowed according to the selection rules, which apply in the quantum mechanical model.

Anharmonic oscillator: Anharmonicity originates from the vibrations about the equilibrium position being non-symmetric in real molecules. Thus, the potential energy function is approximated by the Morse function (see Fig. 1.1). The deviation from harmonicity is most significant for the chemical bonds of a high vibrational amplitude, such as bonds that connect a large atom with the small hydrogen atom.

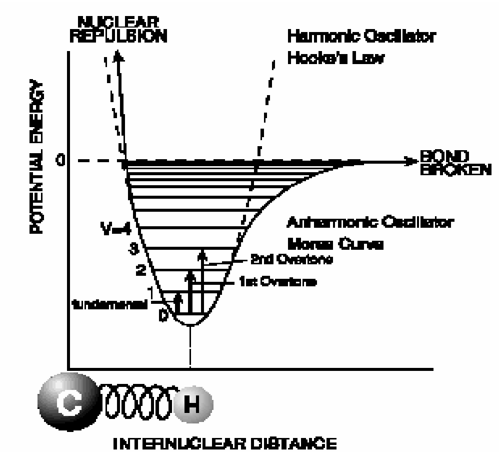


Fig. 1.1. Energy levels, according to the anharmonic oscillator model.
The Morse function (full line) shows the potential energy for the anharmonic oscillator.
Also the potential energy according to the harmonic oscillator model is shown (broken line).
Illustration from [Murray, 2004].

The anharmonicity provides the way for overtones and combination bands, which appear in NIR spectra. (see Fig. 1.2). Anharmonicity affects both the electrical and mechanical properties of the bond, and the overtone and combination bands become active as a consequence of either the electrical or the mechanical anharmonicity. A consequence of the anharmonicity is that allowed vibrational energy levels are not equally spaced but converge with increasing quantum number V as expressed in Eq. 1.3.

(Eq. 1.3) $$E_v = h\nu_m(V + \frac{1}{2}) - h\nu_m x(V + \frac{1}{2})^2$$

In Eq. 1.3*, x* accounts for the mechanical anharmonicity. The energy required for the transition 0→V is given as:

(Eq. 1.4) $$\Delta E = h\nu_m V(1-(V+1)x) \text{, V=1,2,3....}$$

From Eq. 1.4, the energy required for the fundamental transition (V=1) is found as $\Delta E=h\nu_m(1-2x)$, and the energy required for the first overtone transition (V=2) is $\Delta E=h\nu_m(2-6x)$.

Due to anharmonicity, NIR spectra of polyatomic molecules also show combination bands at frequencies that are additions of multiples of the fundamental frequencies. These bands arise when the energy of an absorbed photon is shared between two or more fundamental transitions. The requirement is that at least one of the vibrations is infrared active (described below).

Even though a high number of overtone and combination bands generally are present in NIR spectra, these often have a smooth appearance with a few broad bands, due to band overlapping and averaging out. The infrared spectrum of a protein-rich sample is shown in Fig. 1.2. The chemical information provided by the fundamentals in the MIR region (Fig. 1.2B) is repeated in the combinations and overtones in the NIR region (Fig.1.2A).

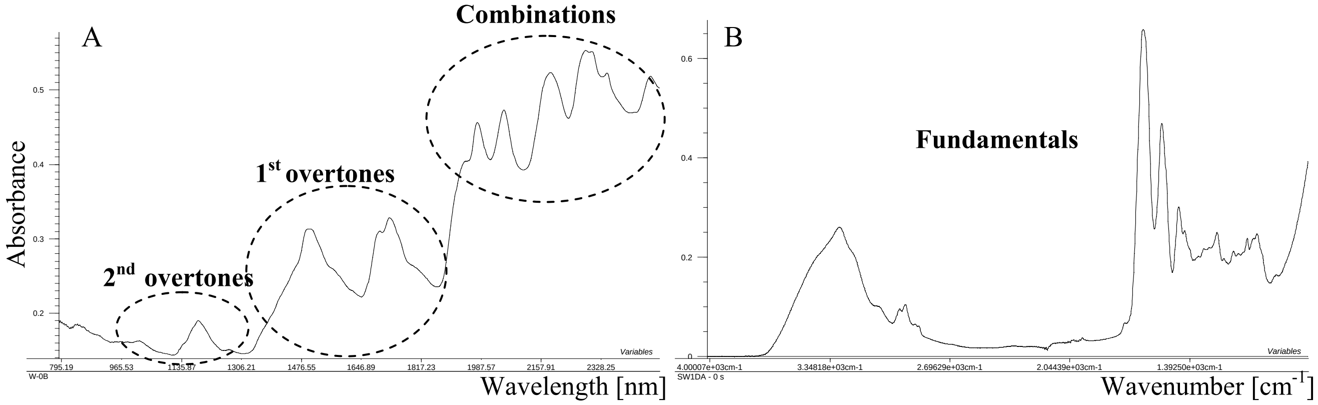


Fig.1.2. The infrared spectrum of gluten. A) NIR region, containing the combination and overtone bands. B) MIR region, containing the fundamental bands. For the NIR spectrum it is common to use the wavelengths (λ), whereas for the MIR spectrum it is common to use the wavenumbers (υ) . The relations to frequency (ν) are: λ = c/ ν and υ= ν /c (c is the speed of light).

Since the probabilities of overtone- and combination transitions are low, the absorptivities in the NIR are 10-100 times weaker than in MIR, and they also get progressively weaker at increasing overtone level (see Fig. 1.2A). The weak absorptions in NIR compared to MIR means that the light, reaching the detector, contains information of a larger amount of sample. However, NIR is still a surface method. For example, the NIR-light only penetrates a few mm into a dough [Wesley, 1998].

The different chemical groups also cause bands of very different intensities. For the exchange of energy to take place, it is a prerequisite that the vibration is associated with a change in dipole moment i.e. a transition dipole moment (TDM). Therefore, vibrations that maintain a centre of symmetry in the molecules are infrared inactive, and, as the absorptivity shows proportionality to the TDM, the chemical groups of high bond polarity are the most absorptive in the MIR [Swanton, 1986]. For example, the C=O stretching gives rise to an intense band in the MIR. However, this does not apply to the NIR spectrum, for which the anharmonicity of the chemical bond has influence on the absorptivity as well (as the NIR absorptions are based on anharmonicity). The consequence is that overtones and combinations, arising from bonds with a high electrical anharmonicity, exhibit the most intense absorptions in NIR [Bokobza, 2002; Murray, 1987]. Therefore, the absorptions from the CH, NH and OH groups dominate, and the C=O groups give rise to very weak absorptions in the NIR region.

Quantitative chemical information

The sample spectrum is obtained from the intensity of radiation entering (I_0) and leaving (I) the sample. I is either the radiation that is transmitted through the sample or the radiation that is reflected back from the sample, giving rise to transmittance ($T=I_T/I_0$) and reflectance ($R=I_R/I_0$) spectra, respectively. (As I_0 is not easily obtained, I is ratioed to a background measurement e.g. obtained from the empty cuvette).

In the ideal transmission situation, the light passing through a sample of thickness l is absorbed according to the absorption coefficient (μ_a) in the sample at that specific wavelength. The attenuation of light in the sample can then be expressed according to Eq. 1.5.

(Eq. 1.5) $\ln(I/I_0)= -\mu_a l$

The absorption coefficient μ_a depends on the analyte concentration (c) [mol/L] and the molar absorptivity (ε) [$L*mol^{-1}*cm^{-1}$], and these parameters are applied in Beer's law, which states proportionality between absorbance (A) and analyte concentration (Eq. 1.6). In Eq. 1.6, the optical path length l has the unit of cm and absorbance is expressed in absorbance units [AU].

(Eq. 1.6) $A = -\log_{10}(I / I_0) = \varepsilon cl$

Beer's law can seldom be used in a univariate manner. The overlapping of bands from various chemical groups and baseline variations instead ask for a multivariate analysis (MVA) method for unravelling the chemical information. These methods are based on Beer's law and the additivity statement in Eq. 1.7.

(Eq. 1.7) $A=\varepsilon_1 c_1 l+\varepsilon_2 c_2 l….$

Due to detector non-linearities Beer's law may not be valid for high absorbencies (>1 AU). For transmission measurements, deviations from Beer's law may also result from instrument-drift, stray light and interactions between components, of which the latter may cause shifts in the absorbance peaks [Murray, 1987].

1.4. Physical information

NIR spectra are rich in physical information about the samples, due to the light scattering effects.

Light scattering effects

NIR has the potential to measure particulate samples (turbid liquids, semi-solids or solids) by use of diffuse transmission, diffuse reflection or transflection. For example, flours have been analysed widely by NIR diffuse reflection [Law, 1977; Delwiche, 1994, 1998; Sato, 2001; Wesley, 2001] and intact grains by use of diffuse transmission [Munck, 2001]. Diffuse reflectance is the light that is reflected back to the detector from a particulate sample after having experienced reflections, refractions and diffractions at the particle surfaces [Birth, 1987]. Beer-Lamberts law is originally not defined for these measurements, since the path length is unknown. Nevertheless, the proportionality between absorbance and concentration may also hold in this situation, i.e. $A=\log(1/R) \propto c$ [Birth, 1987].

The light scattering property of a sample has influence on the effective path length of the light. The higher the scattering property of the sample (reflected in the scatter coefficient μ_s), the shorter is the effective path length and the lower is the apparent absorbance level [Pasikatan, 2001]. Consequently, the spectra show multiplicative scaling effects, i.e. multiplication of every wavelength by a factor (see Fig. 1.3). Also additive offset variations may arise due to constant loss of light at all wavelengths.

The light scattering property is dependent on the sample composition, as the refractions and reflections in the sample depend on the refractive index differences between the particles and the medium. The smaller this difference is, the lower is the μ_s [Doyle, 2001]. For example, addition of water to a dry sample decreases the refractive index differences and the thereby decreased scattering properties could lead to a lack of diffuse reflectance from the sample [Doyle, 2001]. The light scattering property is also highly dependent on the particle concentration, particle size, particle size distribution, particle shape and surface texture as well as the crystalline form and packing density etc. [Murray, 1987]. An increased particle size diminishes the light scattering interfaces, thus reducing the μ_s and leading to a larger effective path length [Pasikatan, 2001]. However, the increased particle size may also decrease the absorption coefficient for strong absorbers, since the light does not penetrate so deeply into strongly absorbing particles. Instead the light is specularly reflected i.e. reflected from the surface and is without information on the sample composition. This ‘hidden mass’ effect means that the absorption from a large particle can be smaller than the absorption from several small particles of the same mass [Burger, 1997]. Due to the described complexity, the different particle sizes may lead to different spectral shapes [Ventura, 1999].

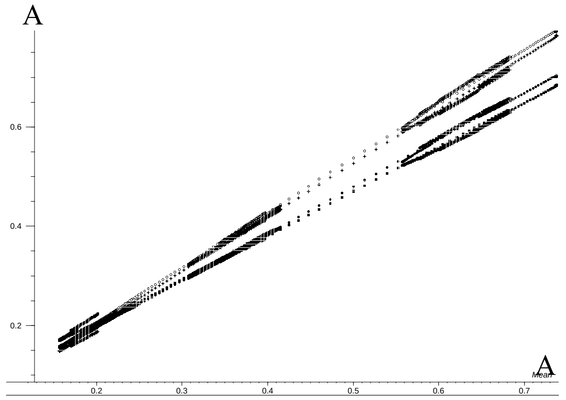


Fig. 1.3. Plot of four diffuse reflectance spectra against their mean. The different slopes reflect the multiplicative effects. Some chemical variation at high absorbance is seen.

Wavelength dependent scatter results from a dependency of μ_s on the wavelength (μ_s usually decreases with wavelength), as this dependency is affected by physical properties such as the particle size [Schmitt, 1996]. This results for example in varying baseline slopes.

Separation of physical and chemical variations

Although the light scattering can be handled implicitly in the multivariate calibration models, the presence of the physical effects in the spectra demand more calibration samples and result in models of higher complexity. Therefore, the alternative approach to use a preprocessing method to separate out the light scattering from the chemical variations will usually improve the data analysis.

Multiplicative Signal Correction (MSC) and Extended MSC (EMSC) are methods with the purpose of modelling and subtracting the unwanted spectral variations that result from light scattering, variable path lengths and baseline drifts [Martens, 2003]. An advantage of these methods is the storage of the scatter information in the output parameters, which can then give information on the physical sample properties. In the MSC/EMSC model the sample spectrum is expressed as the ideal chemical spectrum modified by various physical effects. In the usual MSC approach, an offset (a_i) and a slope (b_i) coefficient are estimated for each spectrum *i* by regression of the spectrum ($\mathbf{z}_i$) to the ideal chemical spectrum $\mathbf{z}_{i,chem}$ according to the linear physical model shown in Eq. 1.8.

(Eq. 1.8) $$\mathbf{z}_i = a_i\mathbf{1}+b_i\mathbf{z}_{i,chem}+\boldsymbol{\varepsilon}_i$$

The MSC coefficients a_i and b_i reflects the additive and multiplicative light scattering effects, respectively, whereas $\boldsymbol{\varepsilon}_i$ accounts for the residuals (measurement errors). In EMSC, the equation is extended with corrections for linear and quadratic wavelength effects i.e. the wavelength-dependent scatter variations. These are represented by the additional terms in Eq. 1.9, where $\boldsymbol{\lambda}$ represents the wavelength range.

(Eq. 1.9) $$\mathbf{z}_\mathrm{i} = a_i\mathbf{1}+ d_i\boldsymbol{\lambda} + e_i\boldsymbol{\lambda}^2+ b_i\mathbf{z}_\mathrm{i,chem} +\boldsymbol{\varepsilon}_i.$$

The regression returns the MSC/EMSC parameters and these are used for correction of the spectrum for the light scattering effects (Eq. 1.10).

(Eq. 1.10) $$\mathbf{z}_{i,corr}=(\mathbf{z}_i-a_i\mathbf{1}-d_i\boldsymbol{\lambda}-e_i\boldsymbol{\lambda}^2)/b_i$$

For determination of the model parameters, the ideal chemical spectrum is taken as a variance $\boldsymbol{\delta}_\mathrm{i}$ around a reference spectrum **m**. In MSC, this variation is unknown and ignored, and the mean spectrum is simply used as the ideal chemical spectrum. A disadvantage of this approach is that the obtained MSC parameters may be influenced by the chemical variations, and MSC may thus remove some of the relevant information [Martens, 2003]. For avoidance of this, the MSC parameters have to be estimated from spectral regions with a minimum of chemical information or the absorption peaks need to be down-weighted.

EMSC/causal modelling: The difficulties described above may also be overcome by applying the spectra of the chemical constituents to express the true chemical spectrum. This spectrum may now be expressed from the J constituent spectra ($\mathbf{k}_1$, $\mathbf{k}_2$,…, $\mathbf{k}_\mathrm{j}$) and the constituent concentration differences (Δc_{i1} , Δc_{i2},…,$\Delta c_{i\,\mathrm{j}}$) relative to the reference sample, according to Eq. 1.11.

(Eq. 1.11) $$\mathbf{z}_{i,\mathrm{chem}} = \mathbf{m} + \Delta c_{i1}\mathbf{k}_1' + \Delta c_{i2}\mathbf{k}_2' + \ldots\ldots \Delta c_{ij}\mathbf{k}_j'$$

Eq. 1.11 is inserted into the physical model (Eq. 1.9), and the EMSC parameters a_i, b_i, d_i and e_i can then be estimated independently of the constituent concentrations by least squares regression. An important feature of EMSC is the possibility to subtract interfering absorption patterns from the sample spectra, as their concentrations are obtained from the estimated EMSC parameters. Whereas the interference spectra are termed “bad spectra”, the chemical constituent spectra are termed “good spectra”, since the latter represent the desired variations. The input spectra may be determined empirically from a calibration set by use of e.g. principal component analysis (PCA, see section 1.6). The application of EMSC with input “good” and “bad” spectra is demonstrated and explained in more detail in Paper III (section 2.5), which considers the removal of temperature- and salt-effects from MIR spectra of aqueous samples. A similar model-based preprocessing method for removal of interfering atmospheric absorptions in MIR spectra is described in Paper II (Appendix II). The EMSC method also remove physical information from the MIR spectra, which are sensitive to the physical state of the samples. When not otherwise mentioned, the term EMSC refers to the standard procedure with the mean spectrum used as the ideal chemical spectrum.

Other preprocessing methods: The choice of preprocessing method for removal of light scattering effects is empirical and the different preprocessing methods and their combinations can be tried out. The derivative transformations are also able to remove baseline offsets and linear baseline variations but cannot get rid of the multiplicative effects. However, some of the residual scatter effects may be removed by use of MSC/EMSC [Pedersen, 2002]. In this work, 2nd derivative transformation is used for obtaining better resolution of the overlapping peaks in both NIR and MIR spectra, as it works as a band-narrowing technique.

1.5. Influence of hydrogen bonding interactions on the infrared spectrum

The dependency on the bond force constant means that any factor that alters the electron density in the bond also alters the vibrational frequency. Likewise, factors that affect the dipole moment associated with the vibration influences the absorptivity. As the hydrogen bonding interaction has pronounced influence on the electron distribution, the MIR and NIR possess the ability to bring information about the strengths of hydrogen bonds [Scheiner, 1997; Czarnik-Matusewicz, 2005], and the methods are recognised as powerful methods for the study of hydrogen bonding interactions [Vanderkooi, 2005; Ozaki, 2002]. The general effect of hydrogen bonding on MIR and NIR spectra is outlined in the following section. The hydrogen bonding interactions in liquid water is dealt with in chapter 2.

The classical hydrogen bond

The hydrogen bond is often described as a special case of a strong dipolar interaction. It links a proton donor molecule (AH) to a proton acceptor molecule (B) and causes the two molecules to share a hydrogen atom. The donor molecule contains a hydrogen atom covalently bound to an electronegative atom (e.g. AH= OH or NH), and the acceptor molecule contains an electronegative atom B with at least one lonepair (e.g. B= O: or

N:) [Scheiner, 1997]. The physical forces that keep the hydrogen bonded molecules together and contribute to the hydrogen bonding energy have been decomposed into several components that include: electrostatic (coulomb), exchange, induction (or deformation) and dispersion (van der Waals) energies. The exchange energy results from the exchange of electrons in molecule A with electrons in molecule B and is a repulsive force in contrast to the other forces, which are acting to contract the hydrogen bond [Scheiner, 1997].

The distance between the hydrogen bonded molecules reflects the strength of the hydrogen bond, and an exponential decrease in hydrogen bonding strength with increasing length of the hydrogen bond is observed. The strength of the hydrogen bond is also influenced by the hydrogen bond angle but to a smaller extent [Scheiner, 1997; Chaplin, 2005]. The linear hydrogen bond has the bridging hydrogen atom positioned on a straight line between A and B, and the more it is bend away from this geometry, the weaker is the hydrogen bond [Scheiner, 1997].

General effect of hydrogen bonding on MIR spectra

Upon hydrogen bond formation, the bridging hydrogen atom is pulled away from the proton donor AH, and the covalent A-H bond is extended and weakened. As a result, the fundamental AH stretching band shifts to lower frequencies, and the size of the vibrational shift ($\Delta\nu_s$) reflects the hydrogen bonding strength [Scheiner, 1997]. The linear relation between $\Delta\nu_s$ and the hydrogen bonding strength for the donor molecule is known as the Badger Bauer rule. It applies for many chemical groups and for hydrogen bonding strengths above a certain value (the shifts are very small for the weakest hydrogen bonds) [Scheiner, 1997]. The size of the shift also depends on the hydrogen bonding angle and is largest for the linear hydrogen bonds. The vibrational frequency of the acceptor molecule is usually not affected as much by the hydrogen bonding as in the case of the donor molecule [Scheiner, 1997]. However, also the binding of the peptide C=O groups to NH-groups causes the C=O stretching band to shift to lower frequencies [Barth, 2002].

The hydrogen bonding strength also has influence on the intensity of the fundamental band for both hydrogen donor AH and acceptor molecule B, due to the perturbations of their dipole moments. When the hydrogen bond is formed, the lonepair on B is dragged towards the AH-molecule, and small electron density redistributions take place. About 0.01-0.03 electrons are transferred from B to A upon hydrogen bonding [Scheiner, 1997]. The electrons from B ‘bypass’ the bridging hydrogen atom, and a consequence of hydrogen bonding is a lowered electron density on this atom [Scheiner, 1997]. The bridging hydrogen atom hereby obtains a higher positive charge, and the dipole moment of the stretching vibration increases for the acceptor groups that possess an electropositive hydrogen atom (e.g. OH and NH). Thus, for these groups, the hydrogen bonding leads to higher intensities of the stretching bands. However, for e.g. the PH group (with an electronegative hydrogen atom), the effect of hydrogen bonding is a decrease in the dipole moment of the stretching vibration and a decrease of the band intensity [Scheiner, 1997].

Also for the acceptor molecule B, the transfer of electron density results in a change of the dipole moment, but the effect on the band intensity is dependent on the chemical group. For example, for the C=O peptide group, the intensity increases upon increasing hydrogen bonding strength [de Jongh, 1996].

The bending vibrational frequency of an XH_2 group reflects the energy necessary for overcoming the repulsion between the hydrogen atoms or between the lonepair electrons on X [Šoptrajanov, 2000]. Hydrogen bonding normally increases the stiffness of the H-X-H angle in the donor molecule. However, this effect may be counteracted by an increase in the X-H bond lengths, which leads to reduced repulsion between the hydrogen atoms [Šoptrajanov, 2000]. Thus, the shifts upon hydrogen bonding are often minor for bending bands [Scheiner, 1997], whereas their absorptivity generally decreases [Vanderkooi, 2005; Scheiner, 1997].

General effect of hydrogen bonding on NIR spectra

The position of an overtone band is affected by hydrogen bonding in the same way as the position of the fundamental band, i.e. a stretching overtone band shifts to lower frequencies/higher wavelengths [Graener, 1991]. Hydrogen bonding may also cause combination bands to shift to higher wavelengths due to the dominating shift of the stretching modes.

The effects of hydrogen bonding on the shape and intensity of overtone and combination bands are different from the case of fundamental bands. The polarisability and thereby the electrical anharmonicity of the bond is affected by hydrogen bonding. Upon hydrogen bonding, the contributions of electrical and mechanical anharmonicity to the overtone transition dipole moment of the stretching vibration can compensate each other, and the consequence is a small decrease in intensity of the stretching overtone band [Graener, 1991]. Thus, the monomers, which are of higher anharmonicity than the polymeric counterpart, produce the most intense overtone bands, and this makes the NIR region suitable for studies of polymer-to-monomer dissociations [Ozaki, 2002; Katsumoto, 2002]. This is in opposition to the case of MIR spectra, in which the free unbound groups have a rather low intensity compared to the hydrogen bonded counterpart, as described above.

Also, the electrical anharmonicity has been shown to affect the shape of overtone bands, and variations in anharmonicity could be partly responsible for the structures (side lopes) that are sometimes observed in overtone bands [Graener, 1991].

Assessing molecular interactions in NIR/MIR spectra

In the spectroscopic study with the purpose of probing macromolecule interactions, the possibility of identifying spectral variations that result from these interactions is essential. The occurrence of chemical interactions is manifested as a difference between the measured sample spectrum and the sample spectrum, reconstructed from the individual constituent spectra. For indicating the presence or absence of macromolecule interactions, the involvement of the pure constituent spectra is thus demanded. EMSC with use of “bad spectra” is a potential approach, as the subtraction of the contributing constituent spectra would leave the variation caused by their interactions in the EMSC residuals (the corrected spectra). Variations in the residuals could then be taken as a confirmation of molecular interactions occurring in the mixture. However, several factors obstruct the implementation of this approach. As the macromolecular interactions often require the presence of water, also changes in the water-macromolecule interactions may be the cause of possible spectral alterations (as hydration affects the spectra of both constituents). Furthermore, the complex light scattering effects in NIR may also contribute to differences between the constituent spectra obtained alone and in the mixture, and non-linearities (i.e. influence of analyte concentration on the spectral shape) add additional difficulties to the above approach. Therefore, it is essential to have some *a priori* information of the spectral changes that can be expected due to interactions. This information is obtained by use of a reference method or by analysis of some chemically well-described systems, such that interpretation of the spectral alterations is possible.

1.6. Multivariate analysis (MVA)/chemometrics

The MVA methods (which are also called chemometrics in case of chemical data analysis) were developed predominantly during the 1970s and provided the way for NIR in this period [Heise, 2002b; Geladi, 2003].

Chemometrics is the application of mathematical and statistical tools to analyse complex chemical data of a multivariate nature i.e. with many variables and/or many types of variation [Martens, 2000a]. A spectral data set usually contains more variables than samples and also has highly correlated variables. This means that the classical multiple linear regression is not suitable for analysing the full NIR spectrum [Murray, 2004]. On the other hand, the bilinear factor methods PCA and partial least squares regression (PLSR) take advantage of the correlations to find a few ‘latent’ non-measurable variables and thereby remove the redundant information [Wold, 2001; Martens, 2000a]. The new latent variables provide an overview of the main variations in the data and facilitate the recognition of otherwise hidden structures. The methods further have the advantages to allow outliers be identified, to allow ‘missing values’ and to separate out the noise from the signal [Wold, 2001]. The stability against noise result from the possibility to use the full spectrum instead of a few selected variables, which is also advantageous for the separation of physical and chemical variations. Thus, the bilinear factor methods have become the most popular methods for NIR analyses [Heise, 2002b]. On the other hand, variable selection may also improve a calibration model.

Here, the basic methods; PCA and PLSR are described, as these methods are being used throughout this work for analysis of a single set or two related sets of variables, respectively. Both are bilinear methods (linear in samples and variables) and generally based on the linear relation between absorbance and concentration, according to Beer’s law. However, PLSR can handle some non-linearity

Principal component analysis (PCA)

PCA is as a non-supervised classification technique. It is used for exploration of patterns and groupings in one data set (**X**) without using any prior knowledge, i.e. the data is allowed to "speak for itself". PCA is described in more details in [Esbensen, 2000a; Martens, 2000b; Geladi, 2003].

To explain the principles of PCA, the objects (described by m variables) are first visualized as points in the m-dimensional space with each axis represented by one variable. PCA reduces this multidimensional space to fewer dimensions by replacing the original variable axes with new principal component (PC) axes, which are linear combinations of the original variables. This is done by singular value decomposition. The orthogonal PCs are determined successively in such a way that PC1 describes as much as possible of the variance in the original data, and each successive PC accounts for as much as possible of the remaining variance. The last PCs only describe the noise in the data. Hence, only the 'A' PCs that make up the structured part and have the noise part separated out are subject for further interpretation. Each of the A PCs is build from a loading- (**p**) and a score vector (**t**) as expressed in Eq. 1.12. Here, **E** holds the residual unmodelled variance (noise) and **X** is often the mean-centred data.

(Eq. 1.12) $$\mathbf{X} = \mathbf{t}_1 \cdot \mathbf{p}_1^{'} + \mathbf{t}_2 \cdot \mathbf{p}' + \ldots + \mathbf{t}_A \cdot \mathbf{p}' + \mathbf{E} = \Sigma \mathbf{PC}_i + \mathbf{E}$$

(Eq. 1.13) $$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}' + \mathbf{E}$$

The score vectors $\mathbf{t}_1$, $\mathbf{t}_2$ …$\mathbf{t}_A$ hold the score values for the objects obtained by projection onto the new PC axes, and the 2D score plots of $\mathbf{t}_1$ vs. $\mathbf{t}_2$, $\mathbf{t}_1$ vs. $\mathbf{t}_3$ etc. provide the possibility to discover patterns, groupings, and outliers in the data, as they show the relations between the objects. The loading vectors: $\mathbf{p}_1$, $\mathbf{p}_2$ …$\mathbf{p}_A$ hold the signature of the PCs, i.e. they express the relations between the original variables and the new PC variables. A high loading value (positive or negative) signifies a variable with a high contribution to the PC, while a loading value close to zero signifies a variable without much contribution. The PCA model is summarised in Eq. 1.13. Here, **T** and **P** are the score- and loading matrices, respectively.

PCA works as a kind of curve resolution technique, which have the purpose to find the pure chemical spectra and their concentrations in a mixture. However, even though the loading vectors can be subject to interpretation to explain the chemical/physical meaning of the PC, the loading vectors from spectroscopic analyses seldom have a pure chemical meaning, due to the constraints of orthogonality [Geladi, 2003]. In addition, a PC does not necessarily result from a single source of variation. However, the PCA results may be used as initial estimates for the pure chemical spectra, which can then be found by adding constraints such as non-negativity, unimodality etc. and optimizing by alternating least squares [Czarnik-Matusewicz, 2005].

A curvefitting approach can also be used for finding concentrations of the individual constituents in the mixture, but in contrast to the curve resolution methods, it applies assumptions of the number and shapes of the individual sub bands and is less flexible. The curvefitting is seldom used for NIR spectra but more frequently for MIR spectra.

Partial least squares regression (PLSR)

In multivariate calibration, the relationship between the independent variables **X** and the dependent variables **Y** is established by use of a calibration set with known Y-values. The obtained model, which is summarised in the regression coefficients **B**, can then be used for prediction of **Y** for a data set of unknown Y-values (see Eq. 1.14). When the spectra are used as the independent data set to predict e.g. chemical concentrations, the spectroscopic method can replace a more labour-intensive chemical analysis.

(Eq. 1.14) $\mathbf{Y}=\mathbf{XB}+\mathbf{F}$

The inspection of B-vectors can be helpful in spectroscopic analyses, where they can draw attention to the spectral regions of high correlation to **Y** and thus help in the band assignment [Šašić, 2000]. The B-vectors may be found by regression of latent variables against the Y-variables, e.g. the score matrix from a PCA can be used. However in PLSR, both **X** and **Y** are modelled by latent variables (see Eq. 1.15), and the Y-variance is used as a guide for decomposition of **X,** as the algorithm seek to increase the covariance of the X- and Y-scores [Martens, 2000f]. This ensures that the first PLS-components (these are also termed PCs) are relavant to **Y**.

(Eq. 1.15) $\mathbf{X}=\mathbf{TP'}+\mathbf{E}$ $\mathbf{Y}=\mathbf{TQ'}+\mathbf{F}$

In addition to the calibration purposes, PLSR may also be applied for determination of the influence of the design factors on some response values (hypothesis-testing method), or for an entirely explorative analysis of how different variables relate to each other’s (hypothesis-generating method). Relationships between several X- and Y-variables can be inspected from the two-dimensional correlation loading plots (See Fig. 1.4) [Martens, 2000d]. These plots show the correlations of each original variable to the two latent variables that are plotted against each other’s. If the main variation is explained in the two PCs, the correlation between variables can be inspected from the plot: The positively correlated variables are placed close together, whereas negatively correlated variables are placed oppositely. Variables close to the origin are poorly explained by the PLS-components and contain no useful information [Martens, 2000b].

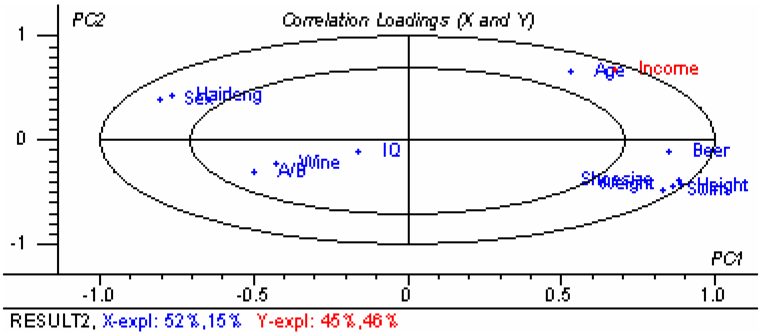


Fig. 1.4. Correlation loading plot from a PLS-regression of several X-variables to one Y-variable (income). The inner circle represents 50 % explanation and the outer circle represents 100 % explanation. The plot shows e.g. that income is highly correlated to age.

PLSR is described in more details in [Wold, 2001; Esbensen, 2002b, Geladi, 2003 and Martens, 2000b, 2000f].

Validation: It is important to validate the PLSR models in order to prevent overfitting and wrong conclusions to be drawn. The external validity of the model is checked by employing a priori knowledge. The internal validation considers the reliability of the results and is usually done by test set validation and/or cross validation (CV). The validation methods assess the stabilities of the model parameters and estimate the prediction errors that can be expected in the future predictions [Martens, 2000e, 2000f]. CV simulates test set validation by consecutively taking out subsets from the calibration set and using these as temporary test sets in several predictions. The stability of the model parameters against the perturbations in CV reflects the reliabilities of the parameters. Thus, reliability ranges for e.g. the B-coefficients can be estimated on the basis of CV results (by means of jack-knifing), and X-variables with a significant influence on **Y** may be identified in this way [Martens, 2000e, 2000f].

The predictive ability of the PLSR model is assessed from the root mean square error of Y (RMSE(Y)), which sum up all Y-residuals and is a measure of both precision and accuracy. The future prediction result can be reported as Y±2*RMSE(Y), which is an empirical interval [Martens, 2000e]. A criterion for a successful calibration may be taken as a specific ratio of RMSE(Y) to the original Y-variance [Oberg, 2004].

1.7. Contents of this thesis

The next chapter deals with the infrared water spectrum, as this has a great influence in the spectroscopic analyses of most biological/food samples (chapter 2). The subsequent parts of the work are concerned with protein structure and interaction analyses using the two infrared spectroscopic techniques, and the experiments cover the simple case of pure protein solutions (chapter 3) and the case of more complex samples, which are of concern in food analyses (chapter 4 and 5). In the latter chapters, the gluten protein network is used as a model system for investigation of the spectral- macromolecule function relationship. Finally, the results are discussed and concluded upon (chapter 6).

1.8. References

Allain, A.F., Paquin, P., Subirade, M. (1999). Relationships between conformation of beta-lactoglobulin in solution and gel states as revealed by attenuated total reflection Fourier transform infrared spectroscopy. *Int. J. Biol. Macromol.* 26, 337-344.

Barth, A., Zscherp, C. (2002). What vibrations tell us about proteins. *Q. Rev. Biophys.* 35, 369-430.

Barton II, F. E., Himmelsbach, D. S., Archibald, D.D. (1996). Two-dimensional vibration spectroscopy. V: Correlation of mid- and near infrared of hard red winter and spring wheats. *J. Near infared Spec.* 4, 139-152.

Birth, G.S., Hecht, H.G. (1987). The Physics of Near-infrared Reflectance. In: *Near-infrared Technology in the Agricultural and Food Industries,* (Wiliams, P. & Norris, K., eds.). AACC, Inc., St. Paul, Minnesota, pp. 1-15.

Blanco, M.,Villarroya, I. (2002). NIR spectroscopy: a rapid-response analytical tool. *Trac-Trend Anal. Chem.* 21, 240-250.

Bokobza, L. (2002). Origin of near-infrared absorption bands. in: *Near-infrared spectroscopy: Principles, instruments, applications,* (Siesler, HW., Ozaki, Y., Kawata, S., Heise, HM., eds.). Wiley-VCH, Weinheim, pp. 11-41.

Bro, R., van den Berg, F., Thybo, A., Andersen, C.M., Jørgensen, B. M., Andersen, H. (2002). Multivariate analysis as a tool in advanced quality monitoring in the food production chain. *Trends Food Sci. Technol.* 13, 235-244.

Burger, T., Kuhn, J., Caps, R., Fricke, J. (1997). Quantitative determination of the scattering and absorption coefficients from diffuse reflectance and transmittance measurements: Application to pharmaceutical powders. *Appl. Spectrosc.* 51, 309-317.

Chaplin, M. (2005). Structure of liquid water. (http://www.sbu.ac.uk/water/).

Ciurczak, E.W. (2001). Principles of near-infrared spectroscopy. In: *Handbook of Near-infrared analysis,* (Burns, DA., Ciruczak E.W., eds.). 2nd edition, Marcel Dekker, Inc., pp. 7-18.

Czarnik-Matusewicz, B., Pilorz, S., Hawranek, J.P. (2005). Temperature-dependent water structural transitions examined by near-IR and mid-IR spectra analyzed by multivariate curve resolution and two-dimensional correlation spectroscopy. *Anal. Chim. Acta*. 544, 15-25.

de Jongh, H.H.J., Goormaghtigh, E., Ruysschaert, J.M. (1996). The different molar absorptivities of the secondary structure types in the amide I region: An attenuated total reflection infrared study on globular proteins. *Anal. Biochem.* 242, 95-103.

Delwiche, S.R.,Weaver, G. (1994). Bread quality of wheat flour by near-infrared spectrophotometry: feasibility of modeling. *J. Food Sci.* 59, 410-415.

Delwiche, S.R., Graybosch, R.A., Peterson, C.J. (1998). Predicting protein composition, biochemical properties, and dough-handling properties of hard red winter wheat flour by near-infrared reflectance. *Cereal Chem.* 75, 412-416.

Doyle W.M. (2001). Tecnical note: Comparison of near-IR and Raman analysis for potential process application. http://www.anatec.be/comm/suppliers/Axiom/axiom.htm.

Esbensen, K.H. (2000a). Principal Component Analysis (PCA) - Introduction. In: *Multivariate Data Anlysis - in Practice: An introduction to multivariate data analysis and experimental design*, (Esbensen, K.H., ed). CAMO ASA, Oslo, pp. 19-72.

Esbensen, K.H. (2000b). Multivariate Calibration (PCR/PLS). In: *Multivariate Data Analysis -in practice: An introduction to multivariate data analysis and experimental design*, (Esbensen, K.H., ed). CAMO ASA, Oslo, pp. 115-154.

Geladi, P. (2003). Chemometrics in spectroscopy. Part 1. Classical chemometrics. *Spectrochim. Acta B.* 58, 767-782.

Graener, H. (1991). Anharmonicity and overtone spectra of OH stretching vibrations. *J. Phys. Chem.* 95, 3450-3453.

Halverson, J., Zeleny, L. (1988). Criteria of wheat quality. In: *Wheat: Chemistry and Technology*, (Y. Pomeranz, eds.). AACC, Inc., St. Paul, Minnesota. pp. 15-45.

Heise, H.M. (2002a). Application of Near-Infrared Spetroscopy in Medical Sciences. In: *Near-infrared spectroscopy. Principles, instruments, applications*, (Siesler, H.W., Ozaki, Y., Kawata, S., Heise, H.M., eds.). Wiley-VCH, Weinheim, pp. 289-333.

Heise, H.,M., Winzen, R. (2002b). Chemometrics in near-infrared spectroscopy. In: *Near-infrared spectroscopy. Principles, instruments, applications*, (Siesler,H.W., Ozaki, Y., Kawata, S., Heise, H.M., eds.). Wiley-VCH, Weinheim, pp. 125-162.

Jackson, M., Mantsch, H.H. (1995). The use and misuse of FTIR spectroscopy in the determination of protein structure. *Crit. Rev. Biochem. Mol.* 30, 95-120.

Katsumoto,Y., Adachi, D., Sato, H., Ozaki, Y. (2002). Usefulness of a curve fitting method in the analysis of overlapping overtones and combinations of CH stretching modes. *J. Near Infrared Spec.* 10, 85-91.

Law, D.P., Tkachuk, R. (1977). Near infrared diffuse reflectance spectra of wheat and wheat components *Cereal. Chem.* 54, 256-265.

Lewis, R.N.A.H., McElhaney, R.N. (2000). Calorimetric and spectroscopic studies of the thermotropic phase behavior of lipid bilayer model membranes composed of a homologous series of linear saturated phosphatidylserines. Biophys. J. 79, 2043-2055.

Martens, H., Martens, M. (2000a). Why Multivariate Data Analysis?. In: *Multivariate Analysis of Quality. An Introduction*, (Martens, H., Martens, M., eds.). John Wiley & Sons, LTD, Chichester, pp. 3-23.

Martens, H., Martens, M. (2000b). Analysis of One Data Table X: Principal Component Analysis. In: *Multivariate Analysis of Quality. An Introduction*, (Martens, H., Martens, M., eds.). John Wiley & Sons, LTD, Chichester, pp. 93-110.

Martens, H., Martens, M. (2000c). Analysis of Two Data Tables X and Y: Partial Least Squares Regression (PLSR). In: *Multivariate Analysis of Quality. An Introduction*, (Martens, H., Martens, M., eds.). John Wiley & Sons, Ltd, Chichester, pp. 111-125.

Martens, H., Martens, M. (2000d). Interpretation of Many Types of Data X<=>Y: Exploring Relationships in interdisciplinary Data Sets. In: *Multivariate Analysis of Quality. An Introduction*, (Martens, H., Martens, M., eds.). John Wiley & Sons, Ltd, Chichester, pp. 139-156.

Martens H., Martens, M. (2000e). Validation X? Y??. In: *Multivariate Analysis of Quality. An Introduction*, (Martens, H., Martens, M., eds.). John Wiley & Sons, Ltd, Chichester, pp. 177-206.

Martens, H., Martens, M. (2000f). Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food Qual. Prefer.* 11, 5-16.

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. *Anal.Chem.* 75, 394-404.

Munck, L., Nielsen, J.P., Møller, B., Jacobsen, S., Søndergaard, I., Engelsen, S.B., Nørgaard, L., Bro, R. (2001). Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics. *Anal. Chim. Acta*. 446, 171–186.

Murayama, K., Yamada, K., Tsenkova, R., Wang, Y., Ozaki, Y. (1998). Determination of human serum albumin and γ-globulin in a control serum solution by near-infrared spectroscopy and partial least squares regression. *Fresen. J. Anal. Chem.* 362, 155-161.

Murray, I., Williams, P. C. (1987). Chemical principles of near-infrared technology. In: *Near-infrared technology in the agricultural and food industries,* (Williams,P., Norris, K., eds.). AACC, Inc., St. Paul, Minnesota, pp. 17–34.

Murray, I. (2004). Scattered information: philosophy and practice of near infrared spectroscopy. In: *Near Infrared Spectroscopy: Proceedings of the 11th international conference,* (Davies, A.M.C., Garrido-Varo, A, eds.). NIR publications, Chichester, pp. 1-12.

Oberg, K.A., Ruysschaert, J.M., Goormaghtigh, E. (2004). The optimization of protein secondary structure determinatino with infrared and circular dichroism spectra. *Eur. J. Biochem*. 271, 2937-2948.

Osborne, B.G. (2000). Near-infrared spectroscopy in Food analysis. in: *Encyclopedia of analytical chemistry: Application, Theory and Instrumentation,* (Meyers, R. A., ed.). John Wiley & Sons, Ltd, Chichester, pp. 4069-4082.

Ozaki, Y. (2002). Applications in Chemistry. In: *Near-infrared spectroscopy: Principles, Instruments, Applications*, (Siesler, HW., Ozaki, Y., Kawata, S., Heise, HM., eds.). Wiley-VCH, Weinheim, pp. 179-211.

Pasikatan, M.C., Steele, J.L., Spillman, C.K., Haque, E. (2001). Near infrared reflectance spectroscopy for online particle size analysis of powders and ground materials. *J. Near Infrared Spec.* 9, 153-164.

Pasquini, C. (2003). Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. *J. Braz. Chem. Soc.* 14, 198-219.

Pedersen, D.K., Martens, H., Nielsen, J.P., Engelsen, S.B. (2002). Near-infrared absorption and scattering separated by extended inverted signal correction (EISC): Analysis of near-infrared transmittance spectra of single wheat seeds. *Appl. Spectrosc.* 56, 1206-1214.

Pizzi, N., Choo, L.P., Mansfield, J., Jackson, M., Halliday, W.C., Mantshc, H.H., Somorjai, R.L. (1995). Neural-network classification of infrared-spectra of control and Alzheimers diseased tissue. *Artif. Intell. Med.* 7, 67-69.

Reich, G. (2005). Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications. *Adv. Drug Deliv. Rev.* 57, 1109-1143.

Šašić, S., Ozaki, Y. (2000). Band assignment of near-infrared spectra of milk by use of partial least-squares regression. *Appl. Spectrosc.* 54, 1327-1338.

Sato, T., Morishita, T., Hara, T., Suda, I., Tetsuka,T. (2001). Near-infrared reflectance spectroscopic analysis of moisture, fat, protein, and physiological activity in buckwheat flour for breeding selection. *Plant Prod. Sci.* 4, 270-277.

Schmitt, J.M., Kumar, G. (1996). Spectral distortions in near-infrared spectroscopy of turbid materials. *Appl. Spectrosc.* 50, 1066-1073.

Schrader, B. (1995). Infrared and Raman spectroscopy: Methods and applications. Wiley-VCH, Weinheim.

Scheiner, S. (1997). Hydrogen bonding, a theoretical perspective. (Truhlar, D.G., ed.). Oxford University Press, New York.

Sevenou, O., Hill, S.E., Farhat, I.A., Mitchell, J.R. (2002). Organisation of the external region of the starch granule as determined by infrared spectroscopy. *Int. J. Biol. Macromol.* 31, 79-85.

Shaw, R.A., Mantsch, H.H. (1999). Vibrational biospectroscopy: from plants to animals to humans. A historical perspective. *J. Mol. Struct.* 481, 1-13.

Šoptrajanov, B. (2000). Very low H-O-H bending frequencies. I. Overview and infrared spectra of $NiKPO_4*H_2O$ and its deuterated analogues. *J. Mol. Struct.* 555, 21-30.

Swanton, D.J., Bacskay, G.B., Hush, N.S. (1986). The infrared-absorption intensities of the water molecule – a quantum chemical study. *J. Chem. Phys.* 84, 5715-5727.

Tolstoguzov, V. (1996). Structure-property relationships in foods. In: *Macromolecular Interactions in Food Technology*, (Parris, N., Kato, A., Creamer, L.K., Pearce, J., eds.). Am. Chem. Soc., Washington DC, pp. 2-14.

Tolstoguzov, V. (2003). Some thermodynamic considerations in food formulation. *Food Hydrocolloid.* 17, 1-23.

Vanderkooi, J.M., Dashnau, J.L., Zelent, B. (2005). Temperature excursion infrared (TEIR) spectroscopy used to study hydrogen bonding between water and biomolecules. *BBA-Proteins Proteom.* 1749, 214-233.

Ventura, C., Papini, M. (1999). Analysis of the reflectance of granular materials in the near-infrared wavelength range. *J. Quant. Spectrosc. Ra.* 61, 185-195.

Wesley, I.J., Larsen, N., Osborne, B.G., Skerritt, J.H. (1998). Non-invasive monitoring of dough mixing by near infrared spectroscopy. *J. Cereal Sci.* 27, 61-69.

Wesley, I.J., Larroque, O., Osborne, B.G., Azudin, N., Allen, H., Skerritt, J.H. (2001). Measurement of gliadin and glutenin content of flour by NIR spectroscopy. *J. Cereal Sci.* 34, 125-133.

Wold, S., Sjöström, M., Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Sys.* 58, 109-130.

Chapter 2: Effects of water in infrared spectroscopy

The solubilising properties of water result from its high ability to form hydrogen bonds, and MIR and NIR, being sensitive to the hydrogen bonding state of molecules, allow the observation of the hydrogen bonded network of water and its perturbations -even in complex systems [Gergely, 2003; Marechal, 1997]. The infrared water spectrum may thus be very informative and bring information on the structure of water and on solute-water interactions. On the other hand, some of the spectral water variations are irrelevant (e.g. reflect the measurement conditions) and hampers the study of the interesting phenomena. Sections 2.3 and 2.4 deal with the spectral variations caused by temperature and salts, and in section 2.5 is shown how EMSC pretreatments can be applied on MIR spectra for removal of these irrelevant variations, thus leading to better spectroscopic analyses of biological specimen. First, the infrared water spectrum is introduced.

2.1. The infrared water spectrum

In MIR, water appears almost 'black' and let no light pass unless the path length is extremely short (few microns). One way to obtain the short path length is to use the attenuated total reflection (ATR) mode. In this mode, the sample is placed in contact with a crystal e.g. of ZnSe, Ge or diamond, in which internal reflections generate an evanescent wave. Each time the radiation impinges on the crystal-sample interface, it penetrates a few microns into the sample and is attenuated before being transmitted to the detector.

In NIR, the absorptivity from water is much less than in MIR, and water studies do not require the very short path lengths.

The infrared water spectrum

The H_2O molecule has three normal modes of vibrations: Symmetric stretching vibrations (ν_1), bending (scissoring) vibrations (ν_2) and antisymmetric stretching vibrations (ν_3). These are all infrared active, and, in ATR-spectra, the fundamental absorption bands of bulk liquid water appear at around 3405 cm^{-1} (ν_3), 3210 cm^{-1} (ν_1) and 1635 cm^{-1} (ν_2) [Max, 1999]. The water bands are somewhat downshifted in ATR-FTIR spectra compared to transmission spectra owing to an 'anomalous dispersion effect' [Grdadolnik, 2002]. Although this effect has influence on both shape and position of the water bands, the ATR mode is considered a powerful method for water analysis [Chen, 2004; Marechal, 1991,1993].

Hydrogen bonding between water molecules in the liquid state has a huge influence on the vibrational frequencies, so the OH-stretching (OH-str.) bands of gaseous water are shifted several hundred wavenumbers up compared to those of liquid water. Thus, the water spectrum also reflects the hydrogen bonding state of liquid water, and, as the degree of hydrogen bonding decreases linearly with increasing temperature, the temperature variations are reflected as shifts of the water bands. This effect is dealt with in sections 2.3.

A broadness of the water bands owes to anharmonic coupling between vibrations of similar energy on neighbouring molecules and probably to coupling of the intramolecular vibrations to the intermolecular vibrations between hydrogen bonded molecules [Marechal, 1991]. Also, the wide range of hydrogen bonding configurations in liquid water causes inhomogeneous line broadening [Chaplin, 2006]. Thus, the ν_1 and ν_3 vibrations appear as one broad absorption envelope for liquid water. In the spectrum of liquid water is, close to the far-infrared region, seen a libration band (ν_L) at ~800 cm^{-1}, which is due to reorientation of the water molecules in the hydrogen bonded network, and which is sensitive to the dynamic properties of water [Gaiduk, 2004]. A minor band at ~2165-2127 cm^{-1} has been ascribed to the combination of ν_2 and ν_L [Marechal, 1991; Chaplin, 2006] and some very weak bands at ~1300 cm^{-1} and 3200-3260 cm^{-1} to the 1st overtones of ν_L and ν_2, respectively [Fischer, 2001; Marechal, 1993].

In NIR spectra, six water bands are evident, the assignments of which are shown in Table 2.1. In addition, a very weak band at 2083 nm has been reported as the 2nd overtone of HOH-bending (3*ν_2) [Wang, 1998].

Wavelength [nm]	Wavenumber [cm^{-1}]	Common assignment		Alternative assignment
1930	5180	OH-str.+HOH-bend.	$\nu_{1,3}+\nu_2$	
1440	6940	1st overtone of OH-str.	$2*\nu_{1,3}$	
1190	8400	OH-str.+HOH-bend.	$2*\nu_{1,3}+\nu_2$	$\nu_{1,3}+3*\nu_2$
970	10,310	2nd overtone of OH-str.	$3\nu_{1,3}$	
840	11,900	OH-str.+HOH-bend.	$3*\nu_{1,3}+\nu_2$	$\nu_{1,3}+5*\nu_2$
760	13,160	3rd overtone of OH–str.	$4*\nu_{1,3}$	

Table 2.1. Water absorptions and assignments in the NIR region [Chaplin, 2006].

In contrast to the liquid water bands, the water vapour bands are of a high-frequent structure. The fundamental rotation-vibration bands are centred at 3755 (ν_3), 3657 (ν_1) and 1594 cm^{-1} (ν_2) [Chaplin, 2006]. These bands result from vibration of the bonds simultaneously with rotation of the molecules. (The theory of rotation-vibration spectra from gaseous molecules is described in Heise et al (1995). The rotation-vibration band associated with ν_2 overlaps with some important protein bands, and its removal is essential in the MIR protein structure analysis. A method for removal of the water vapour spectrum from Fourier Transform Infrared (FTIR) spectra is described in Paper II (Appendix II).

2.2. Water structure

Even though water is one of the most studied chemical systems, its hydrogen bonding configuration and dynamic, which is thought to govern the anomalous properties of water, is far from understood.

Hydrogen bonding of H_2O in liquid water

Water is a highly dipolar molecule and it has the capacity to participate in hydrogen bonds as donor and acceptor simultaneously. The water molecule may participate in maximum four hydrogen bonds: The oxygen atom may act as acceptor of two hydrogen bonds and the OH groups may act as hydrogen donors in two other hydrogen bonds [Vanderkooi, 2005]. The fully coordinated water molecule has a tetrahedral arrangement around the oxygen atom, and the resulting hexagonal structure is the primary arrangement found in ice [Vanderkooi, 2005]. See Fig. 2.1.

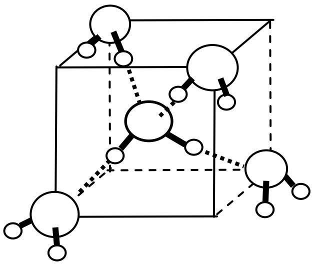


Fig. 2.1. Outline of the tetrahedral arrangement of the fully hydrogen bonded water molecule in ice.

Liquid water has a high degree of hydrogen bonding as well; it has been estimated that less than 5% of OH groups in water at room temperature are free (non-hydrogen bonded) [Marechal, 1993]. Many studies suggest that liquid water consist of the hexagonal structures found in ice, though with many defects [Vanderkooi, 2005]. The new emerging comprehension is that the presence of bend or distorted (but not broken) hydrogen bonds distinguishes liquid water from ice and provide water its fluidity, and the existence of weak bifurcated hydrogen bonds in liquid water is supported from several studies [Walrafen, 1989; Giguére, 1987; Khoshtariya 2002; Rull, 2002]. The bifurcated configuration involves weak hydrogen bonding of a hydrogen atom to two other water molecules [Walrafen, 1989].

Recently Wernet et al (2004) suggested that the dominating configuration of H_2O in liquid water is asymmetric: It has a strong donor- and a strong acceptor hydrogen bond, while the remaining possible hydrogen bonds of the water molecule are weak or broken (single donor configuration) [Wernet, 2004]. This configuration is favoured by the cooperative and anticooperative effects. According to the authors, also the fully coordinated water molecule, as seen in ice, constitutes a small part (20 %) of liquid water in a broad temperature range.

Models of the water structure

The models for water structure can broadly be classified into two classes, namely the mixture models and the continuum models. The original mixture model describes water as being composed of a limited number of water species (different ring structures or cluster types) that coexist in a temperature-and pressure-dependent equilibrium [Eisenberg, 1969]. The continuum model on the other hand implies that hydrogen bonds are equally distributed in the water sample, and that there is a continuous weakening of the hydrogen bonds with increasing temperature [Wall, 1965]. None of the two models has been finally proved to be the correct model and the different models need to be taken into account in order to explain the many anomalous properties of water [Khoshtariya, 2002].

Some recent two-component models invoke that water is composed of two coexisting phases: a low-density phase and a high-density phase with a high local tetrahedral ordering [Rønne, 2000; Chaplin, 2006]. One of the theories involves the rapidly interconverting dodecahedral and isocahedral clusters, consisting each of 280 water molecules [Chaplin, 2006], and an outer two-state model describes water as a mixture of different ice-like species (Ice Ih and Ice II type), which can rearrange on a pico second scale [Urquidi, 1999]. Thus, the hydrogen bonded structures in water are considered to be transient, with hydrogen bonds constantly breaking and reforming. A common feature among the recent models is the presence of interstitial (weakly bound) water molecules sitting in the cavities of a distorted tetrahedrally bonded network, resulting in a higher density than the pure tetrahedral arrangement [Rønne, 2000; Wernet, 2004].

2.3. Temperature effects

The hydrogen-bonded network structure in liquid water and its dependence on temperature has been subject to several MIR and NIR studies. The presence in the spectra of nearly isosbestic points, i.e. frequencies at which the absorbance is independent of the temperature, has usually been taken as support of the mixture model. However, MIR and NIR only provide indirect evidence as to the water structure.

Temperature-dependent spectral changes

Fifteen properties of water have been estimated from NIR spectra of water at different temperatures. These include e.g. density, refractive index, dielectricity constant, relative viscosity, isothermal compressibility, thermal expansitivity, thermal conductivity, surface tension, vapour pressure, enthalpy, entropy, free energy and ionisation constant [Lin, 1994]. The excellent correlations to all properties reflect the sensitivity of NIR to both hydrogen bonding and the packing density of water. Whereas variations in packing density cause scaling effects, the hydrogen bonding variations affect position and relative intensities of the water bands.

The effect of temperature on the fundamental and some overtone and combination bands are shown in Fig. 2.2 and 2.3.

Shifts: For both the MIR and NIR spectra, the effect of increasing temperature is apparently shifts to higher frequencies (lower wavelengths) of the bands that involve OH–str. vibrations ($\nu_{1,3}$). (As described in chapter 1, low-frequency shifts occur proportional to the strength of the hydrogen bonds). Nearly isosbestic points are seen in the NIR spectra in Fig. 2.3. but also appears for the MIR spectra. The position of the bending band (ν_2) is only little affected, and even though a low-frequency shift has been frequently reported for the ν_2 band, this shift has been shown to be caused by the temperature variations of the bands on which ν_2 is superimposed (ν_2+ν_L and 2*ν_L) [Marechal, 1993].

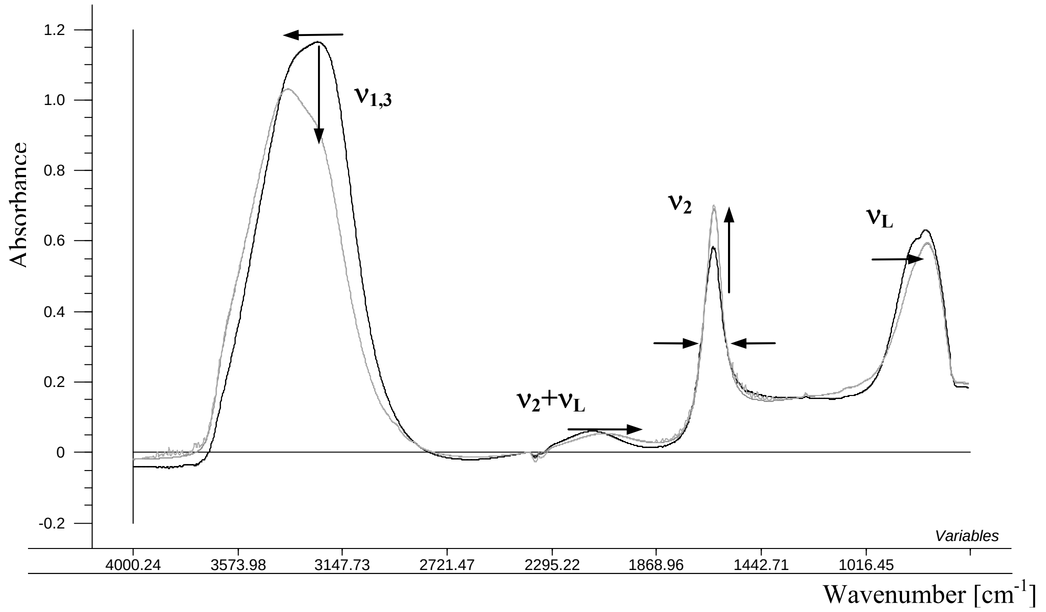


Fig. 2. 2. ATR MIR spectra of water at temperatures at 9 °C and 63°C. Some effects of increasing temperature is shown by the arrows.

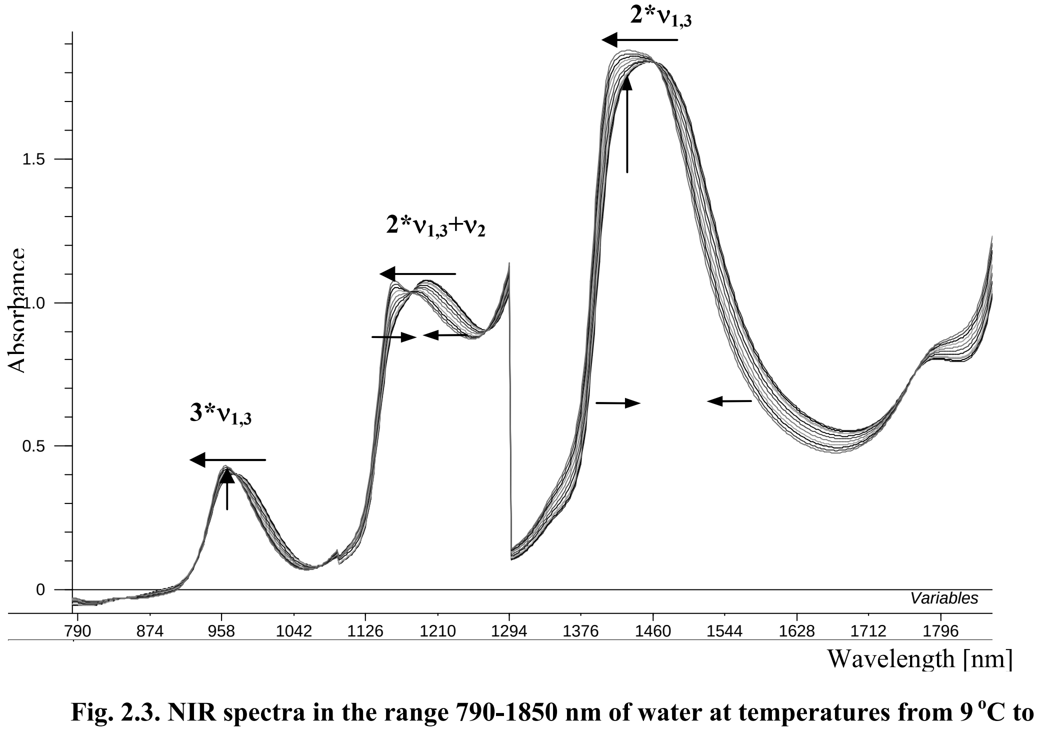


Fig. 2.3. NIR spectra in the range 790-1850 nm of water at temperatures from 9 °C to 63°C. Some effect of increasing temperature is shown by the arrows.

Thus, the shifts of the combination bands are caused by shifts of the stretching or libration bands. For example, the shift of $\nu_2+\nu_L$ in MIR is governed by ν_L, which shifts to lower frequencies upon increasing temperatur [Marechal, 1993]. Even though the reorientational motion that results in the ν_L band is frequently depicted to be governed by unspecific interactions, it is essentially influenced by the hydrogen bonding state and thus by the temperature [Gaiduk, 2004], and the shift to lower frequencies upon increasing temperature is due to the weakening of the water structure [Marechal, 1993]. The combination bands in NIR show high-frequency/low-wavelength shifts similar to the $\nu_{1,3}$ band (Fig. 2.3) although the shifts may be smaller [Chaplin, 2006].

Differences between NIR and MIR: The intensities of the fundamental bands are not affected in the same way by temperature as are the intensities of overtone and combination bands. As the temperature increases, the intensity of the fundamental $\nu_{1,3}$ band decreases. In contrast, an opposite increase is observed for its overtone and combination bandsl as a result of anharmonicity affecting the two regions in different ways (see section 1.5) [Graener, 1991]. Similar to the overtone bands, the fundamental ν_2 band increases with increasing temperature (in fact, the ν_2 band is almost vanished for ice) [Chaplin, 2006; Vanderkooi, 2005]. In contrast, due to coupling with the $\nu_{1,3}$ band, the minor ν_2 $+\nu_L$ band decreases at increasing temperature [Marechal, 1993]. Also, with increasing temperature, the fundamental OH-str. band shows minor broadening, and the HOH-bending band becomes ~20 % narrower (Fig. 2.2) [Marechal, 1991; Czarnik-Matusewicz, 2005; Toft, 1996]. In contrast to the fundamental OH-str. bands, the overtone OH-str. bands also become narrower with increasing temperature [Czarnik-Matusewicz, 2005; Libnau, 1994]. The opposite effects of temperature on the bandwidths in the two regions may result from the inter- and intramolecular anharmonic couplings [Marechal, 1991].

The opposed effects of hydrogen bonding on band intensities and shapes in NIR and MIR offer complementary information to the two infrared regions and a few studies have combined information from the two regions [Czarnik-Matusewicz, 2005; Libnau, 1994].

Assessment of the water structure from temperature studies

Analyses of the MIR and NIR spectra at different temperatures by use of PCA, PLSR and multivariate curve resolution techniques have provided indications as to the number of water configurations in the temperature-dependent equilibrium. In PCA/PLSR studies of water between 6-80°C, the main spectral variation as described in PC1 and accounting for more than 99 % of the variation has indicated an interchange between two groups of strongly and weakly hydrogen bonded species, respectively [Segtnan, 2001; Libnau, 1994]. For example in an analysis of the temperature variations in the OH-str. overtone band, the first loading vector reflected a shift between 1492 nm (from the strongly hydrogen bonded structure) and 1412 nm (from the weakly hydrogen bonded structure) [Segtnan, 2001]. Also, in an analysis of the fundamental OH-str. band, two spectral profiles could be extracted and these showed OH-str. maxima at 3410 and 3240 cm^{-1} from the two components increasing and decreasing, respectively, with temperature [Libnau, 1994]. Likewise, a rank two-result in a curve resolution study of the ν_2 band has supported the two-component model [Toft, 1996]. Minor deviations from the two-component models have been ascribed for example to the presence of a third species (as suggested for water below 20°C), a continuum of hydrogen bonding interactions or to band broadening and small shifts [Libnau, 1994; Czarnik-Matusewicz, 2005]. However, most authors conclude that water can be seen as a pseudo two-component system [Libnau, 1994; Toft, 1996, Šašić, 2002].

Additional information on the water structure has been obtained from band deconvolution and calculation of the sub band areas in order to assess the relative concentrations of the differently hydrogen bonded OH-groups at different temperatures. Maeda et al (1995) analysed the first OH-str. overtone at 1440 nm of water between 5 and 85°C and found from the 2nd derivative spectra five sub bands, which did not change much in position with temperature. These were assigned to water with zero to four hydrogen bonds: S0 (1410 nm), S1 (1439 nm), S2 (1456 nm), S3 (1553 nm) and S4 (1642 nm), unlike the common assignment made to the four sub bands of the fundamental OH-str. band (See Table 2.2).

Sub band position [cm^{-1}]	OH species	Hydrogen bonding
3250	Icelike	Strong
3380	Icelike-liquid	Normal
3540	Liquid-like amorphous	Defect
3670-3650	Free	Non hydrogen bonded

Table 2.2. Common assignments for the sub bands obtained in the deconvolution of the OH-str. band [Khostariya, 2002].

A decrease of the species S2, S3 and S4 and an increase of S0 were observed with increasing temperature. Likewise, five sub bands have been fitted to the water bands at 970 nm and 1190 nm [Abe, 2004]. From the integrated sub band areas of the 2nd overtone band (970 nm), Abe (2004) estimated the average number of hydrogen bonds (N_h) as a function of temperature and obtained good agreement with data obtained from X-ray data (N_h =2.63 at 0°C and 2.12 at 98°C). The five components were also classified into a strongly hydrogen bonded group and a weakly hydrogen bonded group in agreement with the two-component model.

In a study of the band at 1930 nm, an observed interconversion between two major sub bands with temperature was attributed to a redistribution between OH-groups with normal hydrogen bonds (2026 nm) and OH groups with temperature-bifurcated hydrogen bonds (1914 nm) [Khoshtariya, 2002]. Two other minor sub bands assigned to ice-like species with strong hydrogen bonds (2146 nm) and to free non-hydrogen bonded species (1835 nm), respectively did not change much with temperature. Thus, in the temperature-dependent equilibrium outlined by Khoshtariya et al (2002), only the 2026 and 1914 nm OH-groups were involved. An additional temperature-insensitive OH-group (at 1970 nm) was thought to represent a moderate distortion of the hexagonal ice structure. The authors found the data to support the tetrahedral displacement mechanism, in which the net outcome is the conversion of a normal hydrogen bond to a bifurcated hydrogen bond as a tetrahedrally bound water molecule moves to an interstitial position [Khostariya, 2002; Agmon, 1996]. This reaction, which was suggested important for the anomalous properties of water, was also suggested to be the mechanism underlying the temperature-dependent equilibrium between normal hydrogen bonds and bifurcated hydrogen bonds [Khoshtariya, 2002].

2.4. Salt effects

When a salt is dissolved in water, the anions and cations interact with water and become hydrated. The interactions cause perturbations of the water spectrum, since the vibrational properties of the bound water molecules are different from those of bulk water. The perturbation is salt specific, as are the effect of salts on various phenomena in biology (described more in chapter 5).

The Hofmeister series

Franz Hofmeister et al published in 1880-1890 several papers concerning the study on what he called 'the water withdrawing power of salts' [Kunz, 2004a]. Hofmeister first showed that the amount of salt needed for the precipitation of hen egg white lysozyme depended on the type of salt. Then he arranged cations and anions in series after their effectiveness on protein precipitation, since he observed the same order of the salts for hen egg white globulin and blood serum globulin.

The effect of salts on protein solubility is a balance between salting-out of hydrophobic groups on the proteins and salting-in of polar peptide groups, and the Hofmeister series arranges anions and cations according to these properties [Baldwin, 1996]. The anion and cation series are shown below.

PO_4^{3-}>HPO_4^{2-}=SO_4^{2-}>citrate->acetate$^-$> Cl^->Br^->NO_3^->ClO_4^->SCN^-
Kosmotropic------------------------------------ chaotropic

NH_4^+>K^+>Na^+> Li^+>Mg^{2+}>Ca^{2+}
Chaotropic ---------- kosmotropic

The first ions in the two series reduce the non-polar solubility in water, stabilise protein conformation and cause precipitation of proteins in their native form (salting-out). This salting-out property diminishes throughout the series, and the last ions increase the non-polar solubilities and cause protein solubilisation and denaturation (Salting-in) [Kalra, 2001]. The series are not definitive as the order can vary for different proteins, pHs, temperatures and counter-ions. In certain cases, the series can even be reversed [Ebel, 1999; Wiggins, 2000; Chaplin, 2006]. The sequences also reflect the efficiency of the ions on numerous other phenomena of concern in chemistry and biology, e.g. ion binding to micelles, protein and colloid stability and protein conformations [Kunz, 2004b]. The effect on protein conformations is demonstrated in chapter 5.

Hofmeister reasoned that the precipitation effectiveness was related to the attracting forces between the salt and water molecules leading to solvent removal from the protein [Kunz, 2004]. In fact, the two series reflect the abilities of the ions to interact with water molecules. Based on these properties, the ions have been classified into structure-makers (kosmotropes) and structure-breakers (chaotropes). The kosmotropes are typically ions smaller than Cl^- (for anions) or K^+ (for cations) with a high charge-density that bind water strongly and are highly hydrated, whereas the chaotropes are larger monovalent ions with a low charge-density that bind water weakly and are less hydrated [Collins, 1997; Chaplin, 2006]. It was originally hypothesized that structure-makers could act as nucleation sites for ice-like structures in the liquid, whereas structure-breakers would disrupt these structures. However, the concept of long-ranging structure-breaking or structure-making effects of ions has been questioned, and evidence points to the fact that there is no large-scale effects on the water structure, and that only water molecules in the vicinity of the ions are affected [Omta, 2003]. The increased viscosity induced by kosmotropic ions, as expressed by a positive Jones-Dole B-coefficient, is explained from the rigidity of the solvation structure of the ion and its first solvation shell and not by an increased hydrogen bonding in bulk water [Omta, 2003].

For both anions and cations, there is electrostatic interaction with nearby water molecules. These try to align their dipole moment in the direction of the ion, so that the positive end points towards anions and the negative end towards cations [Hribar, 2002]. The water-cation interaction is a complex dipolar interaction that involves the free lonepair on the water molecules, while the water-anion interaction may entail a hydrogen bonding interaction with the hydrogen atom of water [Symons, 1975; Kropman, 2003].

Due to the ion-water interactions, water arranges in coordination/hydration shells around the ions. In the first coordination shell of the ion, there is a competition between the electrostatic ion-water interaction and the water-water hydrogen bonding interaction, so, depending on the charge density of the ion, the water become more or less oriented towards the ion [Hribar, 2002]. For small anions with a high charge density (kosmotropes), the electrostatic interactions dominate over water-water interactions and water molecules are highly ordered, whereas for larger anions (chaotropes), the weaker electrostatic interaction with water causes the water molecules to be less strictly aligned towards the ion [Hribar, 2002]. Thus, the kosmotropic and chaotropic properties are associated with low and high entropies, respectively, of the hydration water [Hribar, 2002]. The chaotropic property of K^+ has been corroborated in a molecular dynamic simulation study, which showed that the exchange of water molecules in the hydration shell of these ions is higher than in bulk water [Tongraar, 2004].

Salt-dependent spectral changes

Adding salt to water influences the physical and chemical properties of water and causes various effects in the infrared water spectrum. At increasing salt concentration, the water concentration decreases and the refractive index increases, and as a result, additive baseline variations and multiplicative effects appear in the spectra. On the other hand, the ion-water interactions affect the band shapes and positions. In agreement with this, both a baseline variation and some distinct peaks have been shown of importance for the discrimination of mineral waters of different salt compositions by use of the 1100-1800 nm spectra [Tanaka, 1995]. The peaks seen in the 2nd derivative NIR spectra were thought to result from the perturbation of the water structure by the salts, since most inorganic ions cause no absorption peaks by itself in the NIR range.

The perturbations of the infrared water spectra have been found to be rather characteristic of each electrolyte [Lin, 1994; Fischer, 2001; Liu, 2005; Wei, 2005]. For example, it has been possible to estimate the individual concentrations of Cl^-, Na^+, K^+ and SO_4^{2-} in seawater samples using the 1100-1800 nm range with correlations of more than 0.7 [Chen, 2003]. The spectral effects represent the perturbations due to both anion and cation and generally the two effects are additive an independent [Chen, 2004; Fischer, 2001].

Anion effects: The chaotropic and kosmotropic anions cause different shifts in the infrared water bands, due to the different strengths of water-anion interactions for the two groups. The perturbations of the vibrational frequencies may result directly from the water-anion hydrogen bonding interaction [Fischer, 2001]. The chaotropic KCl has been found to produce spectral changes similar to those observed from a temperature-increase, demonstrating the structure-breaking ability of the salt [Max, 1999]. Likewise, in a Raman study of the weakly chaotropic perchlorate anion, the OH str.-band of 1.5M $LiClO_4$ at 45°C could be modelled from the Gaussian components obtained from the pure water spectrum, and this spectrum coincided almost perfect with the pure water spectrum at 60°C [Neto, 2003]. In order to obtain the spectral signature of the ion hydration water, the ATR-MIR difference spectra (after subtraction of the pure water spectrum) have been looked upon [Wei, 2005; Liu, 2005].

The difference spectra have revealed a decrease at 3203 -3196 cm^{-1} and an increase at ~3585 cm^{-1} in the OH-str. band, upon increasing concentration of ClO_4^- [Wei, 2005]. These effects were thought to reflect the breakage of a strong hydrogen bonds in the icelike structure and the substitution with a weak hydrogen bond to ClO^-, according to the assignments in Table 2.2 [Wei, 2005]. With decreasing size of the anion (more kosmotropic), the positive peak in the difference spectra is found at lower frequencies [Fischer, 2001].

Sulphate, which is known as a kosmotrope, results in ion-water hydrogen bonds of almost similar strengths as between water molecules [Wei, 2005]. Thus, the only spectroscopic effect of the sulphate ions is a decrease of the high-frequency component at 3670-3650 cm^{-1}, where the sub band from free OH-groups appear, as weak hydrogen bonds from monomeric water are substituted with hydrogen bonds to the anions [Jin, 2003; Wei, 2005].

Cation effects: Also cations affect the shape of the water bands due to their electrostatic interactions with the surrounding water molecules. The effects are not merely shifts but may be rather complex changes, causing various positive and negative peaks in the difference spectra [Fischer, 2001; Wei, 2005].

2.5. Removal of temperature and salt effects (Paper III)

The following paper describes a preprocessing method based on EMSC for removal of the irrelevant temperature and salt effects in ATR-FTIR spectra. The paper also concerns the extraction of relevant sample information from the water spectrum as regards hydration phenomena.

Correction for temperature- and salt- effects in water in FTIR biospectroscopy

Harald Martens (1,2,3), Susanne Bruun (4) and Achim Kohler (2,5)

(1) Center for Biospectroscopy and Data Modelling, Matforsk, Norwegian Food Research Institute, N-1430 Aas, Norway. (2) CIGENE/IKBM, Norwegian University of Life Sciences, N-1430 Aas, Norway
(3) The Royal Veterinary & Agricultural University, 1870 Frederiksberg C, Denmark. (4) Biochemistry and Nutrition Group, BioCentrum-DTU, Technical University of Denmark, building 224, Søltofts Plads, DK-2800 Kgs. Lyngby, Denmark. (5) Unité de Sensométrie et de Chimiométrie, ENITIAA, BP 82225, 44322 NANTES CEDEX 3, FRANCE .

Submitted to Journal of Chemometrics, March 15, 2006.

KEYWORDS: ATR-FTIR; Extended Multiplicative Signal Correction; water; temperature; salts.

SUMMARY

Multivariate modelling methods are presented for stabilizing Fourier transform infrared spectroscopy (FTIR) in biospectroscopy. First, the irrelevant gas contributions from water vapour and CO_2 in the instrument light path is modelled and removed. By use of Extended Multiplicative Signal Correction (EMSC), variations in the Attenuated Total Reflection (ATR)-FTIR spectrum of water caused by temperature and salts ($MgSO_4$, $NaClO_4$, NaCl) are estimated and removed from the spectra of biological samples in order to improve their analysis. These effects were described quantitatively for use in EMSC based on ATR-FTIR spectra of water solutions, and the models were tested successfully in the in-vivo monitoring of *Candida albicans* growing on the ATR crystal of the same instrument, as well as for gluten/water/salt mixtures measured in a rather different ATR-FTIR instrument. The spectral variations remaining after subtraction of the estimated temperature and salt effects are likely to reveal relevant chemical information of the samples

Keywords: ATR-FTIR, Water, salt effects, temperature effects.

1. INTRODUCTION

In biological sciences, Fourier transform infrared spectroscopy (FTIR) has proven to be an important tool for measuring a chemical fingerprint of very different samples [1-3]. The water spectrum often dominates in FTIR spectra of biological specimens, as water is a major part of living cells and is highly absorptive in the mid-IR. Due to the high absorptivity of water in the mid-infrared region, a very thin transmission cell or the Attenuated Total Reflection (ATR) technique is needed for IR analyses of aqueous samples in order to prevent saturation of the water peaks. The ATR-FTIR technique provides spectra of a high signal to noise ratio, is easy to use, and thus has found widespread application for analysis of biological/microbial samples [1-3].

The sensitivity of FTIR towards hydrogen bonding means that perturbations of the hydrogen bonded network in liquid water has great impact on the spectra. ATR-FTIR is therefore very suitable for studying hydration phenomena, as well as the hydrogen bonded structures in water and its temperature-dependent variations [4, 5]. Liquid water seems to maintain different hydrogen-bonded structures, the distribution of which is affected by the other constituents (proteins, carbohydrates, lipids, nucleic acids, salts etc.), as well as the physical state of the system (temperature) [4,5]. Thus, the FTIR spectra, reflecting the state of water, may bring highly relevant information on the system.

On the other hand, a frequent problem encountered in FTIR analyses of biological samples relates to temperature fluctuations during the measurement series, as the hereby introduced variations are irrelevant and hamper the analysis of the phenomena of interest. The temperature variations result in apparent frequency shifts of the water bands and also affect their intensities. Furthermore, varying types and concentrations of ions present in the samples may cause unwanted spectral variations, since the hydration water of ions may obtain different vibrational properties compared to bulk water [6].

Varying physical properties of the samples may cause other unwanted spectral variations, such as additive and multiplicative effects, since the penetration depth of the IR light depends on the refractive index of the sample and is wavenumber-dependent. However, these effects can be separated from the chemical variations prior to the analysis by means of Extended Multiplicative Signal Correction (EMSC) [7]. In addition, this method provides the possibility to remove the unwanted spectral variations caused by chemical perturbants or temperature variations etc. This is done by use of empirical model spectra, describing the unwanted phenomena.

The present paper develops empirical models for the effect of temperature and of salts on FTIR spectra of water. The models are based on ATR-FTIR spectra from controlled experiments in water solutions with various temperatures and salt contents. The results are tested on two sets of biologically relevant aqueous samples. The water-temperature model are tested on spectra of a colony of *Candida albicans* growing or decaying directly on the same ATR crystal used for the above water measurements. The salt models are tested on a simplified bread dough model: mixtures of aqueous samples of wheat storage protein (gluten) with various known levels of salts added, measured in another country in a different brand of FTIR instrument.

Effect of temperature and salts on IR spectra of water

The temperature-induced variations in the IR water spectrum have been investigated in a number of studies with the purpose of increasing the knowledge of the water structure, which is still far from understood. The IR studies have mainly provided indications to the number of water species (with different hydrogen bond configurations) present in a temperature-dependent equilibrium in liquid water and to their relative concentrations [8-10]. With decreasing temperature of liquid water, the water bands show apparent frequency shifts, reflecting the increased hydrogen bonding of the water structure. For example, a low-frequency shift of the OH stretching band $\nu_{1,3}$ appears upon hydrogen bond, since the interaction weakens the covalent O-H bond [11]. Hydrogen bonding further has the effect of increasing the $\nu_{1,3}$ band intensity due to an increase of the dipole moment [11].

In addition to temperature, the presence of various solutes, interacting to different extents with water, affects the infrared water spectrum. For example, ions in the aqueous samples bring about spectral perturbations, which are greatly determined by types and concentrations of the ions. Both anions and cations have long been classified into structure-makers (kosmotropes) and structure-breakers (chaotropes) based on their ability to interact with water and affect its properties. Although the effect of kosmotropes and chaotropes on the bulk water struture is questioned [12], it is certain that water in the vicinity of ions responds in different ways to the ions. The kosmotropic and chaotropic property is thus associated with a decrease and increase, respectively, of the entropy of the hydration water: Water molecules surrounding the kosmotropes become strictly oriented towards the ions due to strong electrostatic interactions, whereas a more disordered structure is allowed for water surrounding the chaotropic ions [13].

The kosmotropic anions are typically ions smaller than Cl^- of high charge densities such as sulphate and phosphate ions, and they form strong hydrogen bonds to water [14, 15]. For example, the sulphate ion is thought to cause the exchange of weak water-water hydrogen bonds with normal strength ion-water hydrogen bonds. This is e.g. reflected in the IR water spectrum as a minor decrease at 3670-3650 cm^{-1} in the OH-str. band, where the weakly bound water absorb [6]. On the other hand, the chaotropic anions posses low charge densities and are characterised by weak hydrogen bonds to water. In accordance, the effect on the IR water spectrum of the chaotropic Cl^- and ClO_4^- have been found similar to that observed from temperature increases [16-17]. An observed decrease at 3203 -3196 cm^{-1} and an increase at ~3585 cm^{-1} induced by ClO_4^- have been interpreted as the breakage of a strong hydrogen bond in the 'fully hydrogen bonded five molecule tetrahedral nearest neighbour structure', and the substitution with a weak hydrogen bond to ClO_4^- [6].

Cations influence the IR water spectrum as well, due to the electrostatic interactions, which affect the vibrational properties of the bound water molecules. The higly kosmotropic Mg^{2+} is found to increase a low-frequency component at 3136 cm^{-1} of the OH-str. band, whereas, the weaker kosmotrope Na^+ is found to cause an increase at 3423 cm^{-1}. The low-frequency shift of the OH-str. band seen for cations, interacting strongly with water, may be due to an effect on the H-O-H bond angle and the charge separation in the water molecules [18]. In case of divalent cations, which polarise the bound water, an increased interaction of water in the first hydration shell with water in a second hydration shell has also been described as a cause of the observed shift [15].

Notation and terminology

Spectroscopically, I means measured light intensity, while A means absorbance, defined as $\log_{10}(I)$, or $\log_{10}$ (I/I0), when mentioned explicitly. I(ν) means measured intensity at ν cm^{-1}.

For modelling, upper-case and lower-case bold characters represent matrices and vectors, respectively; italics represent scalars. An absorbance spectrum # i is denoted by row vector $\mathbf{z}_i = z_{i,k}$, k=1,2,...,K, where *k* represents wavenumber channel # (here at 2 cm^{-1} intervals).

2. MATERIALS AND METHODS

2.1 Experiments A and B: Water at different temperatures

ATR-FTIR spectra Milli-Q water were obtained on a Bruker Equinox 55 FTIR spectrometer, recording from 4000 to 600 cm^{-1}. Samples at different temperatures were measured on an ATR-crystal (ZnSe) mounted in a closed sample cell, which allowed for temperature control by circulation of water from a water bath with heating and cooling ability. Each spectrum was the result of the co-addition of 128 or 256 scans, obtained in single beam mode, with a resolution of 4 cm^{-1} and a data interval of ~2 cm^{-1}. The absorbance spectra were calculated by subtraction of the background spectrum obtained on the empty ATR-crystal at room temperature in the beginning of each measurement series.

Experiment A (calibration set A-1, test set A-2)

Temperature-scanning series were obtained by automatic spectrum collection at specific time intervals, while the temperature in the water bath either increased or decreased at a fixed rate. The temperature range was ~10-35°C for the calibration set A-1 and 10-60°C for the test set A-2.

Experiment B (calibration set B)

Spectra of water at known temperatures (8-63°C) were obtained by measuring the sample temperature by use of a digital thermometer dipped into the sample cell immediately before and after collection of the spectrum. The mean temperature was used. The temperature drift in each measurement was less than 0.5°C.

2.2 Experiment C: *Candida albicans* growth and decay

C. albicans (strain SC 5314, ATCC collection) was grown in 10 ml Sabouraud Medium (bioMérieux, France) for 24 hours. 300 µl of this culture were added to 3 ml of fresh medium and placed on the ATR crystal at room temperature (21°C) and a biofilm was allowed to develop. Biofilm growth was monitored by ATR-FTIR during 19 hours and a total of 58 spectra were collected at intervals of 20 minutes throughout the growth period. 64 scans were coadded for obtaining each final spectrum. Then a toxin was added, and 64 spectra were collected during decay over a 20 hours period.

2.3 Experiment D: Salt solutions at different temperatures (test set D)

Different salt solutions ($NaClO_4$, $MgSO_4$, NaCl) were prepared at 0.05, 0.2, 0.5 and 1 M concentrations in Milli-Q-water. These were measured at temperatures of 15, 22 and 29°C on a Bruker Equinox 55 spectrometer as in Experiment A and B.

2.4 Experiment E: Gluten measured on a different instrument (test set E)

Different salt ($MgSO_4$, NaCl) solutions were prepared at 0.1, 0.2, 0.5 and 1.0 M concentrations in Milli-Q-water. 25 ml solution was each added to an aliquot of 10 mg gluten powder, and the samples were mixed gently until homogeneity. After ~4 hours at room temperature, the samples were centrifuged for 15 min at

~340 g, and excess solution was discarded. Slices of gluten were measured by ATR-FTIR. The spectra from 4000 cm^{-1} to 748 cm^{-1} were recorded on a Bomen FTIR spectrometer equipped with a horizontal ATR-crystal (ZnSe), at a resolution of 4 cm^{-1} and with coaddition of 128 scans. The data interval was 1.93 cm^{-1}.

2.5 Gas modelling and smoothing

Gas modelling: All spectra were corrected for irrelevant water vapour and CO_2 absorption by the method, software and parameters presented by Bruun et al (2006) [9]. Concentrations of water vapour and CO_2 in the spectra is in principle estimated by least squares regression, based on the model of the absorbance spectrum $\mathbf{z}_{0,\mathrm{i}}$ of sample *i* (containing wavenumber channels *k*=1,2,…1764) by previously estimated [9] gas components with absorbance in two wavenumber regions for water vapour and two regions for CO_2, as shown in eq. 1:

$$\mathbf{z}_{0,i} = \mathbf{c}_{Gas,i}\mathbf{K}'_{Gas} + \mathbf{d}_i + \mathbf{e}_i \qquad (1)$$

$$= c_{H_2OVap\_,1,i}\mathbf{K}_{H_2OVap\_,1} + \mathbf{c}_{H_2OVap\_,1,i}\mathbf{K}_{H_2OVap\_,1} + \mathbf{c}_{H_2OVap\_,2,i}\mathbf{K}_{H_2OVap\_,2} + \mathbf{c}_{CO_2\_,1,i}\mathbf{K}_{CO_2\_,1} + c_{CO_2\_,2i}\mathbf{K}_{CO_2\_,2} + \mathbf{d}_i + \mathbf{e}_i$$

$$= \sum_{j=1|}^{JH_2OVap\_,1} c_{H_2OVap\_,1,ij}\mathbf{k}_{H_2OVap\_,1,i} + \sum_{j=1|}^{JH_2OVap\_2} c_{H_2OVap\_2,ij}\mathbf{k}_{H_2OVap\_2,j} + \sum_{j=1|}^{JCO_2\_1} c_{CO_2\_1,ij}\mathbf{k}_{CO_2\_1,j} + \sum_{j=1|}^{JCO_2\_2} c_{CO_2\_2,ij}\mathbf{k}_{CO_2\_2,j} + \mathbf{d}_i + \mathbf{e}_i$$

where $\mathbf{c}_{Gas,i}$ =[$c_{i,j\ ,\ \mathrm{j}=1,2,\ldots}$] represents the concentration or "score" of the gas elements with spectra $\mathbf{K}_{Gas}$ =[$\mathbf{k}_{j\ ,\ \mathrm{j}=1,2,\ldots}$] (water vapour or CO_2, each on two wavenumber regions), $\mathbf{d}_i$ represents the "interesting" chemical and physical absorption effects of the sample, while $\mathbf{e}_i$ is the residual representing measurement error. The reason for allowing two independent model elements for water vapour and two for CO_2 is that local instrument artifacts or vapour-like sample absorbance effects may interfere differently with the gas concentration estimation at different wavenumber regions, so this provides some robustness and flexibility to the gas modelling. Three component spectra were used for the water vapour high range (H_2OVap_1, 4000-3300 cm^{-1}), three component spectra for the water vapour low range (H_2OVap_2, 2200-1200 cm^{-1}), three component spectra for CO_2 absorption in its high range (CO_2_1, 2450-2200 cm^{-1}) and two for CO_2 absorption in its low range (CO_2_2, 750-600 cm^{-1}), in total 11 gas model component spectra.

Even though the gas model spectra $\mathbf{K}_{Gas}$ have been estimated previously, it is risky to estimate their concentrations $c_{i,j}$ in a sample directly from eq.1, because the "interesting" chemical and physical absorption information $\mathbf{d}_i$ is usually unknown at this stage of the pre-processing. If these unknown effects are large, and ignored in the estimation of the gas scores, then they may create large alias errors in the gas score estimation. Therefore the gas scores estimation is based only on the high-frequency part of the spectra, where the "non-gas" sample constituents and other phenomena that constitute $\mathbf{d}_i$ usually have smoother features than the gas elements. (This assumption may not be correct in all parts of the spectrum, especially for protein absorptions in the second water vapour region, see below). Each input spectrum $\mathbf{z}_{0,\mathrm{i}}$ was thus transformed into estimated negative second derivatives (convolution by $\mathbf{u}_1$=[-1 2 -1]), mean centred (to remove low-pass information) and regressed (full rank) on the correspondingly transformed gas model spectra $\mathbf{K}_{\mathrm{Gas}}$. The regression coefficients were taken as estimated gas component scores $\mathbf{c}_{\mathrm{Gas,i}}$. The gas estimates were then subtracted by:

$$\mathbf{z}_{1,\mathrm{i}} = \mathbf{z}_{0,\mathrm{i}} - \mathbf{c}_{\mathrm{Gas,i}}\,\mathbf{K}_{\mathrm{Gas}}' \qquad (2)$$

Smoothing: While most of the water vapour and CO_2 contributions were removed from input spectra $\mathbf{z}_{0,\mathrm{i}}$, some minor gas contributions were evident in $\mathbf{z}_{1,\mathrm{i}}$. To ensure that the subsequent water model spectra were devoid of incidental gas contributions, the spectra $\mathbf{z}_{1,\mathrm{i}}$ were slightly smoothed (convolution with $\mathbf{u}_2$= [1 1 1]). Finally, the remaining CO_2 contribution at 2450-2200 cm^{-1} was considered small enough that the spectrum could be replaced by a straight baseline in this limited region:

$\mathbf{z}_{2,\mathrm{i}}$ =conv($\mathbf{z}_{1,\mathrm{i}}$, $\mathbf{u}_2$), with a straight line between 2450 and 2200 cm^{-1} (3)

2.6 Model of water and its temperature variations

Variation patterns due to water temperature: Since general physical baseline-changes with water temperature were considered irrelevant at this stage, a simple baseline-correction was used at this stage: Each of the gas-corrected and smoothed spectra $\mathbf{Z}_2$ =[-$\mathbf{z}_{2,\mathrm{i}}$, i=1,2,...,39] in Experiment A-1 was base-line corrected by subtraction of the absorbance at an apparently uninformative position (2625 cm^{-1}= channel # n=714).

$$\mathbf{z}_{3,\mathrm{i}} = \mathbf{z}_{2,\mathrm{i}} - z_{2,\mathrm{i,n}} \qquad (4)$$

To reduce the impact of the possible instrument drift during the course of Experiment A-1, temporal difference spectra were computed (i.e. estimated first derivative over time):

$$\mathbf{d}_{\mathrm{i}} = \mathbf{z}_{3,\mathrm{i}} - \mathbf{z}_{3,\mathrm{i-1}} \qquad (5a)$$

This differentiation was done within each of the three short time series in the experiment yielding matrix $\mathbf{D}_{\mathrm{m}}$=[$\mathbf{d}_{\mathrm{i}}$,i=1,2,...], m=1,2,3

A weighted singular value decomposition (svd) of the matrix of difference spectra $\mathbf{D}_{\mathrm{m}}$ was performed within each of the three time series, and the two first PCs were saved:

$$\mathbf{D}_{\mathrm{m}}\mathbf{W} = \mathbf{u}_1 s_{11}\mathbf{v}_{1,\,\mathrm{m}}' + \mathbf{u}_2 s_{22}\mathbf{v}_{'2,\,\mathrm{m}}' + \mathbf{E}_{\mathrm{m}} \qquad (5b)$$

where weights $\mathbf{W}$=diag($\mathbf{w}$) is 1 in the 4000-810 cm^{-1} range but 0.01 for 810-600 cm^{-1} because of apparent instrumental problems at this end of the spectrum. Water temperature spectra were defined as $\mathbf{k}_{\mathrm{watertemp,1,\,m}} = \mathbf{v}_{1,\,\mathrm{m}}'\mathbf{W}^{-1}$ and $\mathbf{k}_{\mathrm{watertemp,2,\,m}} = \mathbf{v}_{2,\,\mathrm{m}}'\mathbf{W}^{-1}$.

Such svd modelling was also done for the difference spectra for Experiment A-1 combined,

$$\mathbf{D}=[\mathbf{D}_{\mathrm{m}}, m=1,2,3] = \mathbf{u}_1 s_{11}\mathbf{v}_1' + \mathbf{u}_2 s_{22}\mathbf{v}_2' + \mathbf{E} \qquad (5c)$$

and water temperature effect component spectra, $\mathbf{k}_{\text{watertemp},1} = \mathbf{v}_1'\mathbf{W}^{-1}$ and $\mathbf{k}_{\text{watertemp},2} = \mathbf{v}_2'\mathbf{W}^{-1}$.

Water reference spectrum at room temperature: Although the exact temperature of the water in the ATR cell was not measured during the temperature-equilibrations in Experiment A, the use of a preliminary temperature calibration model (not shown here) indicated that 8 of the spectra in Experiment A-1 were obtained at 20 +/-2 °C. Their spectra $\mathbf{z}_{3,\text{i}}$ were saved, and their mean was used as reference spectrum **m** for subsequent EMSC modelling.

2.7 EMSC model

The EMSC model as described in Martens et al. (2003) [7] of each gas-corrected absorbance spectrum $\mathbf{z}_\text{i}$ may be written in terms of the "true" chemically based absorbance of the sample, $\mathbf{z}_{\text{i,chem}}$, modified by various physical effects such as an additive polynomial baseline (due to instrument variations or to optical properties of the sample) and a multiplicative signal scaling (e.g. changing optical path length due to changes in refractive index of the sample):

$$\mathbf{z}_\text{i} \approx a_\text{i} + d_\text{i}\boldsymbol{\eta} + e_\text{i}\boldsymbol{\eta}^2 + b_\text{i}\mathbf{z}_{\text{i,chem}} \qquad (6a)$$

where $\boldsymbol{\eta} = [\eta_k, k=1,2,...,K]$ represents the wavenumber range ν defined as –1 to +1 over the channels. The baseline of sample # *i* is thus represented by coefficients a_i, d_i and e_i, and the path length by coefficient b_i. The "true" chemical signal, assumed to reflect the absorbances by various chemical species in the sample, may be written

$$\mathbf{z}_{\text{i,chem}} = c_{\text{i1}}\mathbf{k}_1' + ... + c_{\text{ij}}\mathbf{k}_\text{j}' + ... + c_{\text{iJ}}\mathbf{k}_\text{J}' \qquad (6b)$$

where $c_{\text{ij}}\mathbf{k}_\text{j}'$ represents the concentration and spectrum of chemical constituent # j (e.g. various species of water, salts etc.). This can be equivalently written as

$$\mathbf{z}_{\text{i,chem}} = \mathbf{m} + \Delta c_{\text{i1}}\mathbf{k}_1' + \Delta c_{\text{i2}}\mathbf{k}_2' + ... + \Delta c_{\text{iJ}}\mathbf{k}_\text{J}' \qquad (6c)$$

where **m** is the reference spectrum and Δc_{ij} represents the difference in constituent # *j*'s concentration between the sample *i* and reference **m**. In the pure water samples, the constituents are expected to involve various water structures, the variations between which are described in the obtained water temperature effect component spectra, so an equivalent, simplified bi-linear model may be written:

$$\mathbf{z}_{\text{i,chem}} = \mathbf{m} + \Delta c_{\text{watertemp,i,1}}\,\mathbf{k}_{\text{watertemp,1}}' + \Delta c_{\text{watertemp,i,2}}\,\mathbf{k}_{\text{watertemp,2}}' \qquad (6d)$$

Combining eq. (6a) and (6c) yields the linear model

$$\mathbf{z}_\text{i} = a_\text{i}\mathbf{1} + d_\text{i}\boldsymbol{\eta} + e_\text{i}\boldsymbol{\eta}^2 + b_\text{i}\mathbf{m} + h_{\text{i1}}\,\mathbf{k}_1' + h_{\text{i2}}\,\mathbf{k}_2' + ... + h_{\text{iJ}}\mathbf{k}_\text{J}' + \boldsymbol{\varepsilon}_\text{i} \qquad (6e)$$

where vector **1** =[1,1,1,....,1] is introduced for matrix formality, and $h_{\text{ij}} = b_\text{i}\cdot\Delta c_{\text{ij}}$. Vector $\boldsymbol{\varepsilon}_\text{i}$ is added to represent the residual spectrum of sample *i*, containing random measurement noise and possible unmodelled spectral structures. The parameters in eq. (6c) are estimate by weighted least squares regression; the diagonal weight matrix **W** described above was used in order to down-weight the problematic lower part of the wavenumber range.

Once estimated, the model parameters are used for EMSC correction. One alternative is to retain the estimated chemical information, $h_{i1}\,\mathbf{k}_1' + h_{i2}\,\mathbf{k}_2' + ... + h_{iJ}\mathbf{k}_J'$ in the spectra:

$$\mathbf{z}_{i,\mathrm{corrected}} = (\mathbf{z}_i - a_i - d_i\,\eta - e_i\,\eta^2)/b_i \qquad (7a)$$

Alternatively,

$$\mathbf{z}_{i,\mathrm{corrected}} = (\mathbf{z}_i - a_i - d_i\,\eta - e_i\,\eta^2 - (h_{i1}\,\mathbf{k}_1' + h_{i2}\,\mathbf{k}_2' + ... + h_{iJ}\mathbf{k}_J'))/b_i \qquad (7b)$$

removes also the estimated chemical variability around the reference spectrum **m**, while

$$\mathbf{z}_{i,\mathrm{corrected}} = (\mathbf{z}_i - a_i - d_i\,\eta - e_i\,\eta^2 - (\mathbf{m} + h_{i1}\,\mathbf{k}_1' + h_{i2}\,\mathbf{k}_2' + ... + h_{iJ}\mathbf{k}_J'))/b_i \qquad (7c)$$

removes the reference spectrum itself as well.

Moreover, once the parameters have been estimated, Δc_{ij} may be estimated as:

$$\Delta c_{ij} = h_{ij}/b_i,\ j=1,2,..,J \qquad (7d)$$

Specifics of the EMSC modelling will be given for each dataset in the Results and Discussion section.

2.8 Temperature prediction from EMSC scores (Experiment B)

The EMSC water temperature scores $\Delta c_{\mathrm{watertemp},i,1}$ and $\Delta c_{\mathrm{watertemp},i,2}$ obtained by EMSC of pre-processed spectra $\mathbf{Z}_3$ (eq. 2-4) followed by eq. (7d) were used in a polynomial multivariate calibration model

$$\mathbf{y} = \mathbf{Xb} + b_0 + \mathbf{f} \qquad (8)$$

where $\mathbf{y}$ =temperature$_i$ and $\mathbf{X} = [\Delta c_{\mathrm{watertemp},i,1}, \Delta c_{\mathrm{watertemp},i,1}^2, \Delta c_{\mathrm{watertemp},i,2}]$ and **f** is residual.

The three regression coefficients in **b** and b_0 were estimated by ordinary least squares regression, using the i=1,2,..,84 spectra from Experiment B, which had been measured at known temperatures between 8 and 63°C. The temperature calibration model was subsequently used for predicting temperature in new samples from their pre-processed spectra $\mathbf{z}_i$, i=1,2,... .

2.9 Salt-effect estimation (Experiment D)

The spectra corrected for water vapour and CO_2 were used for determination of the salt-effect spectra. The spectra of each salt at each temperature were EMSC corrected individually (only subtraction of channel number) with the corresponding pure water spectrum used as reference spectrum. The weights for the range 1300 cm^{-1}-600 cm^{-1} were set to zero in order to down-weight the sulphate and perchlorate absorptions at ~1093 cm^{-1}. After EMSC correction, the difference spectra were calculated by subtraction of the spectrum of pure water at the same temperature. The difference spectra seemed not influenced much by the temperature, and only by salt type and concentration, as well as some day-to-day variance. For each salt type, a PLS regression was made, for which X=salt concentrationtemperature, day and Y=difference spectra. Thereby the B-coefficient for all Y-variables showed the spectral characteristics of each salt.

3. RESULTS AND DISCUSSION

3.1 Experiment A: Estimation of temperature induced changes in water based on temperature-equilibrating samples with unknown temperature.

The temperature-induced variations in the ATR-FTIR spectrum of water was estimated on the basis of 39 spectra of pure water at various unknown temperatures (in the range 10-35^{o}C). These were obtained in three temperature-scanning series (calibration set A-1). The measured absorbances $\mathbf{Z}_0$ are shown in Figure 1a,b before and after mean-centring. The major band at 3300 cm^{-1} results from the OH stretching modes ($\nu_{1,3}$) of liquid water, whereas the smaller bands at ~1640 cm^{-1} and 800 cm^{-1} result from the HOH bending (ν_2) and the librational (ν_L) modes of liquid water, respectively. A minor band at 2125 cm^{-1} is ascribed to the combination of ν_2 and ν_L. The water bands are affected in different ways by the temperature.

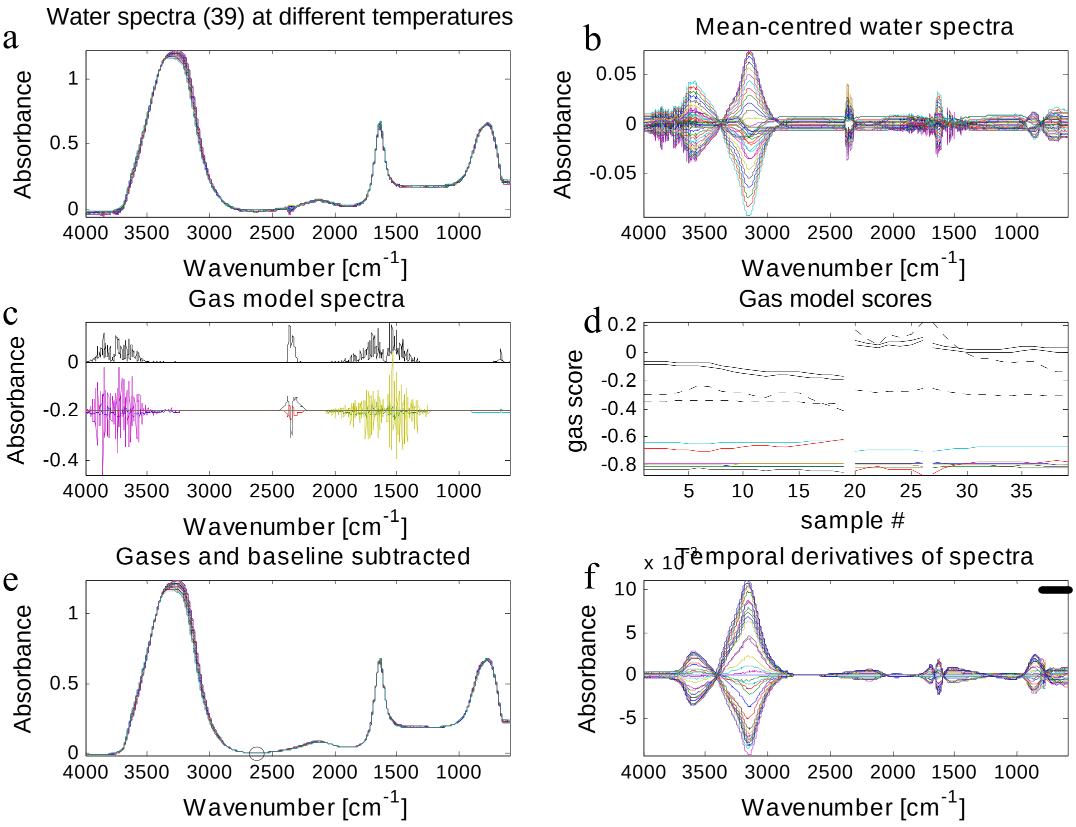


Figure 1. Estimation of water temperature model spectra from Experiment A, Part 1: 39 consecutive ATR-FTIR spectra of pure water from three temperature-drift time series. Top: Measured absorbance spectra $\mathbf{Z_0}$, before (a) and after (b) mean-centring. Middle: Model-based correction for water vapour and CO_2 in terms of (c) model

component spectra vs. wavenumber and (d) their estimated component scores vs. sample # (i.e. time). For visual clarity in c) the 2nd and 3rd component spectra are down-shifted by -0.2 compared to the 1st component spectrum. In d) two upper solid curves and the two upper dashed curves represent the 1st components of water and of CO_2 respectively; the 2nd and 3rd components have been down-shifted by 0.5 for visual clarity.
Bottom: Pre-processed absorbance Z3, after gas - and baseline correction, before (e) and after (f) mean-centring.

In addition, high-frequency signatures characteristic of water vapour are evident in Figure 1(b) the at 4000-3300 cm^{-1} and 2200-1200 cm^{-1}. Likewise, variations in the characteristic peak of CO_2 at 2450-2200 cm^{-1} are seen. Figure 1(c) shows the gas model spectra obtained from Bruun et al (2006) [19], in terms of the four average gas component spectra, seen as four non-overlapping non-negative spectral features and the seven additional gas modification spectra (down-shifted by 0.2 for visual clarity). According to Bruun et al (2006) [19] the former are expected to provide concentration estimates for the gases, while the latter should account for unidentified, but systematic effects in the gas concentration modelling. Based on the gas model (eq. (1)), the gas concentration scores were estimated from spectral second derivatives and they are shown in Figure 1(d) for the three time series. The figure shows that the main water vapour concentration estimates from the two wavenumber regions (solid curves) follow each other very well, while the two main CO_2 wavenumber regions give rather different CO_2 concentration estimates (dashed lines). The estimated gas contributions were subtracted by eq. (2), yielding spectra $\mathbf{Z}_1$, followed by mild local smoothing etc (eq.(3)), yielding spectra $\mathbf{Z}_2$. Remaining undesired baseline variations evident in Figure 1(b) were removed by eq. (4), yielding the spectra $\mathbf{Z}_3$ shown in Figure 1 (e) and (f) before and after mean-centring.

For estimation of the temperature-induced spectral changes, the $\mathbf{Z}_3$ spectra of the three time series in Experiment A-1 were used for calculation of temporal difference spectra $\mathbf{D}_\mathrm{m}$, m=1,2,3 (eq. (5a)) and these were submitted to weighted svd, both separately (eq. (5b)) and jointly (eq. (5c)). Two systematic principal components were found each time, the first singular value being about 22 times greater than the second one each time. Together they explain most of the variance in **D.** Figure 2(a) and (b) show the two components $\mathbf{k}_\mathrm{watertemp,1}$ and $\mathbf{k}_\mathrm{watertemp,2}$, while Figure 2(c) shows the water spectrum **m** at room temperature for graphical comparison. Given the variations in Figure 1(e), the shape of $\mathbf{k}_\mathrm{watertemp,1}$ in Figure 2(a) is not unexpected: A major "shift" in the 3300-2900 cm^{-1}, a clear negative peak resembling a second derivative of the mean spectrum around 1600 cm^{-1}, a positive peak around 800 cm^{-1} plus a weak pattern in the 2500-2000 cm^{-1} range. This reflects that the $\nu_{1,3}$ band changes in intensity and shifts in position when the temperature changes. The ν_L and ν_2+ν_L bands seem to shift in the opposite direction of the $\nu_{1,3}$ band, whereas the postion of the ν_2 band is not affected much. However, an effect on the peak width is seen for this band. As expected, the baseline of $\mathbf{k}_\mathrm{watertemp,1}$ is close to zero both at 4000 and at 2625 cm^{-1}. However, it is below zero at the lowest end of the spectrum, i.e. in the extreme range 810-600 cm^{-1} down-weighted because of apparent instrument problems. The independent replicate estimates of $\mathbf{k}_\mathrm{watertemp,1,m}$, m=1,2,3 (dashed curves) are so similar that they can hardly be distinguished from the over-all estimate $\mathbf{k}_\mathrm{watertemp,1}$(solid curve), which will be used in EMSC models for all the experiments in the paper.

Even for the small second water temperature effect $\mathbf{k}_{\mathrm{watertemp},2}$ (Figure 2(b)), the reproducibility is reasonable except in the 810-600 cm^{-1} range. Therefore, this component is considered reasonably reliable. It will be included in subsequent EMSC models of pure water, but since it appears to reflect both "chemical" peak changes and "physical" baseline changes and its effect is small, it will not be used in modelling the more complex sample types to maintain interpretability of the resulting EMSC-treated spectra.

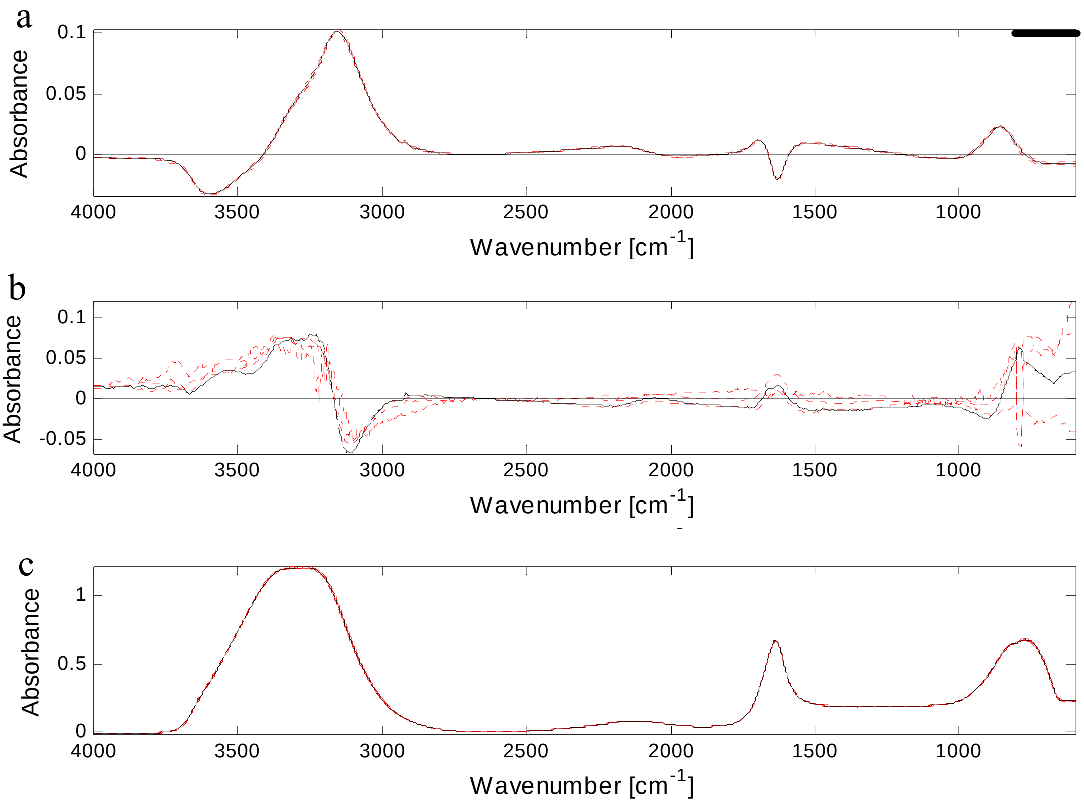


Figure 2. Estimation of water temperature model spectra (continued). First component $\mathrm{k}_{\mathrm{watertemp},1}$(a), second component $\mathrm{k}_{\mathrm{watertemp},2}$ (b) and mean water spectrum m at 20°C (c), obtained from Z2 in Experiment A-1. Solid curves: final model estimate; dotted curves: Replicate estimates from (a, b) three independent time series and (c) eight different samples in the range 20±2°C.

The fact that one PC, $\mathbf{k}_{\mathrm{watertemp},1}$ described most of the temperature effect on pure water is in agreement with several infrared spectroscopic studies of water at different temperatures (these also apply the overtone and combination bands in the near-infrared (NIR) region). A PCA analysis of the variation in the OH-str. overtone band at 1440 nm with temperature (6-80°C) has shown more than 99 % explanation of the spectral variation in the first PC; the associated scores and loading vectors were interpreted as reflecting a continuous inter-conversion with temperature between two water species with weak and strong hydrogen bonds, respectively [9]. This interpretation is in agreement with a mixture model describing water as being composed of a limited number of water species (different ring structures or cluster types) that coexist in a temperature-and pressure-dependent equilibrium [20]. On the other hand, the alternative continuum model implies that hydrogen bonds are equally distributed in the water sample, and that there is a continuous weakening of the hydrogen bonds with increasing temperature [21]. The presence of nearly-isosbestic points in the infrared spectra of water at different temperatures has usually been taken as support of the mixture model. However, both models need to be taken into account in order to explain the many anomalous properties of water [22].

A two-state model of water, involving weakly hydrogen bonded and strongly hydrogen bonded groups of water molecules, is supported from water dynamic studies, which have shown groups with fast rotation (weakly bonded) and groups with slower rotation (strongly bonded) [23]. NIR and MIR studies by Šašić et al (2002) [10] and Libnau et al (1994) [8] have also lead to the conclusion that water can be seen as a closed two-component system with two water species (in a strongly and loosely hydrogen bonded state, respectively) existing in equilibrium.

Nevertheless, some deviations from the two-state model have been described in the spectroscopic studies. In the study by Segtnan et al (2001) [9], additional variation in the OH-str. overtone band occurred from 6 to 26°C and was described in a minor PC3 variation (accounting for less than 1% of the variation) [9]. Disruption of the two-component model at temperatures below 20°C has been suggested from other studies as well, and e.g. D'Arrigo et al (1981) [24] proposed the existence of an additional tetrahedrally bonded form of water below 20°C. Furthermore, Segtnan et al (2001) [9] and Czarnick-Matusewitcz et al (2005) [25] both observed a minor variation over the whole temperature range, for which scores showed a parbolic shape with temperature. Band broadening and change in refractive index with temperature have been suggested as possible causes to these deviations from the two-component model.

For simplicity, we interpret this variation pattern $\mathbf{k}_{\text{watertemp},1}$ as the difference spectrum between the two major species of water ("free" and "bound" water) present at different ratios in pure water at different temperatures. The interpretation of $\mathbf{k}_{\text{watertemp},2}$ is not yet clear. It appears to be a combination of the difference between minor water species plus some wavenumber-dependent baseline effect, e.g. due to an effect of temperature-induced changes in refractive index on the ATR signal. Its features below 1000 cm^{-1} do not seem to be reproducible.

Figure 3 shows the EMSC modelling of the spectra from Experiment A. The gas-and baseline corrected spectra $\mathbf{Z}_3$ (eq. (2) ,(3) and (4)), EMSC corrected with water temperature effects subtracted (eq. (7c)) are shown in Figure 3(a). The 39 spectra appear virtually indistinguishable. After mean-centring some small spectral residuals can be seen (Figure 3(b)). However, these have rather discontinuous nature and appear to reflect changes between the three time series in the FTIR instrument rather than effects related to temperature. Figure 3(c) shows the EMSC parameters [a_i, d_i, e_i , h_{i1}, h_{i2} and b_i] as functions of time within the three time series. Baseline parameters a_i, d_i, e_i and scaling b_i lie quite constant around 0 or 1, respectively, while h_{i1} and some degree h_{i2} change with time. The re-scaled water temperature parameters Δc_{ij} (eq. (7d)) are shown in Figure 3(d). The dominant effect is the first water temperature score Δc_{i1} (solid curve), while the second water temperature score, Δc_{i2} (dashed curve) had to be scaled by a factor of 23 to show comparable variation. The figure shows that the temporal dynamics of the interesting phenomena in the liquid water samples resemble but are distinct from those of the irrelevant atmospheric absorptions in the instrument (Figure 1(d)). This means that without the initial gas correction, the two causally distinct temporal processes would have been confounded.

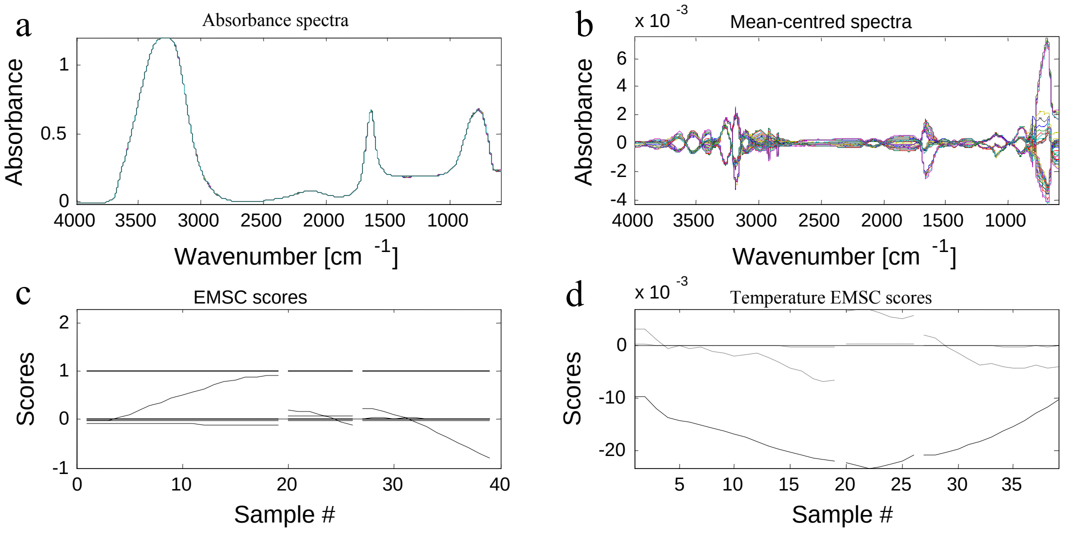


Figure 3. Estimation of water temperature model spectra (continued). Top: 39 EMSC-treated spectra, i.e. after subtraction of the two estimated water temperature component effects from Z_3 (Figure 1(e)), a) before and b) after mean-centring. Bottom: EMSC model scores in the three time series as functions of sample # (i.e. time); c) all EMSC parameters (eq. (6e)), d) the rescaled scores (eq. (7d)) for water temperature components 1 (solid) and 2 (scaled by a factor of 23, dashed).

3.2 Experiment B: Relating the EMSC model parameters to known temperature

In order to find the correlation between the EMSC parameters and the known temperatures, the temperature model was applied on data set B, which contains ATR-FTIR spectra of pure water at different known temperatures (8-63˚C), obtained over a period of two months. The input spectra $\mathbf{Z}_0 = \mathbf{z}_{i,0}, i=1,2,...,84$ are shown in Figure 4 (a) and (b) before and after mean centring. These were gas- and baseline-corrected by eq. (2), (3) and (4), resulting in spectra $\mathbf{Z}_3$ in Figure 4 (c) and (d). The figure shows that most of the characteristic, irrelevant effects of water vapour and CO_2 have been eliminated. However, compared to Figure 1(e), these pure-water samples with known temperature display considerable contributions of other effects, e.g. apparent baseline- and scaling problems. This explains why Experiment B could not be used for estimating the temperature-change spectra of water, as originally planned. The reason for the problems could be that Experiment B was performed over a much longer period of time, in which the FTIR instrument and ATR unit were used repeatedly also for several other purposes.

However, if the problems in Experiment B were of “physical” nature (and not e.g. chemical impurities on the ATR surface), EMSC should be able to correct for at least some of it. Therefore, the spectra in Figure 4(c) were submitted to EMSC modelling with both “physical” baseline- and scaling elements and with the the two water temperature components as “chemical” elements. The score estimation was based on eq. (6d) and (6e) and the correction on eq. (7b). The resulting $\mathbf{Z}_4$= [$\mathbf{z}_{i,\text{corrected}}$,$i$=1,2,...,84] are shown in Figure 4 (e) and (f) before and after mean centring. Again, the temperature-corrected spectra appear virtually indistinguishable except for the lowest, down-weighted wavenumber region. This shows that indeed the EMSC removed many of the baseline- and scaling problems that arose in the instrument over the two-month measurement period. However, the residuals in Experiment B are somewhat larger than those in Experiment A-1, even outside the difficult, lowest wavenumber region.

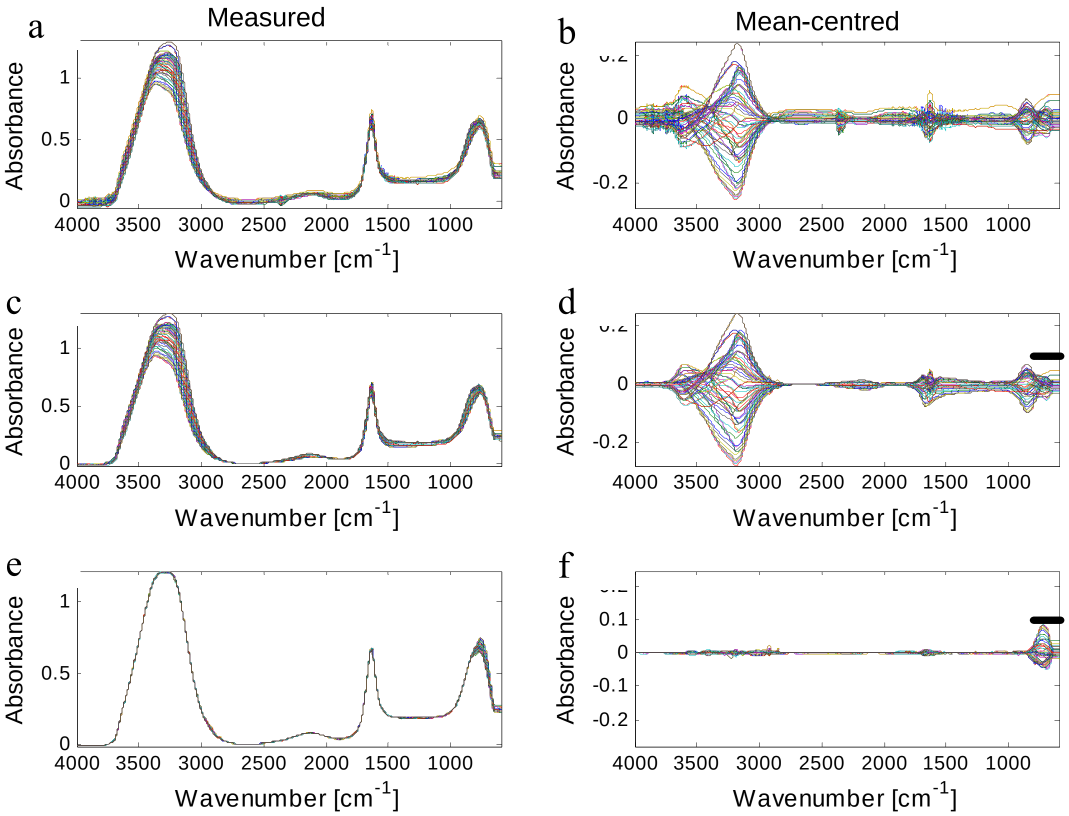


Figure 4. Calibration for water temperature in Experiment B. (84 different ATR-FTIR spectra of pure water at known temperatures, recorded at different dates). Top: Measured absorbance spectra $\mathbf{Z}_0$, before (a) and after (b) mean-centring. Middle: Pre-processed absorbance $\mathbf{Z}_3$, after gas - and baseline correction of $\mathbf{Z}_0$, before (c) and after (d) mean-centring. Bottom: Pre-processed absorbance $\mathbf{Z}_4$, after EMSC correction of $\mathbf{Z}_3$ with water temperature effects subtracted before (e) and after (f) mean-centring. The solid line segments in c) and e) show the wavenumber range down-weighted by a factor of 0.01 due to irrelevant instrument variations.

The parameters $\Delta c_{\text{watertemp},i,1}$ and $\Delta c_{\text{ watertemp},i,2}$ from the EMSC (eq. (7d)) of Experiment B were used for calibrating for temperature. Regression coefficients (eq. (8)) were estimated by least squares. Cross-validated PLS regression showed the full rank solution to have best predictive ability, and this yielded the prediction model:

Temperature = $-11.9\ \Delta c_{\mathrm{watertemp,i,1}} + 0.99\ \Delta c_{\mathrm{watertemp,i,1}}^{2} + 6.8\ \Delta c_{\mathrm{watertemp,i,2}} + 21.1$ (8)

Figure 5 plots the predicted vs. the measured temperature in these 84 samples of Experiment B. The predictive ability is high (cross-validated r = 0.996). Not entirely unexpected, given the serious instrument variations, some unexplained variation is evident particularly at temperatures outside the range 10- 35˚C of Experiment A-1, within which the water temperature spectral components were estimated.

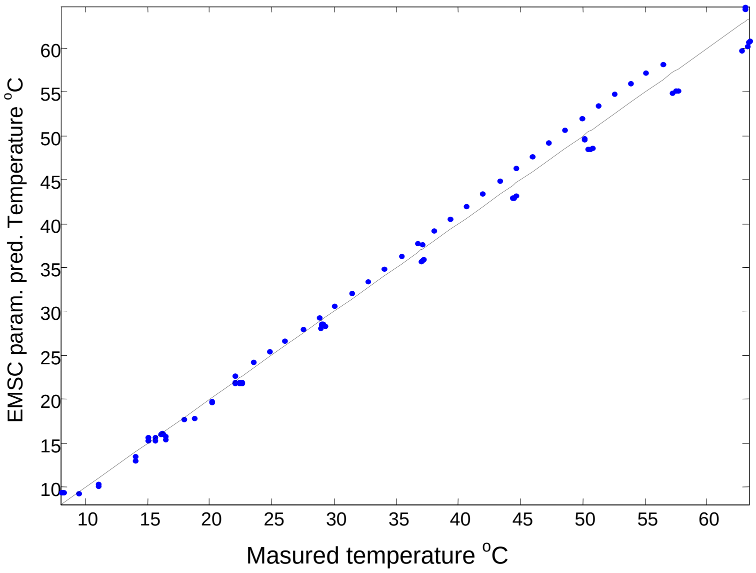


Figure 5. Calibration for water temperature (continued.). Measured temperature (abscissa) vs. predicted temperature (ordinate) obtained by the obtained calibration model (eq. 8) based on EMSC scores in Experiment B. Predictive ability: cross validated r= 0.996.

3.3 Experiment A-2: Testing the EMSC model parameters in independent water samples

Test set A-2 contains ATR-FTIR spectra of pure water obtained in three new temperature-scanning series and is used as an independent test set to evaluate the performance of the EMSC model on ATR-FTIR water spectra in general. Again, the input data in Figure 6(a) and 6(b) show gas contributions as well as large baseline/scaling variations (probably due to instrument drift). After pre-processing by eq. (2) and (3) the spectra $\mathbf{Z}_2$ (Figure 6 (c),(d)) show that the gas contributions have been removed. Since no preliminary baseline correction was applied for these data, the spectra show large baseline variations. However, EMSC is able to remove most of the systematic variations (Figure 6 (e) and (f)), leaving only minor variations in the residuals except for the problematic region below 800 cm^{-1}.

The temperature was predicted for all the samples in Experiments A by (eq. 8). The predicted sample temperatures are plotted in Figure 7 to illustrate the temperature range and the temporal dynamics within the three calibration sets in Experiment A-1 and the three test sets in Experiment A-2. In conclusion, the EMSC model is found to work very well on this kind of ATR-FTIR spectra, measured in pure water at different temperatures with different levels of water vapour and CO_2 contamination and levels of baseline- and scaling problems of the present kind.

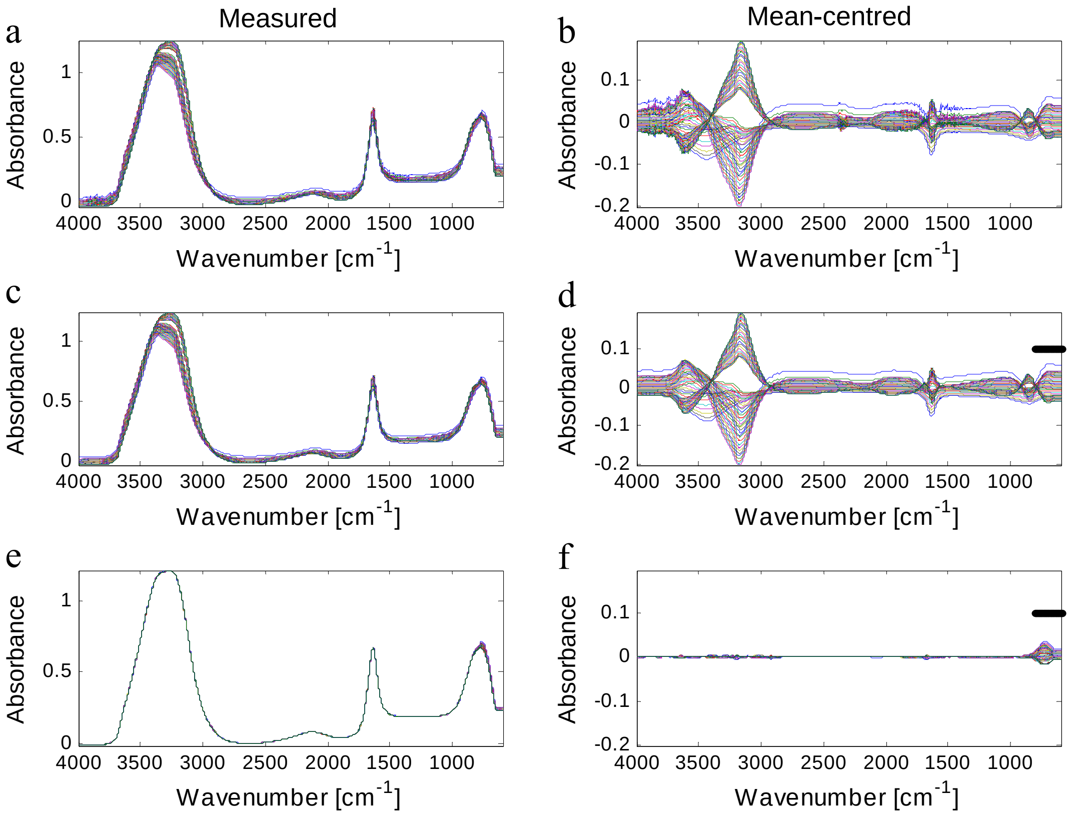


Figure 6. Applying water temperature models in Experiment A-2: (65 consecutive ATR-FTIR spectra of pure water from three new temperature-drift time series). Top: Measured absorbance spectra Z_0, before (a) and after (b) mean-centring. Middle: Pre-processed absorbance Z_2, after gas-correction of Z_0, before (c) and after (d) mean-centring. Bottom: Pre-processed absorbance Z_4, after EMSC correction of Z_2, before (e) and after (f) mean-centring.The solid line segments in c) and e) show the wavenumber range down-weighted by a factor of 0.01 due to irrelevant instrument variations.

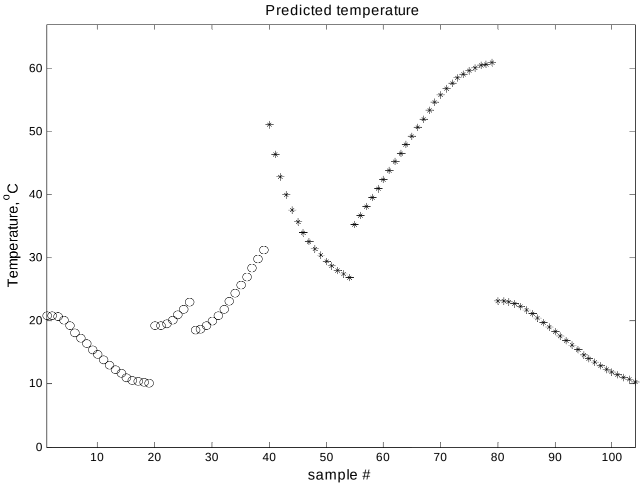


Figure 7. Applying water temperature models (continued). Predicted temperature vs sample # (i.e. time) in the 39 samples from calibration Experiment A-1 (o) and the samples from Experiment A-2 (*).

3.4 Experiment C: Testing the EMSC model parameters in biological samples

Input data: Under the same conditions as for Experiments A and B, the ATR-FTIR absorbance spectra were recorded for pure growth medium (1 spectrum) and of *C. albicans* during growth and decay; 58 and 64 spectra, respectively. The 123 spectra analysed as one consecutive time series. The raw absorbance spectra are shown in Figure 8(a). The general shape of the spectra resembles that of pure water, although some additional signals above 1000 cm^{-1} can be seen. In Figure 8(b) the spectra are shown after subtraction of the spectrum of pure medium. Varying contributions of water vapour and CO$_2$ are seen to dominate the spectra.

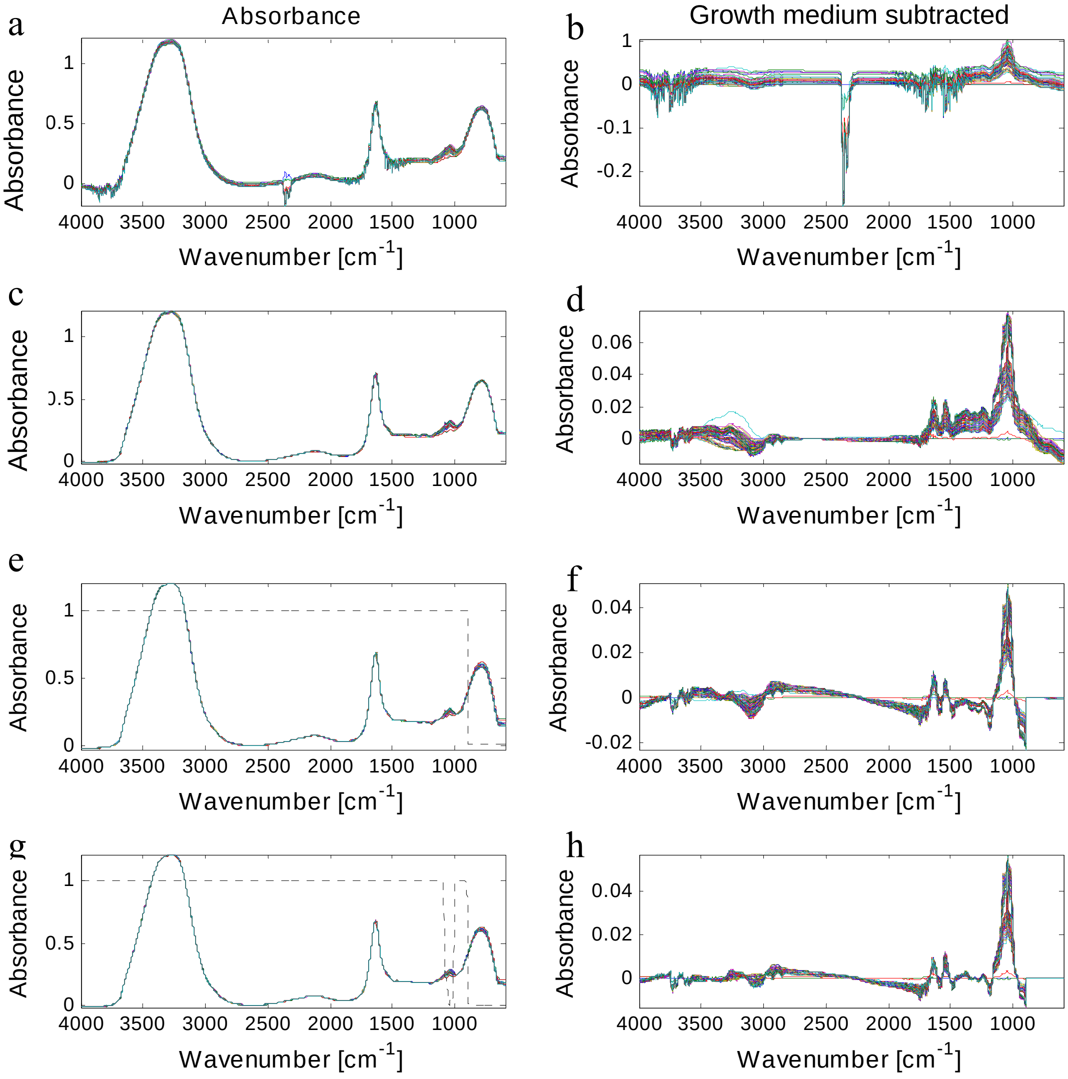


Figure 8. Experiment C: Growth and decay of *C. albicans* before (left) and after (right) subtraction of the pure medium spectrum. a,b) Measured absorbance spectra Z_0. c,d) Pre-processed absorbance Z_1, after gas – and baseline correction of Z_0. e,f) Pre-processed absorbance $Z_{3.1}$, after simple, “physical” EMSC - correction of Z_1. g,h) Pre-processed absorbance $Z_{3.2}$, after reweighted EMSC with subtraction of apparent “water temperature”. The dashed line segments in e) and g) show the wavenumber range down-weighted.

Gas correction: The concentration scores of water vapour and CO_2 components in the wavenumber region above 2000 cm^{-1} were estimated in the second derivative, and used for subtracting these gases in the whole wavenumber region, and remaining variation for the main CO_2 peak was replaced by a straight local baseline, as described by Bruun et al (2006) [19]. The absorbance at 2625 cm^{-1} was taken as a simple baseline estimate and subtracted, all as described above. The obtained gas- and baseline-corrected spectra in Figure 8(c) display some variation patterns in the 2000-1000 cm^{-1} region. After subtraction of the first spectrum Figure 8(d) shows a number of sharp peaks with varying absorbance in this region, where proteins, lipids etc are known to absorb light. A little of this variation may be due to remaining water vapour contributions, since a little waver vapour is also evident above 3500 cm^{-1}. Between 3500 and 3000 cm^{-1}, i.e. inside the main water absorbance peak, systematic variations are also evident. But these are rather smooth, and might thus be due to some physical scaling of the main water peak itself, e.g. caused by changes in the ATR surface and hence in the effective optical path length. Between 3000 and 2000 cm^{-1} a flat but somewhat sloping baseline is evident.

EMSC # 1: The spectra in Figure 8(c) were submitted to a "physical" EMSC for baseline- and scaling correction based on a simplified version of eq. (6e): $\mathbf{z}_i = a_i\mathbf{1} + d_i\eta + e_i\eta^2 + b_i\mathbf{m} + \boldsymbol{\varepsilon}_i$ [7], using water at room temperature (Figure 2 (c)) as reference spectrum **m** and correcting using eq. (7a). The resulting spectra are shown in Figure 8(e); the dotted line shows the weights used in this EMSC parameter estimation. The spectra after medium subtraction (Figure 8(f)) have been scaled by these weights and they show systematic patterns of variation in several wavenumber ranges. The patterns in the 1800-800 cm^{-1} range are now more distinct, compared to Figure 8(d)). Curved baselines are apparent over a wide range, and this may possibly be an artifact from the EMSC "physical" modelling due to an over-simplified baseline model or to estimation alias errors caused by unmodelled chemical absorbance peaks.

The variation pattern inside the main water peak 3500 and 3000 cm^{-1} in Figure 8(f) is still appreciable, although smaller and simpler than in Figure 8(d). It probably reflects water changes in the material (*C. albicans* cells and growth medium) near the ATR surface. But is it due to changing water content or changing water binding?

EMSC # 2: Decker et al. (2005) [26] used a water temperature difference spectrum to improve the classification of *Penicillium camemberti* spectrum by NIR reflectance. The same approach was tried here, based on the rationale that the FTIR difference between free and bound water in pure water might resemble the difference between free and bound water in biological tissue. The water temperature spectrum obtained for pure water was therefore included in a new EMSC modelling of *C. albicans*. Only the first component spectrum (Figure 2(a)) was used, as the second component was considered too complex for the present purpose. Figure 8(g) and 8(h) show the results before and after subtraction of the medium. In order to reduce the impact of the interesting and dominant, but unmodelled spectral variations around 1000 cm^{-1}, additional wavenumber channels have here been down-weighted [7] based on the inverse of the squared mean of the absolute residuals from Figure 8(f) and the EMSC repeated once. The new weights are given by the dashed line in Figure 8(g). Much of the variation in the water peak around 3000 cm^{-1} was thereby eliminated, confirming a similarity between free- and bound water in the two systems. The apparent baseline artifact was also somewhat reduced by this model refinement, leaving very distinct variation signatures in Figure 8(h). This plot reveals the IR fingerprint of *C. albicans* in terms of its changing biomass and –composition and its changing metabolic modification of the surrounding medium.

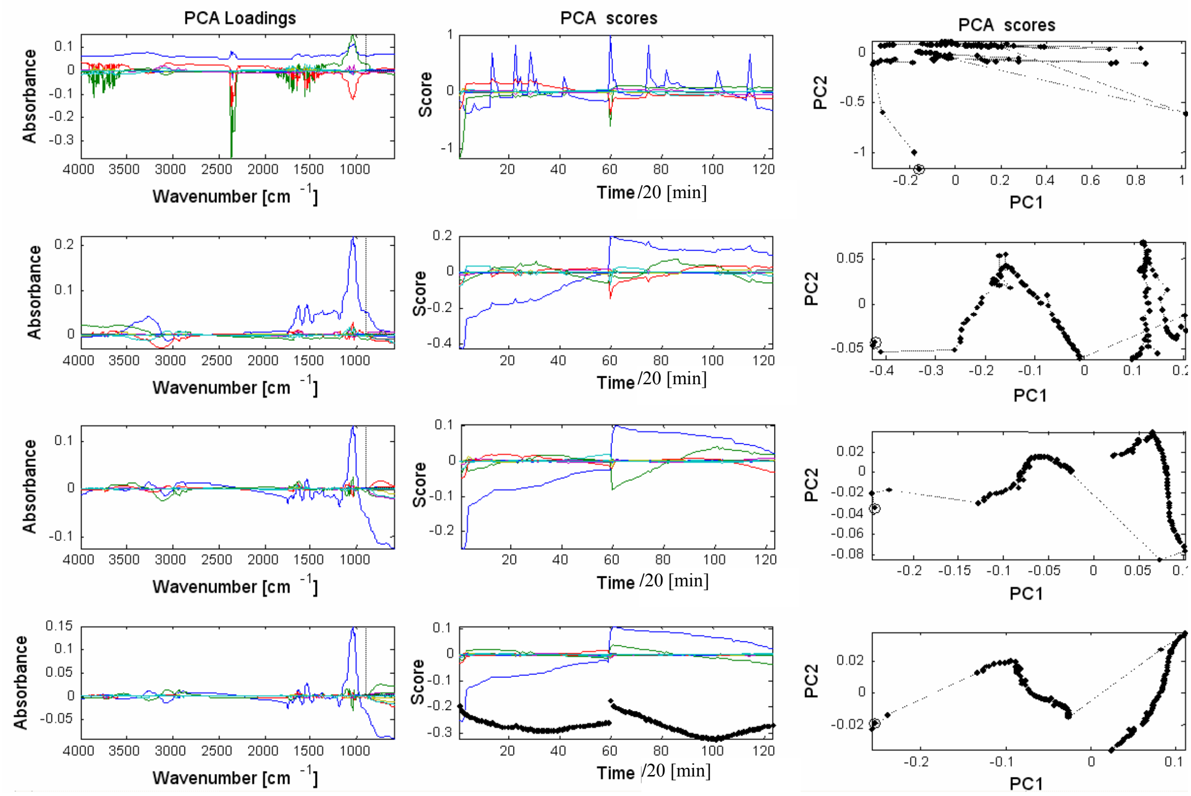


Figure 9. Experiment C, continued. Weighted PCA of the *C. albicans* spectra after subtraction of pure medium (right side of Figure 8). Left: Scaled PC loadings vs wavenumber. The dashed line segments outline the wavenumber range down-weighted by 0.01. Middle: Scaled PC scores vs sample # (i.e. time). The solid curve at the bottom of the last model shows the estimated "free-vs. bound water" from the water-temperature spectrum in Figure 2a).
Right: Scores of PC1 vs. PC2.

Figure 9 compares the main variation patterns remaining after these four different pre-treatments of the *C. albicans* spectra, in terms of PCs from weighted svd of the spectra in Figure 8(b), (d), (f) and (h). The loadings vs wavenumber (left) and scores vs. time (middle) were scaled by the size of the respective singular values. The right-hand side show score vector of PC1 (abscissa) vs. PC2 (ordinate); adjacent points in time are connected; "o" represents the pure medium at the start of the experiment. The rows represent the input spectra, the gas-and baseline-corrected spectra, the spectra after the purely physical EMSC#1 and after the more extended physical and water-temperature EMSC#2. The large jump in the score plot around sample # 64 represents the addition of the toxin to the *C. albicans* on the ATR. The dotted curve in the bottom score plot represent the estimated "water temperature score", assumed to reflect changes in free- vs. bound water. The figure shows that while the input spectra gave a meaningless PCA dominated by CO_2 and water vapour, the three pre-processing methods give increasingly simple and informative results, both in the wavenumber domain and the time domain. The results from the last model appear suitable for biological interpretation and modelling, by e.g. multivariate curve resolution or data base searches of spectra, as well as a multivariate time series input to dynamic mathematical modelling. But that is beyond the scope of the present paper.

3.5 Experiment D: Effect of salts on water

Input data: The spectra in Expriment D were measured over a two-month period. Closer inspection of the input spectra revealed strong day-to-day variation problems of various kinds. Figure 10, left side, shows the difference spectra **Y** of aqueous salt solutions and pure water measured at the same temperature on the same day for various concentrations of $MgSO_4$, $NaClO_4$ and NaCl at 15, 22 and 29°C, respectively. The strong absorbance peak of SO_4^{2-} near 1090 cm^{-1} and of ClO_4^- near 1100 cm^{-1} are clearly evident along with weaker changes in the water peak regions.

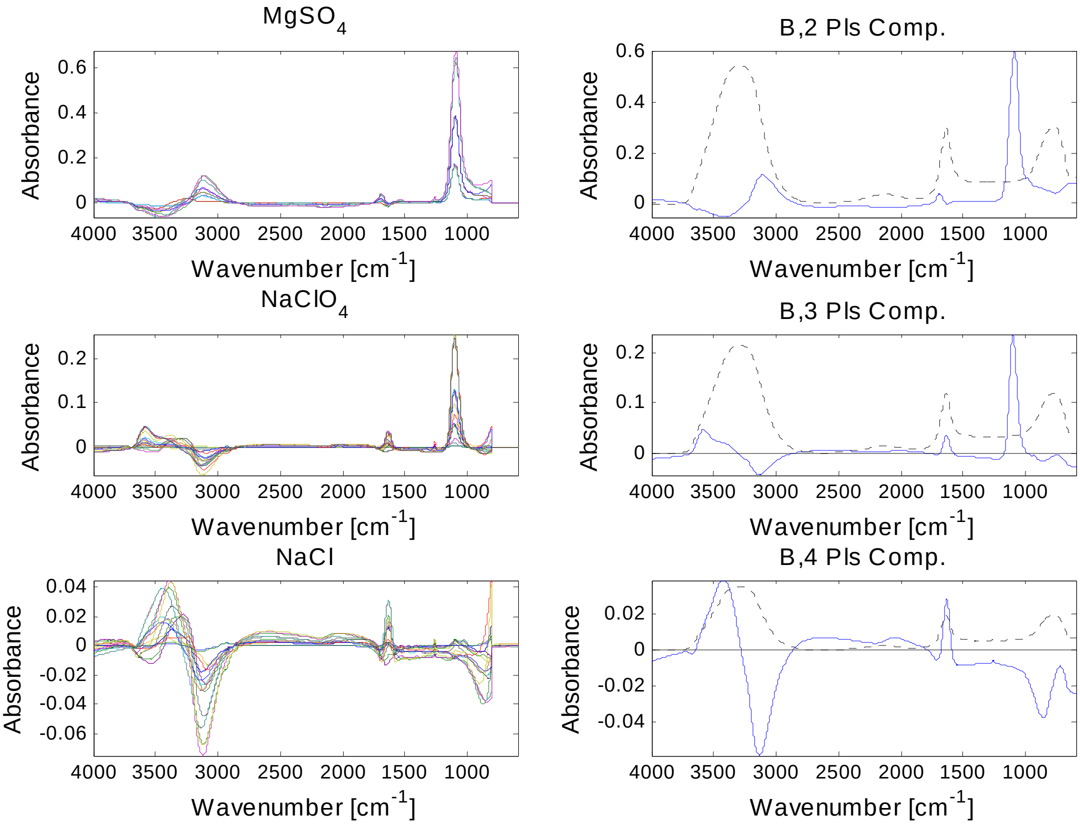


Figure 10. Experiment D: Effects of salts on the IR water spectrum. Top: $MgSO_4$, middle: $NaClO_4$, bottom: NaCl. Left: FTIR spectra of solutions at various salt concentrations, after subtraction of pure water spectrum. Right: Estimated linear effect of salt (solid), estimated by PLSR, compared to water at room temperature (dashed).

Estimated salt effect spectra: The main effect spectrum **k** was for each salt estimated, based on the model:

$$\mathbf{Y} \approx \mathbf{ck}'+\mathbf{tg}'+\mathbf{DH}' = \mathbf{XB} \qquad (9)$$

where vectors **c** =known concentration and **t**=known temperature, and matrix **D**= indicator variables with one column for each day. Coefficient spectra **B**=[**k**, **g**, **H**] were estimated by regressing **Y** on **X**=[**c**, **t**, **D**] by weighted PLS Regression; Y-channels below 893 cm^{-1} were down-weighted by a factor of 0.1 and the X-variable **c** was upweighted by a factor of 100 compared to **t** and **D**. Full leave-one-out cross-validation showed the optimal number pf PLS PCs to be 2,3 and 4 for the three salts, respectively. On the right-hand side of Figure 10 are the main effect spectrum **k** shown for each of the three salts, with the water spectrum **m** (dashed) for visual comparison. The salt*temperature effect estimates appared to be smaller and contaminated by unmodelled instrument problems and are not pursued here.

Each salt effect spectrum reflects both the anion and the cation effect. These have in several studies been found independent of each other, i.e. the cations do not influence the hydration water of the anions and vice versa [Chen, 2004; Fischer, 2001].

The $MgSO_4$ effect spectrum shows an increase at low wavenumbers (3116 cm^{-1}) concomitantly with a decrease at high wavenumbers in the $\nu_{1,3}$ band due to strong interaction of Mg^{2+} with surrounding water molecules. (The sulphate ion may only contribute to a minor negative peak at 3660 cm^{-1}). As expected, the chaotropic $NaClO_4$ effect spectrum shows opposite effects compared to the kosmotropic $MgSO_4$, and an increase at high wavenumbers (3583 cm^{-1}) reflects the formation of weak hydrogen bonds to ClO_4^-. Similarly, the NaCl effect spectrum shows an overall shift to higher wavenumbers, mainly as a result of the weak chaotropic property of Cl^-. However, the effect of NaCl is low, reflecting the low number of water molecules perturbed by the Na^+ and Cl^- ions.

3.6 Experiment E: Testing the salt effect models in a different system

Input data: The temperature- and salt models were based on spectra (Experiments A-D) measured on ZnSe ATR crystal in the Bruker Equinox FTIR instrument in the 4000-600 cm^{-1} range at 1.93 cm^{-1} intervals. In contrast, the spectra in Experiment E were measured on ZnSe ATR crystal in a Bomen FTIR spectrometer in the 4000-748 cm^{-1} range at 1.93 cm^{-1} intervals.

The wavenumber scale of the water-temperature model (Figure 2) and the salt models (Figure 10) was therefore converted to the latter wavenumber scale by local linear interpolation. The uppermost curve in Figure 11(a) illustrates the reference spectrum **m** of pure water at room temperature after this wavenumber conversion. The uppermost curve in Figure 11(b) represents the absorbance spectrum of pure, dry gluten.

The second ensamble of spectra just below **m** in Figure 11(a) is the measured absorbance spectra of water with various concentrations of $MgSO_4$ or NaCl. Likewise, the second ensamble of spectra just below the gluten spectrum in Figure 11(b) is the measured absorbance spectra of gluten/water mixturees with various concentrations of $MgSO_4$ or NaCl. In both cases, the absorbance above 2500 cm^{-1} have been smoothed by a

Savitsky-Golay filter of first degree to reduce the very high measurement noise at the highest water peak in this experimental set-up. In both data sets, the characteristic sulphate peak near 1090 cm^{-1} is evident.

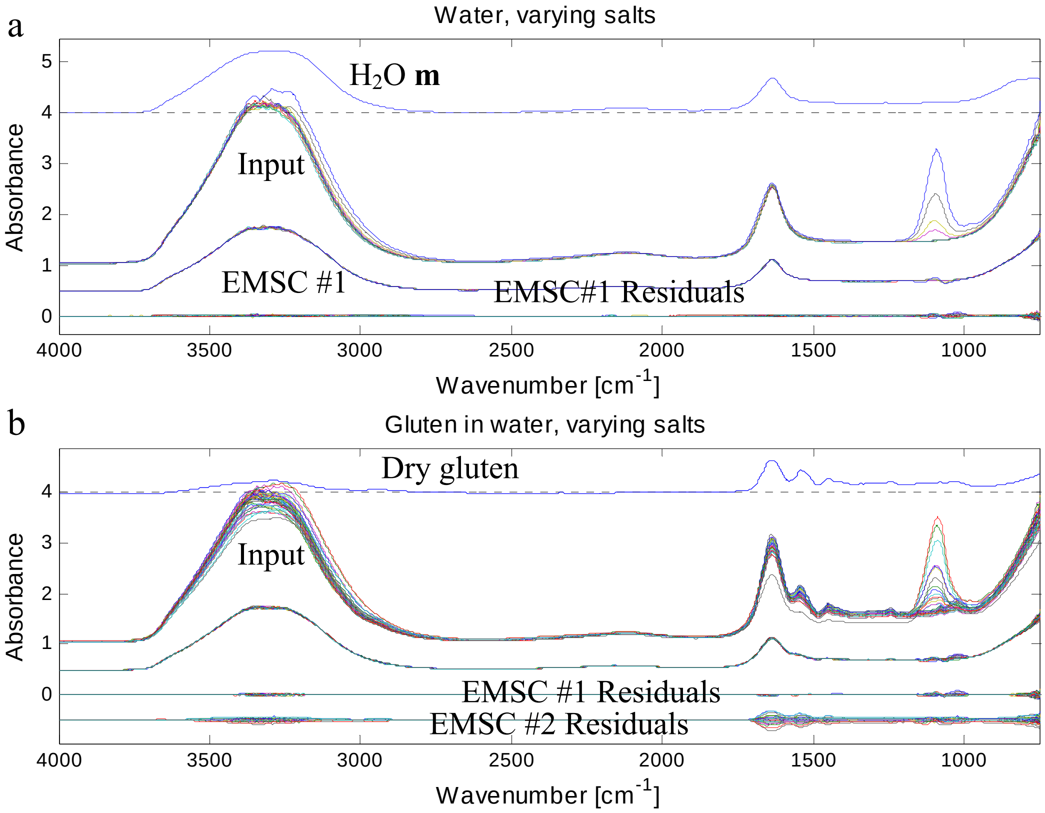


Figure 11. Experiment E: Testing salt effect models in a different system. a) Water with different salts added. From top: Reference spectrum m=water at room temperature; input absorbance spectra; after EMSC; EMSC residuals after subtraction of m. b) Hydrated wheat gluten with different salts added. From top: pure dry gluten; measured hydrated gluten input spectra; residuals after EMSC model# 1(subtracting gluten effect); residual of EMSC#1; residuals after EMSC #2 (retaining gluten effect).

In the water solutions in Figure 11(a), shifts in the water absorbance regions can be seen, primarily at 3200-2900 cm^{-1}. In the gluten slurries Figure 11(b) the variation patterns are more complex and appear to be a mixture of physical and chemical effects. Contributions from gluten can be recognized, e.g. around 1640, 1550 and 1455 cm^{-1}. No disturbing gas contributions from water vapour or CO_2 are seen in these data.

EMSC # 1: Each of the input spectra in Figure 11(a) and 11(b) were submitted to EMSC modelling around the multiplicative reference spectrum **m**. The additive part of the model consisted of the two water temperature component spectra (converted from Figure 2, to account for water binding by the gluten) and the estimated spectra of $MgSO_4$ and NaCl (converted from Figures 10(b) and 10(f), as well as the measured spectrum of dry gluten and a second degree baseline polynomial. The estimated effects of all of these model components were subtracted according to eq. (7c). The EMSC-treated spectra are displayed in Figure 11(a) and 11(b) as the next ensamble below the input spectra. The spectra have become visually almost indistinguishable, and the "residuals" (the EMSC-treated spectra minus **m**, displayed below the EMSC#1 spectra) show very little remaining structure at this graphical resolution. At higher resolution, details about minor peak shifts for the salt and protein peaks with e.g. salt concentration can be seen, but that is beyond the scope of the present paper. This illustrates one way to use the EMSC modelling: estimating and removing all known variation types in order to check one's assumptions and to reveal details.

EMSC # 2: Another way to use the EMSC modelling was tried. The spectrum of dry gluten was treated as a "good spectrum" to be modelled but *not* subtracted in eq. (7c). The bottom ensamble of curves in Figure 11(b) shows clear gluten-contributions in the EMSC residuals after subtraction of **m**.

The EMSC scores for the two salts were rescaled by eq. (7d) and used in prediction of the salt concentrations. The results are shown in Figure 12. For $MgSO_4$ the predictive ability was very good both in absolute and relative terms (r>0.99) for the water solutions as well as the gluten slurries. Even for NaCl the relative prediction ability was reasonable (r>0.9), but the concentrations were generally over-estimated in gluten. Hence, the EMSC models developed for salt solutions in the Bruker FTIR ATR instrument showed validity also for the Bomem FTIR ATR instrument although an offset-and slope-correction of the EMSC calibration models were necessary.

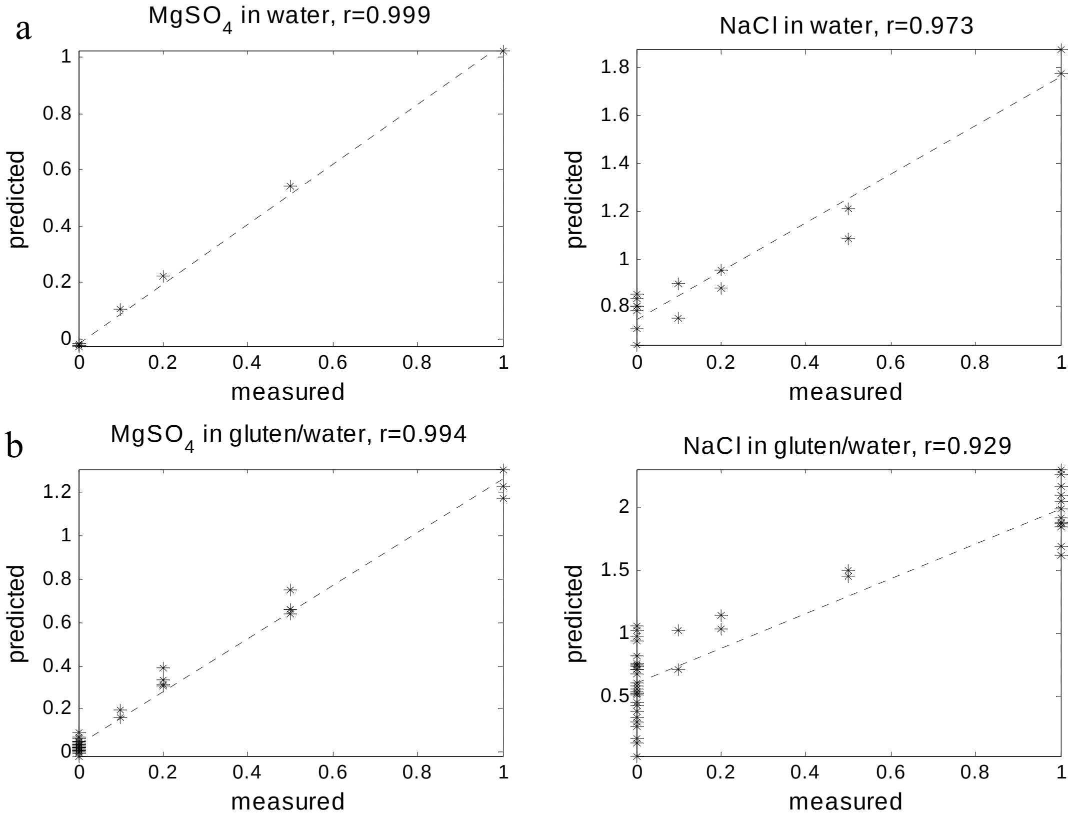


Figure 12. Experiment E: Testing salt effect models in a different instrument, continued. Salt concentrations measured vs. predicted (scaled EMSC scores). a) Water with different salts added. b) Hydrated wheat gluten with different salts added. Left: $MgSO_4$. Right: NaCl.

4. CONCLUSION AND PERSPECTIVES

FTIR is finding an increased importance in high-speed and high-resolution biospectroscopy in e.g. functional genomics and biotechnology [1-3, 28]. Like in NIR, unwanted variations in water content, and in sample temperature, organic molecules, salt content and other things that affect the structure of water can cause a lot of difficulties in FTIR-based biospectroscopy. Moreover, irrelevant water vapour and CO_2 can create nuisance in FTIR measurements. Finally, physical variation in sample or instrument contribute needless baseline- and scaling complexity to FTIR. The present paper has presented multivariate modelling methods (gas removal, EMSC) and estimated model parameters for handling some of these variation sources in FTIR. In addition, the method may improve the analysis of the water in biological samples and provide information on the water-solute interactions. It is increasingly realised that water cannot be ignored as a passive, constant and homogeneous solvent. Instead it is a central, complex reactant or catalyzer in biochemical reactions.

Acknowledgement: Ganesh Sockalingum and Isabelle Adt are thanked for providing the spectra of *C.albicans.*

REFERENCES

1. Kargosha K, Khanmohammadi M, Sarokhani M, Ansari F, Ghadiri M. Application of attenuated total reflectance Fourier transform infrared spectrometry to the determination of cephalosporin C in complex fermentation broths. *J. Pharm. Biomed. Anal* 2003; **31**: 571-577.

2 Mossoba MM, Kramer JKG, Fritsche J, Yurawecz MP, Eulitz KD, Ku Y, Rader JI. Application of standard addition to eliminate conjugated linoleic acid and other interferences in the determination of total trans fatty acids in selected food products by infrared spectroscopy. *J. Am. Oil. Chem. Soc* 2001; **78**: 631-634.

3. Mariey L, Signolle JP, Amiel C, Travert J. Discrimination, classification, identification of microorganisms using FTIR spectroscopy and chemometrics. *Vib. Spectrosc* 2001; **26**: 151-159.

4. Marechal Y. Infrared-spectra of water. 2: Dynamics of H_2O (D_2O) molecules. *J. Phys-Paris II* 1993; **3**: 557-571.

5. Marechal Y. Interaction configurations of H_2O molecules in a protein (Stratum Corneum) by infrared spectrometry. *J. Mol. Struct* 1997; **416**: 133-143.

6. Wei Z, ZhangY, Zhao L, Liu J, Li X. Observation of the first hydration layer of isolated cations and anions through the FTIR-ATR difference spectra. *J.Phys. Chem. A* 2005; **109**: 1337-1342.

7. Martens H, Nielsen JP, Engelsen SB. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. *Anal.Chem* 2003; **75**: 394-404.

8. Libnau FO, Kvalheim OM, Christy AA, Toft J. Spectra of water in the near- and mid-infrared region.*Vib. Spectrosc* 1994; **7**: 243-254.

9. Segtnan VH, Šašić Š. Studies on the structure of water using two dimensional near infrared correlation spectroscopy and PCA. *Anal. Chem* 2001; **73**: 3153-3161.

10. Šašić Š, Segtnan VH, Ozaki Y. Self-modeling curve resolution study of temperature-dependent Near-infrared spectra of water and investigation of water structure. *J. Phys. Chem A* 2002; **106**: 760-766.

11. Vanderkooi JM, Dashnau JL, Zelent B. Temperature excursion infrared (TEIR) spectroscopy used to study hydrogen bonding between water and biomolecules. *BBA-Proteins Proteom* 2005; **1749**: 214-233.

12. Omta AW, Kropman MF, Wouterson S. Influence of ions on the hydrogen-bond structure in liquid water. *J. Chem. Phys* 2003; **119**: 12457-12461.

13. Hribar B, Southall NT, Vlachy V, Dill KA. How ions affect the structure of water. *J. Am. Chem. Soc* 2002; **124**: 12302-12311.

14. Collins KD. Charge density-dependent strength of hydration and biological structure. *Biophys. J* 1997; **72**: 65-76.

15. Chaplin, M. Structure of liquid water. (htpp://www.sbu.ac.uk/water/). [20 January 2006].

16. Max J, Chapados C. Interpolation and extrapolation of infrared spectra of binary ionic aqueous solutions. *Appl. Spectrosc* 1999; **53**: 1601-1609.

17. Neto AMP, Sala O. The effect of temperature and $LiClO_4$ in the water structure: a Raman spectroscopy study. *Braz. J. Phys* 2003; **34**: 137-141.

18. Fischer WB, Fedorowicz A, Koll A. Structured water around ions- FTIR difference spectroscopy and quantum mechanical calculations. *Phys. Chem. Chem. Phys* 2001; **3**: 4228-4234.

19. Bruun SW, Kohler A, Adt I, Sockalingum GD, Manfait, M, Martens, H. Correcting ATR-FTIR spectra for water vapour and carbon dioxide. Submitted to *Appl. Spectrosc*.

20. Eisenberg D, Kauzmann W, eds. The structure and properties of water. Clarendon press: Oxford, UK, 1969.

21. Wall TT, Hornig DF. Raman intensities of HDO and structure in liquid water. *J. Chem. Phys* 1965; **43**: 2079-2087.

22. Khoshtariya DE, Dolidze TD, Lindqvist-Reis P, Neubrand A, van Eldik R. Liquid water (D_2O): a dynamic model emerging from near-infrared DO-D stretching overtone studies. *J. Mol. Liq* 2002; **96-7**: 45-63.

23. Woutersen S, Emmerichs U, Bakker H J. Femtosecond mid-IR pump-probe spectroscopy of liquid water: Evidence for a two-component structure. *Science* 1997; **278**: 658-660.

24. D'Arrigo G, Maisano G, Mallamace F, Migliardo P, Wanderlingh F. Raman-scattering and structure of normal and supercooled water. *J. Chem. Phys* 1981; **75**: 4264-4270.

25. Czarnik-Matusewicz B, Pilorz S, Hawranek JP. Temperature-dependent water structural transitions examined by near-IR and mid-IR spectra analyzed by multivariate curve resolution and two-dimensional correlation spectroscopy. *Anal. Chem. Acta* 2005; **544**: 15-25.

26. Decker M, Nielsen PV, Martens H. Near-infrared spectra of Penicillium camemberti strains separated by extended multiplicative signal correction improved prediction of physical and chemical variations. *Appl. Spectrosc* 2005; **59**: 65-68.

27. Chen Y, Zhang YH, Zhao LJ. ATR-FTIR spectroscopic studies on aqeuous $LiClO_4$, $NaClO_4$ and $Mg(ClO_4)_{(2)}$ solutions. *Phys. Chem. Chem. Phys* 2004; **6**: 537-542.

28. Kaderbhai NN, Broadhurst DI, Ellis DI, Goodacre R, Kell DB. Functional genomics via metabolic footprinting: monitoring metabolite secretion by Escherichia coli tryptophan metabolism mutants using FT-IR and direct injection electrospray mass spectrometry. *Comp. Func. Genom* 2003; **4**: 376-391.

2.6. Discussion and conclusion

From the literature, the ability of NIR and MIR to bring information on the arrangements of intermolecular interactions in a molecular system (water) with extensive hydrogen bonding was outlined.

On the other hand, in analysis of biological samples, where changes in macromolecule composition and interaction is of interest, changes in the water spectrum may be an interference in the analysis. For example, when the macromolecule interactions are perturbed by temperature and salts, the confounding between changes i the water- and the macromolecule spectrum creates problems in the data analysis. As demonstrated in the paper above, models of temperature- and salt- effects developed on basis of ATR-FTIR spectra of water and aqueous solutions were useful for removing these effects from the spectra of biological/food samples (biofilm and gluten). Thus, it appears to be a general preprocessing tool for improving the ATR-FTIR analysis.

The preprocessing method is going to be extended to NIR spectra and, based on the good results in MIR, the method is expected to improve the NIR analyses as well. The data analysis revealed serious drift over time in the ATR-FTIR measurements, causing the demand for short experiments (time series measurements), whereas this problem is thought to be less serious for NIR experiments.

Perturbations of the water spectra with the simple ionic solutes illustrated the information on solute-water interactions that can be obtained from the infrared water spectrum. In analysis of biological/food systems, the water-binding may be of interest, as the interaction of macromolecules with water is of foremost importance to macromolecule functionality and has profound influence on food properties [Tolstoguzov, 1996; Lewicki, 2004]. In the paper above, the variation between ‘free’ and ‘bound’ water in a developing biofilm was shown somewhat similar to a temperature variation. The preprocessing tool could therefore also be useful for studying hydration phenomena in temperature-controlled experiments.

2.7. References

Abe, H. (2004). Estimation of heat capacity and properties of water by spectrum decomposition of the second overtone band of OH stretching vibration. *J. Near Infrared Spec.* 12, 45-54.

Agmon, N. (1996). Tetrahedral displacement: The molecular mechanism behind Debye relaxation in water. *J. Phys. Chem.* 100, 1072-1080.

Baldwin, R.L. (1996). How Hofmeister ion interactions affect protein stability. *Biophys. J.* 71, 2056-2063.

Chaplin, M. (2006). (htpp://www.sbu.ac.uk/water/). [20 January 2006]

Chen J. Y., Matsunaga, R., Ishikawa, K., Zhang, H. (2003). Main inorganic component measurement of seawater using near infrared spectroscopy. *Appl. Spectrosc.* 57, 1399-1406.

Chen, Y., Zhang, Y.H., Zhao, L.J. (2004). ATR-FTIR spectroscopic studies on aqeuous $LiClO_4$, $NaClO_4$ and $Mg(ClO_4)(_2)$ solutions. *Phys. Chem. Chem. Phys.* 6, 537-542.

Collins, K.D. (1997). Charge density-dependent strength of hydration and biological structure. *Biophys. J.* 72, 65-76.

Czarnik-Matusewicz, B., Pilorz, S., Hawranek, J. P. (2005). Temperature-dependent water structural transitions examined by near-IR and mid-IR spectra analyzed by multivariate curve resolution and two-dimensional correlation spectroscopy. *Anal. Chem. Acta.* 544, 15-25.

D'Arrigo, G., Maisano, G., Mallamace, F., Migliardo, P., Wanderlingh, F. (1981). Raman-scattering and structure of normal and supercooled water. *J. Chem. Phys.* 75, 4264-4270.

Ebel, C., Faou, P., Kernel, B., Zaccai, G. (1999). Relative role of anions and cations in the stabilization of halophilic malate dehydrogenase. *Biochemistry.* 38, 9039-9047.

Eisenberg, D., Kauzmann, W. (1969). The structure and properties of water. Clarendon press, Oxford.

Fischer,W.B., Fedorowicz, A., Koll, A. (2001). Structured water around ions- FTIR difference spectroscopy and quantum mechanical calculations. *Phys. Chem. Chem. Phys.* 3, 4228-4234.

Gaiduk, V.I., Tseitlin, B.M., Crothers, D.S.F. (2004). Specific and unspecific interactions in polar fluids in view of wideband dielectric far-infrared spectra. *J. Mol. Liquids.* 114, 63-77.

Gergely, S., Salgo, A. (2003). Changes in moisture content during wheat maturation- what is measured by near infrared spectroscopy? *J. Near Infrared Spec.* 11, 17-26.

Giguére, P.A. (1987). The bifurcated hydrogen bond model of water and amorphous ice. *J. Chem. Phys.* 87, 4835-4839.

Graener, H. (1991). Anharmonicity and overtone spectra of OH stretching vibrations. *J. Phys. Chem.* 95, 3450-3453.

Grdadolnik, *J.* (2002). ATR-FTIR spectroscopy: its advantages and limitations*. Acta.Chem. Slov.* 49, 631-642.

Heise, H.M., Schrötter, H.W. (1995). Rotation-vibration spectra of gases. In: *Infrared and Raman spectroscopy: Methods and Applications,* (Schrader B., ed.).Wiley-VCH, Weinheim, Chap. 4.3, pp. 256-297.

Hribar, B., Southall, N.T., Vlachy, V., Dill, K.A. (2002). How ions affect the structure of water. *J. Am. Chem. Soc.* 124, 12302-12311.

Jin, Y. (2003). Near-infrared spectroscopic study at high temperatures and pressures. *J. Chem. Phys.* 119, 12432-12438.

Kalra, A., Tugcu, N., Cramer, S.M., Garde, S. (2001). Salting-in and salting-out of hydrophobic solutes in aqueous solutions. *J. Phys. Chem. B.* 105, 6380-6386.

Khoshtariya, D.E., Dolidze, T.D., Lindqvist-Reis, P., Neubrand, A., van Eldik, R. (2002). Liquid water (D2O): a dynamic model emerging from near-infrared DO-D stretching overtone studies. *J. Mol. Liq*. 96-7, 45-63.

Khoshtariya, D.E., Hansen, E., Leecharoen, R., Walker, G.C. (2003). Probing protein hydration by the difference O-H (O-D) vibrational spectroscopy: interfacial percolation network involving highly polarizable water-water hydrogen bonds. *J. Mol. Liq*. 105, 13-36.

Kropman, M.F., Bakker, H.J. (2003). Vibrational relaxation of liquid water in ionic solvation shells. *Chem. Phys. Lett*. 370, 741-746.

Kunz, W., Henle, J., Ninam, B.W. (2004a). 'Zur lehre der wirkung der salze' (about the science of the effect of salts): Franz Hofmeister's historical papers. *Curr. Opin. Colloid In*. 9, 19-37.

Kunz, W., Nostro, P.L., Ninham, B.W. (2004b). The present state of affairs with Hofmeister effects. *Curr. Opin. Colloid In***.** 9, 1-18.

Lewicki, P.P. (2004). Water as the determinant of food enginering properties. A review. *J. Food. Eng.* 61, 483-495.

Libnau, F.O., Kvalheim, O.M., Christy, A.A., Toft, J. (1994). Spectra of water in the near- and mid-infrared region.*Vib. Spectrosc*. 7, 243-254.

Lin, J., Brown, C.W. (1994). Novel applications of near-infrared spectroscopy of water and aqueous solutions: from physical chemistry to analytical chemistry. *Trac. Trends. Anal. Chem*. 13, 320-326.

Liu, J., Zhang, Y., Wang, L., Wei, Z. (2005). Drawing out the structural information of the first layer of hydrated ions: ATR-FTIR spectroscopic studies on aqueous NH_4NO_3, $NaNO_3$, and $Mg(NO_3)_2$ solutions. *Spectrochim. Acta A*. 6, 893-899.

Maeda, H., Ozaki, Y., Tanaka, M., Hayashi, N., Kojima, T. (1995). Near infrared spectroscopy and chemometrics studies of temperature-dependent spectral variation of water: relationship between spectral changes and hydrogen bonds. *J. Near Infrared Spec*. 3, 191-201.

Marechal Y. (1991). Infrared-spectra of water. 1. Effect of temperature and of H/D isotopic dilution. *J. Chem. Phys*. 95, 5565-5573.

Marechal Y. (1993). Infrared-spectra of water. 2: Dynamics of H_2O (D_2O) molecules. *J. Phys-Paris II*. 3, 557-571.

Marechal, Y. (1997). Interaction configurations of H_2O molecules in a protein (Stratum Corneum) by infrared spectrometry. *J. Mol. Struct*. 416, 133-143.

Max, J., Chapados, C. (1999). Interpolation and extrapolation of infrared spectra of binary ionic aqueous solutions. *Appl. Spectrosc*. 53, 1601-1609.

Neto, A.M.P., Sala, O. (2003). The effect of temperature and $LiClO_4$ in the water structure: a Raman spectroscopy study. *Braz. J. Phys*. 34, 137-141.

Omta, A.W., Kropman, M.F., Wouterson, S. (2003). Influence of ions on the hydrogen-bond structure in liquid water. *J. Chem. Phys*. 119, 12457-12461.

Rull, F. (2002). Structural investigation of water and aqueous solutions by Raman spectroscopy. *Pure Appl. Chem*. 74, 1859-1870.

Rønne, C. (2000). Intermolecular liquid dynamic studied by THz-spectroscopy. Ph.D. thesis, Aarhus University.

Šašić, Š., Segtnan, V.H., Ozaki, Y. (2002). Self-modeling curve resolution study of temperature-dependent Near-infrared spectra of water and investigation of water structure. *J. Phys. Chem. A*. 106, 760-766.

Segtnan V.H., Šašić, Š. (2001). Studies on the structure of water using two dimensional near infrared correlation spectroscopy and PCA. *Anal. Chem*. 73, 3153-3161.

Symons, M.C.R. (1975). Water structure and hydration. *Phil. Trans. R. Soc. Lond. B*. 272, 13-28.

Tanaka, M., Shibata, A., Hayashi, N., Kojima, T. Maeda, H., Ozaki, Y. (1995). Discrimination of commercial natural mineral waters using near infrared spectroscopy and principal component analysis. *J. Near Infrared Spec*. 3, 203-210.

Toft, J., Sanchez, F.C., van den Bogaert, B., Libnau, F.O., Massart, D.L. (1996). Resolution of overlapping mid-infrared spectra using SIMPLISMA and second-order derivative approach. *Vib. Spectrosc*. 10, 125-138.

Tolstoguzov, V. (1996). Structure-property relationships in foods. In: *Macromolecular Interactions in Food Technology*, (Parris, N., Kato, A., Creamer, L.K., Pearce, J., eds.). Am. Chem. Soc., Washington DC. pp. 2-14.

Tongraar, A., Rode, B.M. (2004). Dynamical properties of water molecules in the hydration shells of Na+ and K+: ab initio QM/MM molecular dynamics simulations. *Chem. Phys. Lett*. 385, 378-383.

Urquidi, J., Singh, S., Cho, C.H. (1999). Origin of temperature and pressure effects on the radial distribution function of water. *Phys. Rev. Lett*. 83, 2348-2350.

Vanderkooi, J.M., Dashnau, J.L., Zelent, B. (2005). Temperature excursion infrared (TEIR) spectroscopy used to study hydrogen bonding between water and biomolecules. *BBA-Proteins Proteom*. 1749, 214-233.

Wall, T.T., Hornig, D.F. (1965). Raman intensities of HDO and structure in liquid water. *J. Chem. Phys*. 43, 2079-2087.

Walrafen, G.E., Hokmabadi, M.S., Yang, W.-H., Chu, Y.C. (1989). Collision-induced Raman scattering from water and aqueous solutions. *J. Phys. Chem*. 93, 2909-2917.

Wang, Y., Murayama, K., Myojo, Y., Tsenkova, R., Hayashi, N., Ozaki,Y. (1998). Two-dimensional Fourier transform near-infrared spectroscopy study of heat denaturation of ovalbumin in aqueous solutions. *J. Phys. Chem. B*. 102, 6655-6662.

Wei, Z., Zhang,Y., Zhao, L., Liu, J., Li., X. (2005). Observation of the first hydration layer of isolated cations and anions through the FTIR-ATR difference spectra. *J.Phys. Chem. A*. 109, 1337-1342.

Wernet, P., Nordlund, D., Bergmann, U., Cavalleri, M., Odelius, M., Ogasawara, H., Naslund, L.A., Hirsch, T.K., Ojamae, L., Glatzel, P., Pettersson, L.G.M., Nilsson, A. (2004). The structure of the first coordination shell in liquid water. *Science*. 304, 995-999.

Wiggins P.M. (2000). High and low density intracellular water. *Cell. Mol. Biol*. 47, 735-744.

Woutersen, S., Emmerichs, U., Bakker, H. J. (1997). Femtosecond mid-IR pump-probe spectroscopy of liquid water: Evidence for a two-component structure. *Science*. 278, 658-660.

Chapter 3: Spectroscopic measurement of protein conformations and interactions

Several methods may provide information on the structure of proteins. The three-dimensional structure is investigated e.g. by use of X-ray crystallography, whereas the contents of secondary structures may be determined by use of MIR and other spectroscopic techniques such as circular dichroism (CD) and Raman. These techniques do not suffer from the limitations of X-ray crystallography, such as the requirement of the protein in the crystallised form, and instead, they may be applied for analysis of protein structures in solution. In addition, MIR may give information on protein structures in a food matrix. NIR may be another potential method for this type of analysis, which can be done by use of only a few other methods (e.g. NMR).

In this chapter, the application of both MIR and NIR in protein structure analysis is introduced, where after the capacity of NIR in this field is further investigated in Experiment III and in Paper IV (Appendix III-1). First an overview of the interactions that define the protein structures is given.

3.1. Protein conformations and interactions

For being catalytically active, a protein needs a specific three-dimensional structure, which is determined from the amino acid sequence.

Secondary and tertiary structures

The polar backbone C=O and NH groups are forced to undergo hydrogen bonding in the hydrophobic protein core, and they widely participate in the periodic main chain hydrogen bonding patterns, which define the different secondary structures. On the average, 89 % of the protein residues are involved in these structures [Darby, 1993]. At the origin of the secondary structures is the Corey-Pauling rules, which establishes the conditions of hydrogen bonding in the backbone and considers the polarity and steric hindrance of the side chains. The conditions apply to the α-helix and β-sheet, which are the principal regular secondary structures in proteins.

Helixes: In the α-helix structure, the protein chain is twisted into repeating turns thereby forming a helix with hydrogen bonding taking place between residues aligned with the helix axis. In a perfect α-helix, all residues participate in *i*→*i*+4 and/or *i*-4←*i* hydrogen bonds (the oxygen in residue *i* form hydrogen bond to the NH group in residue *i*+4). See Fig. 3.1. Favourable dipole interactions provide the α-helix with a high stability. Also, side chains that stick out perpendicular from the helix axis may interact with each other so as to further stabilise the helix [Darby, 1993]. Deviations from the ideal α-helix conformation

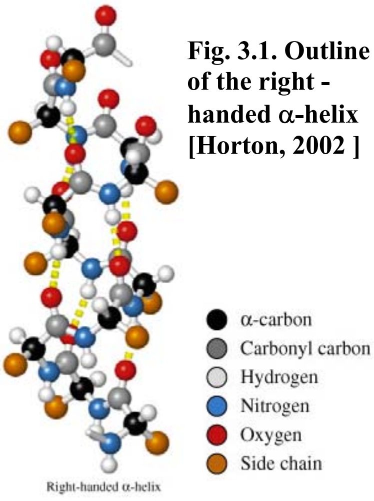


Fig. 3.1. Outline of the right - handed α-helix [Horton, 2002]

occur frequently, wherefore the hydrogen bonding pattern may vary between α-helices [Andersen, 2001a]. Another helix structure, which accounts for about 12 % of all protein helices is called the 3_{10} helix and has an entirely different hydrogen bonding pattern, namely the *i*→ *i*+3 pattern. The ordinary 3_{10} helix is only 3,5 residues long (compared to 11 residues in the α-helix) and may occur as the end-turn of α-helices [Darby, 1993].

β-sheets: The β-sheet is composed of aligned β-strands that run either parallel or antiparallel and therefore form either purely parallel, purely antiparallel or mixed type β-sheets [Darby, 1993]. In the antiparallel conformation, all NH groups in one chain is aligned with C=O groups in the other chain, giving rise to straight interchain hydrogen bonds, whereas in the parallel conformation, the hydrogen bonds are distorted. When residues that are not hydrogen bonded to a partner strand occur, the hydrogen bonding pattern is disturbed and β-bulges are formed [Darby, 1993]. Likewise, the β-sheets may be bend and consequently have dihedral angles that are different from the ideal ones. The side chains of adjacent residues successively point out on different sides of the sheet, where they may interact with each other. The formation of β-sheets can occur independent of the sequence, as side chain interactions do not necessarily take part in this structure [Shukia, 2004].

Turns etc: On the surface of globular proteins are found structures that can reverse the direction of the polypeptide chain. Two structures that accomplish the reversals are the β-turn and the omega-loop, which are formed by ~30 % of all amino acid residues [Darby, 1993]. They contain relatively polar residues and often participate in interactions with other molecules (e.g. the solvent). The β-turn is the most prevalent of the two structures and often connects strands in antiparallel β-sheets. It comprises four residues and can form a 10-membered hydrogen bonded ring [Mantch, 1993]. Several different types of β-turns with different hydrogen bonding and different dihedral angles exist. Regularly repeated β-turns may result in the formation of the β-spiral, which has been evidenced to occur in some proteins containing repeat motives (e.g. gluten proteins) [Pezolet, 1992]. Other non periodic structures (irregular or random structures) include loops and bends, as well as extended β-strands that are not involved in β-sheets but often interact with the solvent.

Tertiary structure: The three-dimensional structure is established by interactions between side chains of amino acid residues that are far apart in the primary sequence. Most of the interactions on the tertiary level are non-covalent, although the disulfide (S-S) bridges may be important to the tertiary structure. The stability of the protein conformation is believed mostly to be a result of the hydrophobic effect. Due to their non-favourable interaction with water, hydrophobic residues agglomerate and become buried in the protein core [Darby, 1993]. The hydrophobic interactions occur between side chains of hydrophobic residues (Met, Ala, Val, Ile, Leu, Phe) [Darby, 1993]. The amino acids with hydrogen bonding potential of their side chains are Glu, Gln, Asp, Asn, Ser, Tyr, Thr, Trp, Lys, His, Arg that may function as hydrogen donors, acceptors or both. The charged amino acid residues (Lys, Glu, Asp, His, Arg) can also form salt bridges. Even a few salt bridges can have profound effect on stabilisation of the tertiary structure [Darby, 1993]. Also the CH group may act as hydrogen donor in a weak CH..O interaction and in an even weaker CH…π interaction, in which the π-electrons from carbonyl groups and aromatic amino acid side chains act as hydrogen acceptors [Brandl, 2001]. It has been estimated that almost 50 % of the aromatic amino acids (Trp, Tyr, Phe) in a protein are involved in CH-π interactions, and that nearly all Trp residues participate in the weak hydrogen bonds (including also NH..π and OH..π interactions) [Brandl, 2001].

Quaternary structure and protein-protein interactions

The quaternary structure of proteins is the assembly of subunits into homo- or hetero-multimers, in which the subunits are held together by protein-protein interactions. The specific protein-protein interactions are essential to almost all levels of the cell function (transport across membranes, signal transduction, gene regulation, muscle contraction etc.), and nearly all proteins participate in protein-protein interactions as part of their functions. Some of the complexes are obligatory, meaning that the monomers do not naturally exist [Ofran, 2003]. On the other hand, enzymes and carrier proteins etc. that can exist as monomers may also be able to form dimers, tetramers or higher oligomers, depending on the environments or external factors. (The folate binding proteins (FBPs) are examples hereof and are dealt with in Paper III).

The binding interface of protein complexes is as tightly packed as in the interior of proteins, and the interactions between side chains as well as desolvation of the protein surfaces lead to the stabilisation of the complexes [Kimura, 2001]. Hydrophobic interactions are main contributors to the complex stability, whereas the rate and specificity of the protein association are determined by long-range specific electrostatic interactions [Kimura, 2001]. The hydrophobic interactions are more frequent in the permanent interactions than in transient interactions [Ofran, 2004]. So, whereas Trp residues are overrepresented in most interfaces, they are less abundant in interfaces involved in transient interactions between identical proteins [Ofran, 2004]. Whilst hydrophobic interactions take place in the core of the interphase, salt bridges and hydrogen are most abundant in the periphery [Ofran, 2004; Darby, 1993].

Complexes having 0-46 intermolecular hydrogen bonds are known. Arg side chains are involved in many of these, but also intermolecular hydrogen bonds form between the peptide backbones [Dou, 2004; Sakurai, 2002]. Thus, intermolecular β-sheets are frequently found in dimers and oligomers [Dou, 2004]. For example, this structure has been shown important to the formation of β-lactoglobulin (BLG) dimers, and it provides the correct positioning of the monomers for binding to each others [Sakurai, 2002]. Dou et al (2004) found that 15.8 % of all the proteins in the protein database have β-sheet interfaces, and thus, intermolecular β-sheet can be considered an important type of protein-protein interaction. A distinct role of intermolecular β-sheets is in the formation of amyloid aggregates, which are involved in diseases like Alzheimer's, Creutzfeldt-Jakob, Huntington's and Parkinson's. The β-sheets in the amyloid aggregates are arranged in a special cross β-structure, in which β-sheets run perpendicular to the fibril axis [Jahn, 2005]. The folding is thought to be inherently different from normal protein folding and probably based on main-chain interactions that are allowed when the normal specific folding is destabilised [Fändrich, 2003; Jahn 2005].

Protein-protein interactions in foods

Aggregates and gels, formed by protein-protein interactions, are important contributors to the structure of many foods, and for example, the whey proteins (BLG, BSA and α-lactalbumin) are used widespread as food ingredients due to their gelling abilities. Gelation is the formation of a continuous well-defined three-dimensional network of particles or polymers, which is able to immobilise water. On the other hand, the unwanted outcome of aggregation, namely precipitation, does not lead to immobilisation of water [Li-Chan, 1996]. Protein aggregation in foods involves unspecific interactions that develops only after the food processing has caused some unfolding of the proteins. Even though heat-treatments at 60-80°C may not affect the secondary structure much, it may lead to increased flexibility and exposure of hydrophobic regions, thus increasing hydrophobic interactions and promoting aggregation [Li-Chan, 1996; Petsev, 2000]. Also, ionic interactions, hydrogen bonding and disulfide bridges are believed to take part in aggregate formation [Li-Chan, 1996]. The aggregates may interact and form strands that again may assemble in a three-dimensional gel network. Different gelling conditions may cause diverse arrangement of the strands and lead to formation of either coagulates, opaque gels, or transparent gels [Li-Chan, 1996; Tolstoguzov, 1996].

In studies of the aggregation process, a conversion of α-helices into random coils and β-sheet structures is often seen [Li-Chan, 1996; Shukla, 2004; Sokolowski, 2003]. However, the secondary structure changes depend on both the protein and its concentration, and β-sheet to random coil transformations take place in some cases [Li-Chan, 1996]. The intermolecular β-sheet, which can form independently of the amino acid composition, has been pointed out as a fundamental part of the aggregation process, as the extended polypeptide chains that are formed upon protein unfolding are allowed to align closely and interact with neighbouring chains by hydrogen bonding. These β-sheet interactions are stabilised by hydrophobic interactions and intermolecular S-S bridges [Srisailam, 2002; Domenek, 2002; Li-Chan, 1996]. The latter may be formed if Cys residues become exposed during the unfolding, and they have been shown e.g. in BSA and BLG gels, where they cause the irreversible denaturation [Gezimati, 1996].

Intermolecular β-sheets may form junction zones that stabilise the networks, but their relative significance in the cross linking of protein networks is not established [Li-Chan, 1996]. However, the β-sheet interactions have been suggested to be the dominating interactions in the gel-state of some proteins [Allain, 1999]. Also, intermolecular β-sheets have been shown of importance to the viscoelastic properties of dough [Wellner, 2003], and Lefevre et al (2003) suggested that the viscoelastic property of protein-stabilised oil-water emulsions could result from the existence of intermolecular β-sheet interactions between proteins, aggregated in the interfaces. On the other hand, hydrophobic interactions have also been pointed out as the most important in cross-linking of protein gels [Li-Chan, 1996].

The detection of protein aggregates can be done by use of e.g. rheology, microscopy, dynamic light scattering or diffusing wave spectroscopy, of which only the latter method is nonintrusive and works for the highly concentrated food systems [Alexander, 2006]. With spectroscopic methods (vibrational, laser light scattering, and mass spectrometry) it is possible to detect the aggregates before they become visible [Li-Chan, 1996].

3.2. The infrared protein spectrum

The MIR protein spectrum

The protein backbone shows nine vibrational modes, which result in nine fundamental absorption bands specific for proteins (and polyamides), and which have been named the amide I-amide VII, amide A, and amide B modes. The amide IV-VII bands have very low intensities in MIR and are not of importance in MIR or NIR analyses. The amide I and amide II modes are not influenced by amino acid side-chains and only depend on protein backbone conformation, whereas the amide III mode is more complex, since it depends also on amino acid side-chains.

Amide I (1700-1600 cm^{-1}) is mainly a C=O stretching vibration with a weak coupling to out-of-phase C-N stretching, NH in-plane bending and CNN deformation. About 70-85 % of the band intensity is from C=O stretching and 10-20 % is from C-N stretching [Barth, 2002].

Amide II (1595-1510 cm^{-1}) is an N-H in plane bending (40-60 %) coupled to C-N stretching (18-40 %) and also to C-C stretching (10 %) and C=O in plane bending [Barth, 2002].

Amide III (1350-1200 cm^{-1}) has usually been described as mainly an C-N-H in-plane bending coupled to C-N stretching, but recent studies have revealed that for molecules containing CH_2 groups, the amide III result mostly from coupled NH and CH_2 deformations. Thus, the mode is somewhat mixed with side chain vibrations [Fu, 1994].

Amide A (3310-3270 cm^{-1}) is an N-H stretching vibration, totally localized to the NH group. The weaker band from amide B (3100-3030 cm^{-1}) is commonly reported as being either the 1st overtone or a combination mode of amide II, which is amplified through Fermi resonance with amide A [Krimm, 1986; Lazarev, 1974]. However, the assignment is subject to controversy, as will be described later [Wang, 1998].

In addition, the amino acid side chains absorb light in the MIR. For example, most amino acids cause absorptions from aliphatic groups. The CH and CH_2 bendings are observed at around 1450 cm^{-1} and the symmetric and antisymmetric CH_3 bending at 1375 and 1465 cm^{-1}, respectively [Barth, 2000]. Some of the CH_2 and CH_3 bending vibrations are relatively uncoupled from other vibrations, but the twisting, rocking and wagging are coupled to other modes. The symmetric and antisymmetric -CH_2 and -CH_3 stretchings give rise to several bands between 2875 and 3100 cm^{-1}. Other side chain absorptions are indicative of specific amino acids.

The NIR protein spectrum

In addition to the bands in MIR, proteins have their complementary fingerprints in the NIR. The band assignment is difficult in this region and is less well established than in the MIR due to the numerous and overlapping bands.

In the combination band region (2000-2500 nm), bands characteristic of the protein amide groups appear at ~2050 and ~2170-2190 nm. Contradictory assignments as to their origins are found in the literature. The 2050 nm band has mostly been attributed to amide A combined with either amide I or amide II (amide A/I or amide A/II). The second option was demonstrated by Wang et al (2004) in an experimental and theoretical study, whereas, a band at 2012 nm has been ascribed to the amide A/I combination [Bai, 2005]. The broad 2180 nm band has commonly been ascribed to the amide B/II combination [Wang, 1994]. However, in 2nd derivative spectra, the band is seen to consist of at least two overlapping bands at ~2170 and ~2190 nm, in the vicinity of which are found several different combination bands. Czarnik-Matusewicz (1999) proposed three assignments in the 2100-2186 nm region: amide B/I at 2100-2124 nm, amide B/II at 2160 nm and amide A/III at 2186 nm. Likewise, the assignment of a weak absorption at 2215 nm to amide A/III was adopted by Murayama et al (2002), whereas Wang et al (1994) assigned the 2206 nm band to the CH str. combined with amide I.

Higher order combinations of the amide bands are found as weaker absorptions in the lower wavelength region where overtones also appear. For example, the 1600 nm band has been assigned to the combination of free NH stretch with the 1st overtone of amide II [Sadler, 1984], and a band at 1255 nm was assigned to the combination of amide II with the 1st overtone of amide A [Czarnik-Matusewicz, 1999].

The amide combination bands outlined above stem from the protein backbone vibrations, and their intensities reflect the total protein concentration, almost independently of amino acid composition (except that the 2190 nm band has been assigned to primary amide groups [Yamashita, 1994]). Therefore, the amide bands have been used frequently for protein concentration analysis, and the 2170 nm band has been found the most suitable for this purpose. Yamashita et al (1994) established that nitrogen-containing amino acid side chains do not contribute to this band. An amide combination band, which on the other hand is dependent on the amino acid composition, is found at 1980 nm and results from the primary amide group from Gln. It has been attributed to amide A/II, and the intensity of this band correlates with the content of Gln in proteins [Holly, 1992].

Above 2220 nm, combination bands originating from amino acid side chains appear, and different amino acid compositions result in some variation in this region. Many bands emerge in the region, since both the C-H symmetric and anti-symmetric stretching can combine with either symmetric or anti-symmetric C-H bending. Furthermore, the CH, CH_2 and CH_3 groups cause absorptions at different wavelengths. Czarnik-Matusewizc recognised the absorptions at 2318, 2335, 2355, 2366-2382 nm as CH-combination bands from protein side chains [Czarnik-Matusewicz, 1999]. A band at 2290 nm has been reported as the combination of

CH_2 stretching and amide III [Murayama, 2000], and absorptions at 2255 nm has been ascribed to the OH-stretching combined with OH-deformation from amino acid side chains (e.g. Ser, Thr) [Murayama, 2002]. Higher order C-H combination bands appear at lower wavelength, e.g. at 1410 nm [Murayama, 1998].

The 1st C-H stretching overtone bands from amino acid side chains are found at ~1660-1780 nm, and again, band shapes and positions depend on the amino acid composition. Aromatic C-H stretching vibrations appear separate from the aliphatic. The 1st N-H stretching overtone bands are found at ~1460-1640 nm. Sadler et al (1984) assigned the 1490 nm band to the 1st overtone of free NH stretching. The 1st overtone of amide A may appear between 1523 and 1584 nm, whereas the 1st overtone of amide B was suggested at 1638 nm [Sadler, 1984; Czarnik-Matusewicz, 1999; Šašić, 2000].

3.3. Protein structure analysis by MIR

The sensitivity of the amide bands to hydrogen bonding strengths and dipolar couplings provide MIR its sensitivity to the secondary structures. However, also interactions with the solvent (usually water) and with other molecules cause perturbations of the protein spectrum.

Introduction to MIR in protein analysis

MIR has been used frequently since 1970s for quantitative protein structure analysis, and it is now a well-established method of increasing importance in the protein research [Jackson, 1995; Haris, 1999]. The FTIR technique is the most widespread MIR technique for this purpose. With FTIR it is possible to detect even small conformational changes in proteins, and the method has been applied for the analysis of structural changes induced in proteins by varying e.g. pH-, salts, ionic strength and temperature [Matsuura, 2001; Murayama, 2001, Perez, 2000; Mohney, 2000; Dong, 1997; Lin,1999], induced by interactions with ligands [Neault, 1998] or by adsorptions to surfaces [Fang, 1997; Lefevre, 2003]. The presence of other macromolecules does not obscure the protein analysis, wherefore FTIR is highly suitable for studying the effect of interactions with other macromolecules. For example, the effect of protein-lipid interactions as regards both protein and lipid conformations has been analysed by FTIR [Lefevre, 2001; Borgrah, 1999]. FTIR is a versatile tool, not limited by the nature of the physicochemical environments or the physical state of the protein, and it has been used for protein studies in membranes, detergents, films, emulsions, gels and in the viscoelastic or semi-solid form by use of the ATR cell [Borgrah, 1999; Fang, 1997; Lefevre, 2003; de Jongh, 1996; Lorenz-Fonfria, 2003; Wellner, 2005]. Raman is an alternative method for measuring turbid samples, whereas the light scattering sensitive UV-, visible-, fluorescence- and CD spectroscopies are not applicable for these samples. Thus, FTIR is an often used method for the analysis of protein gelation and aggregation [Lefevre, 2001; Srisailam, 2002; Allain, 1999].

The consequence of the small path length required in FTIR measurements is that a high protein concentration (10-20 mg/ml) is needed for obtaining the signal to noise ratio (SNR) necessary for conformational analyses [Haris, 1999; Jackson,1995].

Sensitivity of Amide I to secondary structures

The amide bands are all affected by the protein structure. However, the amide I band is the most sensitive to secondary structure, and since it also is the most intense protein band, it is often chosen for protein structure analysis. An empirical relation between protein secondary structure and the shape of the amide I band was discovered by Elliott and Ambrose in 1950 [Elliott, 1950]. The findings were that the frequency of the amide I mode depend on the predominant secondary structure of the protein, and that the different structure types absorb at specific frequency intervals in the broad amide I envelope. Therefore, the amide I band of proteins are usually composed of several overlapping sub bands, resulting in a rather featureless profile.

The hydrogen bonding geometries in the different secondary structures are of foremost importance to their characteristic amide I frequencies, and the strength of the hydrogen bond is one parameter that affect the amide I spectrum. A non-hydrogen bonded amide C=O group is found to have a vibrational frequency of 1660-1666 cm^{-1}, and the stronger the hydrogen bonding in the secondary structure, the lower is the characteristic frequency [Barth, 2002; Jackson 1995]. Consequently, the frequencies follow the order: turn>α-helix >β-sheet>intermolecular β-sheet. Absorptions above 1666 cm^{-1} originate for example from non-hydrogen bonded β-turns, which have a disturbed planarity of the amide group and thus an increased electron density in C=O groups [Jackson, 1995]. Other absorptions in the high-frequency region arise from β-sheets and these absorptions can be explained from dipole couplings that split the amide I mode 50-70 cm^{-1}, as will be described below.

The different strengths of hydrogen bonding in the secondary structures do not provide a satisfactory explanation for their different frequencies. The structural sensitivity of the amide I frequency also arises from vibrational coupling between the individual peptide groups, which may be represented as equivalent amide I oscillators [Torii, 1992, Barth, 2002; Brauner 2005]. Couplings between the oscillators by electrostatic- and through-bond interactions cause the appearance of collective delocalised modes that involve several oscillators, between which the absorbed energy is shuttled back and force. In the collective system, the vibrational transition energies are shifted away from the transition energies of the single peptide group. The size of the shift depends on the couplings between the individual oscillators, and as the couplings are determined by the relative position, orientation and connectivity of the peptide groups, the transition energies reflect the secondary structure and to some extent the tertiary structure of the protein [Barth, 2002; Brauner, 2005; Moran, 2004; Torii, 1992]. The most important coupling is the transition dipole coupling (TDC) [Torii, 1992], which is an electrostatic resonance interaction taking place between oscillating dipoles on neighbouring amide groups when one of the groups is in the exited state. One effect of TDC is splitting of the amide I band (as mentioned) in an out-of-phase and an in-phase band. The size of the splitting depends on the magnitude of the TDM: the larger the TDM, the larger is the coupling and the splitting [Barth, 2002]. TDC can explain the high-frequency band that appears for antiparallel β-sheets, as they have large TDC interactions between peptide groups in different strands [Barth, 2002; Torii]. The α-helix absorptions are only modestly split by TDC [Torii, 1992].

Due to the influence of hydrogen bonding on the amide I frequency, the hydrogen bonding interaction of the amide groups with the solvent also affects the amide I frequencies. Proteins in aqueous solutions interact with the water through ionic interactions and hydrogen bonding. For example, the surface carbonyl groups already involved in helix structure hydrogen bonding can accommodate an extra hydrogen bond to water [Walsh, 2003]. Thereby, the amide I spectrum is affected and can give information on protein hydration and solvation. The amide I vibration shifts to a lower frequency, as the carbonyl group become exposed and binds to the solvent [Walsh, 2003]. The heterogenic interaction with the solvent also leads to broader amide I bands for the exposed groups than for the buried groups [Walsh, 2003]. Thus, the amide I frequencies of the different secondary structures depend on the solvation.

Correspondence between secondary structures and amide I band shape

The usual quantitative amide I analysis involves a deconvolution and curve fitting step, in which several sub bands are fitted to the broad amide I band by an iterative least squares method. Each sub band is then ascribed to a specific secondary structure, and the relative band areas are calculated by integration and translated into relative contents of the secondary structures.

Bands in the frequency interval ~1620-1640 cm^{-1} is usually ascribed to β-sheet structures, with the exact position being influenced by the varying strengths of the hydrogen bonding and transition dipole coupling in different β-strands [Haris, 1999]. For example, β-strands on the edges of β-sheets (β-edges) can interact strongly with other structures, other molecules or the solvent and may thus be found at a lower frequency than the major β-sheet [Haris, 1999; Dong, 2000a,b; Mohney, 2000; Lefevre, 2001, 2003]. The antiparallel β-sheet structure also causes the appearance of a five times less intense high-frequency band at 1670-1695 cm^{-1} due to band splitting by TDC [Torii, 1992; Cheatum, 2004]. The possibility to distinguish between parallel and antiparallel β-sheets in the amide I analysis was e.g. shown by Oberg et al (2004).

Due to the diversity of β-turn geometries, their absorptions appear in a broad frequency range, and bands between 1666 cm^{-1} and 1690 cm^{-1} can be ascribed to non-hydrogen bonded β-turn structure (if not β-sheet structures) [Jackson, 1995]. Furthermore, β-turns with internal hydrogen bonds cause absorptions at 1638-1646 cm^{-1} [Mantsch, 1993].

A band component at 1650-1658 cm^{-1} is usually ascribed to α-helix, but there may be overlap from bands due to random structure at 1648±2 cm^{-1} [Haris, 1999] and loops at 1658-1665 cm^{-1} [Khurana, 2000]. However, α-helix usually contributes the most to this band. The α-helices at least 10 residues long contribute mostly to an A-type combination of individual amide I vibrations, suggested to appear at 1660 cm^{-1} [Al Azzam, 2002]. Short helices do not exhibit a single band because of a weaker coupling, and they cause broader bands with several maxima, although the most intense band may appear at 1650 cm^{-1} [Torii, 1992].

The external α-helices are found at lower frequencies than the internal α-helices, and solvent interactions or distortions of α-helices may cause further low-frequency shifts, so that the α-helix band appears in the 1620-1640 cm^{-1} region [Mohney, 2000].

Absorptions at 1662±3 cm^{-1} have been assigned to the more seldom 3_{10}-helix, but overlap with the β-turn often prevents its detection [Dong, 2000b]. Spectra of the different aperiodic structures like bends, irregular structure and isolated β-strands do not appear to be distinguishable [Oberg, 2004]. The common assignments are shown in Table 3.1.

Wavenumber region [cm^{-1}]	Assignment
1695-1660	β-sheet, β-turns
1660-1650	α-helix
1650-1640	Random/unordered
1640-1620	β-sheet

Table 3.1. Common assignments in the amide I band, according to [Torii, 1992]. The deviations from these empirical rules are described in the text.

Difficulties in the amide I analysis arise from the ambiguous band assignments, the baseline correction and the subtraction of the overlapping water band and amino acid side chain absorptions etc. In addition, the rotation-vibration spectrum associated with the bending mode of gaseous water overlays the amide I band. A method for water vapour and CO_2(g) correction is described in Paper II (Appendix II). A prediction accuracy in the order of ±4 % is considered acceptable at best [Oberg, 2004].

Sensitivity of other amide bands to secondary structures

Like the amide I mode, the amide II band is influenced by hydrogen bonding and TDC. For example, the NH bending frequency of an amide group depends on whether the NH group forms a hydrogen bond to a C=O group or to water [Maréchal, 1997; Maeda, 2000]. However, the correspondence between band shape and secondary structure is complex. Assignments taken e.g. from Lacey et al (1998) and Krimm et al (1986) are β-sheet absorptions at 1517 and 1524 cm^{-1}, β-turn absorption at 1568 cm^{-1} and α-helix absorption at 1545 cm^{-1} with a shoulder at 1517 cm^{-1}. Others report the parallel β-sheet absorption to appear at 1530-1550 cm^{-1}. These bands are also overlapped by random coil absorptions at 1520-1545 cm^{-1}. Due to the above difficulties, the amide II band has not been studied in details for its feasibility in protein structure analysis. However, the use of the amide II in combination with the amide I band has been found to give better estimates of α-helix than with the use of the amide I band alone, and actually Oberg et al (2004) saw a better estimation of α-helix and turns by use of amide II compared to amide I.

The amide III band is found to contain more structural information than the amide II band [Cai, 1999; Fu, 1994]. Only ~30 % of the band results from amide vibrations, due to the high contribution from side chain vibrations (the amide III mode is significantly mixed with CH_2 wagging). In addition, the low intensity is problem, and the band variations are not fully understood. Anyway, it has been possible to estimate secondary structure contents of several proteins in good agreement with X-ray crystallographic data [Cai, 1999; Fu 1994]. The advantages of using the amide III band compared to amide I are: no interference from water and less overlapping of the bands from the different secondary structures, so that no band-narrowing is necessary. The assignments adopted from Cai et al (1994) are: 1340-1295 cm^{-1}: α-helix, 1295-1270 cm^{-1}: β-turns, 1270-1250 cm^{-1}: random structure, 1250-1220 cm^{-1}: β-sheet. Extended chains have been reported to cause a band at 1240-1247 cm^{-1} [Griebenow, 1995].

The amide A is sensitive to hydrogen bonding, which may shift this band to a lower frequency [Lazarev, 1974], and the band is sensitive to the protein secondary structure as well [Lorenz-Fonfria, 2003]. However, the assignments done in this region are contradictory and the band is not used for secondary structure analysis. The amide B may have a complex relation to the hydrogen bonding due to fermi resonance, as described by Lazarev et al (1974): By fermi resonance, the 1st overtone (or a combination mode) of amide II borrows intensity from amide A to give rise to amide B. As the two bands come closer upon increasing hydrogen bonding, they obtain more similar energies and the fermi resonance and intensity-borrowing is increased. Thus the intensity of amide B reflects the hydrogen bonding of the NH group. The frequency of amide B is also determined from the hydrogen bonding, which shifts the band to higher frequencies [Lazarev, 1974]. In opposition to the above description, Wang et al (1998) assigned the amide B to the stretching vibration of the intramolecular hydrogen bonded NH group.

The overlap with the water stretching bands often limits the use of the amide A and B bands in MIR protein analysis.

3.4. Protein interaction analysis by MIR

Although the amide I band may act as a probe for protein three-dimensional structure in some degree, the secondary structure is the main property reflected in the amide I band, and the tertiary structure has seldom been analysed by MIR.

Detection of tertiary structure

Measurements of the native and molten globule states have provided evidence of the ability of MIR to probe tertiary structure changes. The molten globule state possesses native-like secondary structure, while the tertiary structure is disrupted. Matsuura et al (2001) observed a broadening of the amide I sub bands and decreased band resolution, when the CD40 ligand was converted into a molten globule state by lowering of the pH. This was taken as a sign of a more loosened structure. Still, these effects on the amide I band are small compared to those from secondary structure changes. Instead, hydrogen-deuterium (H-D) exchange is a more valuable method for probing the tertiary structure and its perturbations. The exposed structures are most prone to the H-D exchange, so a loosening of the tertiary structure allows for more solvent interaction and increased H-D exchange and is reflected as band-shifts in amide II [Pedone, 2003].

Specific amino acid side chain absorptions, which are influenced by interactions and the dielectric properties of their environments, may change in respond to tertiary structure changes [Jackson, 1995]. The amide groups from Gln and Asn usually cause absorptions at 1660-1690 cm^{-1}, but the frequency depends on the hydrogen bonding interactions of the C=O group. For example, the aspargine ladder, which has hydrogen bonding between Asn side chains, may lead to the Asn absorptions at 1622-1633 cm^{-1} [Khurana, 2000]. The carboxylate group from Glu and Asp causes a C=O-str. band at ~1740 cm^{-1} or ~1696-1705cm^{-1}, depending on the hydrogen bonding state. The unprotonated carboxylate group from Asp gives rise to absorptions at 1570-1598 cm^{-1} (anti-symmetric stretch), and the intensity of this band has been found a good probe for protein unfolding [Jackson, 1995]. The position and intensity of the Tyr, Trp and CH-str. absorption have also been shown to reflect their exposure and the folding state of proteins [Jackson, 1995; Liu, 2001].

Detection of protein-protein interactions

The relation between intermolecular β-sheet formation and the appearance of an amide I component at around 1620 cm^{-1} has been utilized in several studies to show the presence of soluble protein oligomers/aggregates in the native state or formed upon denaturation [Haris, 1999; Dong, 2000b; Mohney, 2000; Lefevre, 2001; Lin S., 1999; Sokolowski, 2003]. The ability of MIR to detect this structure is an advantage over e.g. CD analyses. From calculations, the band due to antiparallel intermolecular β-sheet has been shown to appear at 1624 (± 8 cm^{-1}), but bands as low as 1613 cm^{-1} has as well been ascribed to this structure [Allain, 1999; Lefevre 2001], as the strength of the intermolecular β-sheet interaction determines the exact position of the low-frequency band. Accordingly, fine stranded gels formed from heat treated BLG were found to have stronger hydrogen bonding interactions than particulate gels, since the two types of gels caused absorptions at 1613-1619 cm^{-1} and 1620-1623 cm^{-1} , respectively [Lefevre, 2001]. Involvement of intermolecular β-sheet in the dimerisation of BLG at high protein concentrations was also observed from the amide I band [Lefevre, 2001]. The dissociation of the dimer lead to the disappearance of a band at 1623 cm^{-1}, which consequently was ascribed to the β-sheet interaction involved in monomer-monomer binding [Lefevre, 2001]. Although the low-frequency bands are indicative of intermolecular-β-sheet and aggregation, other factors that increase the hydrogen bonding of CO groups may cause the appearance of these bands. These factors include the interaction with water or other molecules. Extended hydrated structures have been suggested to cause an absorption at ~1619 cm^{-1} in spectra of gluten proteins [Belton, 1995].

As the intermolecular β-sheets are formed, other secondary structures may decrease and therefore further influence the amide I band shape. For example, a decrease in the 1635 cm^{-1} band was found a good probe for detection of ovalbumin (OVA) aggregation [Dong, 2000a].

For interactions taking place between two or more different proteins, information on the structural perturbations of the individual proteins cannot be gained from conventional FTIR spectra due to the overlapping of their amide I bands. Instead, a technique called isotope edited IR, which separate the two amide I bands, can be used. See: [Fabian, 1996]. Conventional FTIR, on the other hand, can be used for demonstrating interactions between two different proteins. This is shown from the difference spectrum between the experimental spectrum and the spectrum constructed from the individual spectra.

Other amide bands are rather insensitive to the intermolecular β-sheet structure, and no distinct frequencies in the amide II band have been assigned to this structure. Studies of different proteins by [Cai, 1999] showed that β-sheet sub bands in the amide III region were less sensitive to the denaturation than α-helix bands. However, the β-sheet band at 1244 cm^{-1} was split into a band at 1247 and at 1235 cm^{-1} after denaturation of BSA [Cai, 1999].

3.5. Protein structure analysis by NIR

It could be expected that the sensitivity of the fundamental modes towards protein secondary structures exists for the overtones and combinations as well. This hypothesis has now been confirmed from several studies, but the relation between the amide combination bands and the secondary structure was not recognised until the early 1990s [Liu, 1994; Kamishikiryo-Yamashita, 1994], and the full potential of NIR in protein structure analysis has yet to be explored. In one of the first studies, Kamishikiryo-Yamashita et al (1994) considered the relative contributions from α-helix, β-sheet and random coil to the intensity of the 2170 nm band and found that α-helix contributes the double of β-sheet and random coil to this band.

Freeze-dried proteins

The majority of NIR-studies on protein structure have been carried out with freeze-dried proteins, and these have established the correspondence between protein secondary structure and band shape and position mainly in the combination band region [Robert, 1999; Miyazawa, 1998, Bai, 2005]. The advantages of measuring proteins in the solid state are the high protein signals and the absence of interfering water bands. Robert et al (1999) studied several dry proteins and used generalized canonical correlation analysis (CCA) to correlate the three data tables: NIR spectra, MIR spectra and the secondary structure contents of the proteins (reference values). Hereby, they identified NIR wavelength that represented the different secondary structures. The assignment was based on measurements of 9 proteins with dissimilar structures, and their different amino acid compositions were accounted for in the analysis. Wavelengths, at which α-helices contribute most, were identified as 2056 nm, 2172 nm, 2239 nm, 2289 nm and 2343 nm with the 2172 nm and 2289 nm being the most important. Bands at 2205 nm, 2264 nm and 2313 nm were characteristic of β-sheet, while 2265 nm was most prominent for random structure. The fact that the combination bands from amino acid side chain also reflect the secondary structure may stem from the side chains in the different structures experiencing different microenvironments. Interactions, solvation and dielectric properties influence the position and intensity of the side chain absorptions [Jackson, 1995]. For example, the OH-combination band at 2255 nm from Ser-, Tyr- and Thr- OH-groups could be affected by the conformation, as the hydrogen bonding interaction of the OH groups depends on the secondary structure. The CH overtone and combination bands could as well be influenced by the C-H..O interactions, as NIR has been found sensitive to this type of weak hydrogen bonding interaction [Katsumoto, 2002]. Also, Sefara et al (1997) observed a coordinated change in the amide B/II band and the side chain combination band at 2257 nm when BLG was unfolded and changed from α-helix to β-sheet (by solvent-denaturation). Other side chain bands (2262 and 2252 nm) changed their intensity upon denaturation, but this happened out of phase with the 2257 nm band.

Spectral changes caused by the denaturation of proteins have confirmed the sensitivity of the NIR combination bands towards the secondary structure. Bai et al (2005) observed changes in the 2nd derivative NIR spectra of freeze-dried α-chymotrypsin and cytochrome c at both 2056 nm and 2168 nm upon denaturation, and a PLSR model showed a good correlation of the NIR variations to the spectral changes in the amide I region. The same two bands were increased when sucrose was added for protection against unfolding during freeze-drying.

Miyazawa et al (1998) examined the 2nd derivative transmission spectra of dried protein films that were cast from solutions of the fibrous silk fibroin. They treated the silk fibroin with methanol and thereby induced more β-sheet structure and diminished the α-helix structure in the protein. Upon methanol treatment, the 2050 nm-band shifted from 2051 nm to 2058 nm, indicating sensitivity of this band to the secondary structure. In opposition, Liu et al (1994) found the amide A/II band to appear at 2055 nm for several globular proteins irrespective of their different secondary structures, demonstrating structural insensitivity of this band. Instead, they proposed that the 2055 nm band could detect the loosening of the hydrogen bonding strength in the structure e.g. upon denaturation. The weaker hydrogen bonds in collagen and peptides of 3_{10}-helix structure were demonstrated by a shift to lower wavelength of the amide A/II band (~2040 nm). Finally, they showed a good correlation between the amide A/II wavelength and the amide A frequency [Liu, 1994].

In the study by Miyazawa et al (1998), in which one fibrous and seven globular proteins were included, the 2180 nm band was resolved by the derivation procedure into six component bands at around 2141, 2168, 2186, 2200, 2209 and 2213 nm, of which the 2209 nm-band showed mostly in the case of proteins with a high sheet content and therefore was ascribed to this structure. Other bands were not ascribed to specific secondary structures, but it was noticed that the bands at 2141 and 2168 nm appeared for all native proteins, while bands at 2186, 2200 and 2213 nm were seen for only some of the proteins. Upon methanol treatment of silk fibroin, the 2141nm band disappeared. In conclusion from this experiment, NIR+ 2nd derivative was found to be a powerful technique for structural analysis and characterisation of proteins in the solid state.

Protein solutions

The study of protein structure-function relationships needs to be carried out in solution or in the hydrated state, such that the protein can obtain its native functional state. Furthermore, NIR-measurements of proteins in aqueous solution cannot only provide structural information but also give information on protein hydration, since NIR allows simultaneous examination of the water and the protein signal [Wu, 2000]. However, for protein solutions the absorbance from the protein is much weaker than that from water. Due to this fact and to the existence of overlapping bands, the analyses is often done by use of chemometric methods or two-dimensional correlation spectroscopy (2DCOS). In 2DCOS, enhanced spectral resolution of overlapping bands is obtained by spreading of the bands in a second dimension. The spectra are measured as a function of a property (concentration, temperature, pressure etc) and 2DCOS generates synchronous and asynchronous maps, in which peaks represent in-phase and out-of phase variations between band intensities, respectively. the advantage of 2DCOS over the chemometric methods is the provision of the asynchronous map showing uncoupled variations between bands (these are not easily observed from loading plots), and from which, the order of the spectral changes can be found [Murayama, 2000].

2D NIR-COS has been applied for analysis of the thermal denaturation processes of OVA and human serum albumin (HSA) [Wang, 1998; Ozaki, 1999; Wu, 2000]. The protein concentrations were varied from 10 to 50 mg/ml, and the concentration-dependent 2D correlation maps were studied. In the study of OVA, the correlation maps of native and heat-treated protein were compared. A spectral change was observed at 67-69^{o}C and interpreted as drastic changes in protein hydration that preceded the secondary structure changes: The in-phase variation between amide A/II and 2nd overtone of OH bending ($3\nu_2$ at 2080 nm) observed in the native state was replaced by an out-of phase variation and also a new out-of-phase variation between amide B/II and $3\nu_2$. At a somewhat higher temperature, shifts in the amide A/II and amide B/II to lower wavelengths were observed and interpreted as secondary structure changes [Wang, 1998].

In the correlation maps of native HSA, asynchronous peaks, indicative of protein hydration or structural changes, were observed [Wu, 2000]. It was suggested that aggregation and association of proteins at the higher concentrations lead to small changes in secondary structure and changes in the microenvironments of the amino acid side chains, which resulted in spectral variation in the range 2041-2381 nm. Between 58 and 60^{o}C, frequency shifts were observed for both amide B/II (at 2174 nm) and amide A/II (at 2062 nm). Amide B/II shifted from 2174 nm to 2170 nm upon denaturation of HSA [Wu, 2000]. The amide B/II combination band position is determined by the position of amide B and amide II fundamental bands, and a shift to lower wavelengths has been taken to mean a weakening of the hydrogen bonding strength [Murayama, 2002].

Other 2D NIR-COS studies have considered pH-dependent structural changes of proteins. HSA exists in a native form at pH 5.0 but is converted to an F-isoform at pH 4.3 and an extended E-isoform at pH 2.7. Thus, the correlation maps at pH 5.0, pH 3.5 and pH 2.4 were compared in a study by Ozaki [Ozaki, 1999]. When lowering the pH from 5 to 3.5, an auto peak at 2162 nm disappeared, and this was interpreted as a change in hydrogen bonding state of amide groups, when α-helix changed to β-structure or random structure in the N-F transition. At pH 2.4, the in-phase correlations between the protein peaks and $3\nu_2$ were changed to out-of phase correlations suggesting a change in hydration upon unfolding of the secondary structure at the lower pH. A 2D NIR-COS study of OVA transition from native to a molten globule-like state showed hydration and secondary structure changes to take place in parallel between pH 2.6 and 2.8. At a pH above the transition pH, a change in amino acid side chain microenvironment was indicated from disappearance of an auto peak at 2342 nm [Murayama, 2002].

Although most NIR analyses of proteins in solution have applied the 2DCOS, also difference spectroscopy and 2nd derivative analysis have been used, and Murayama et al (2000) concluded from a comparison between analysis techniques that these two methods always should be tried. The water spectrum subtraction followed by 2nd derivative calculation was for example used by Izutsu et al (2006). In a 2nd derivative analysis of BSA in solution, a temperature-dependent shift of the amide A/II band was observed as well as a splitting of the amide B/II band in two (2164 and 2176 nm), when increasing the temperature from 45 to 85°C [Yuan, 2003]. Other important information was obtained from the water bands in the difference spectra and concerned the hydration changes of BSA upon protein unfolding.

Drawing up of NIR in protein structure analysis

In conclusion from the above literature study, several bands in NIR have been recognised as sensitive to protein secondary structure and/or the denaturation state. However, the spectral changes upon alteration of secondary structure is much lower in NIR than in MIR, and the relation to secondary structure is still somewhat obscure. Likewise, studies of protein structures in solution are scarce. The following questions will be considered in the subsequent experiments and in paper III.

How much structural information is contained in the NIR spectra of protein solutions ? *(This question is investigated in Experiment III and Paper IV).*

   a. What qualitative information of secondary structure changes can be obtained?
   b. Is it possible to obtain good estimates of α-helix, β-sheet, β-turns and random coil contents?
   c. Is there any information of tertiary/quaternary structure changes?
   d. Is there any information of the intermolecular β-sheet? Or of other interactions ?

2. How much structural information is contained in the NIR spectra of proteins in a complex matrix (such as foods)? *(This question is investigated in Experiment IV and V).*
   a. Are their any unique protein absorptions ? (distinguishable from starch and lipids)
   b. Can protein conformation/interaction changes be detected?
   c. Is it possible to interpret the spectral changes and obtain structural information from NIR?
   d. Can information regarding protein-water interactions be obtained?

<u>3.6. Experiment III: NIR analysis of protein structures in solution</u>

In this experiment, the same set of standard proteins in solution is measured by both FTIR and NIR, and the assignments of NIR-wavelength regions to secondary structures are attempted by the combined FTIR-NIR analysis. Combining NIR analyses with other spectroscopic techniques (e.g. FTIR, Raman, CD) is a common method for improving the interpretation of the NIR spectra [Barton, 1996; Gouti, 1998; Maalouly, 2004; Barros, 1997]. This approach has also been used in a few structural studies of proteins [Robert, Sefara, 1997; Navea, 2003], in which the analysis methods have included e.g. 2D-COS, CCA and evolving factor analysis. Navea et al (2003) found that the combination of the corresponding spectra could not only give useful insight into the NIR region but could even bring information not possible with either method alone.

One main goal of the study was to identify possible spectral signatures of intermolecular β-sheet in NIR spectra. A denaturation study by Yuan et al (2003) of BSA in solution (by evolving factor analysis) has previously indicated an ability of NIR to detect the intermolecular β-sheet, as a turning point was seen at 71°C where this structure commonly appears.

Method

BSA, BLG, lysozyme (LYS), folate binding protein (FBP), Casein (CAS) and OVA solutions were prepared in PBS buffer of pH 7.4 at 10 mg/ml. The reference values of secondary structures as found from the literature are shown in Table 3.2 (no reliable data was found for CAS, and this protein is left out in most analyses, but however is known as a random coil protein). BLG and BSA solutions were heat-treated in water bath at 75-85°C for 30 min. NIR spectra from 790-2500 nm were measured on a Perkin Elmer, Spectrum One FT-NIR instrument, equipped with at DTGS detector and by use of a 1 mm transmission cell (quartz). The resolution was 16 cm^{-1} and 100 scans were co-added. The data interval was 1.67 nm. At least four replicates were measured for each solution. ATR-FTIR spectra were obtained on a Bomen FTIR Spectrometer from 4000 cm^{-1} to 748 cm^{-1}, at a resolution of 4 cm^{-1} and with co-addition of 128 scans. The data interval was 1.93 cm^{-1}. The instrument was continuously purged with dry air and no gas absorption was evident in the spectra. At least two replicates were obtained. No temperature control was applied in either measurement series. Preprocessing and PLSR analyses were carried out in Unscrambler 9.2 (Camo). Outer product analysis (OPA) was also carried out in Matlab. NIR spectra were preprocessed by use of standard EMSC in the 2100-2300 and 1600-1800 nm regions. Then the buffer spectra were subtracted and the spectra were EMSC corrected again before Savitzky Golay 2nd derivative calculation (with 9 or 13 data points) was carried out and the spectra were inverted. Spectral pretreatment of the ATR-FTIR spectra included Savitzky Golay smoothing followed by standard EMSC correction in the 1800-1500 cm^{-1} region and 2nd derivative transformation. The buffer spectrum was then subtracted from each spectrum, and mean normalization was carried out in the range 1700-1600 cm^{-1}.

%	α-helix	Random coil	β-sheet	References
Bovine serum albumin (BSA)	55	45	~0	[Riley, 1953]
β-lactoglobulin (BLG)	15		49	[Creamer, 1983]
Lysozyme (LYS)	45	13	19	[Robert, 1999]

Ovalbumin (OVA)	35		45	[Stein, 1991]
Folate binding protein (FBP)	22	31	30	[Kaarsholm, 1993]
Standard deviation (SD)	16	15	20	

Table 3.2. Secondary structures of the standard proteins.

Results: NIR analyses

The NIR analysis was limited to the 1600-1800 nm and the 2100-2300 nm regions, since earlier experiments had found these regions of the highest protein information, and since these regions are not overlapped much by the water bands. These regions were for example identified by interval-PLSR for protein concentration (see [Nørgaard, 2000]). Likewise, in a preliminary analysis of native and denatured BSA (25-50 mg/ml), where the EMSC corrected were submitted to a PCA, the score plots indicated that denatured BSA could be discriminated from native BSA in both spectral regions, although the 2100-2300 nm region seemed to be the most sensitive towards denaturation. However, these analyses did not allow identification of intermolecular β-sheet regions, and therefore more standard proteins were analysed, as described in the following.

The buffer subtracted and EMSC corrected NIR spectra in the 2100-2300 nm region are shown in Fig. 3.2 for the standard proteins. The EMSC treatment after the buffer subtraction seemed to remove small variations in the protein concentration.

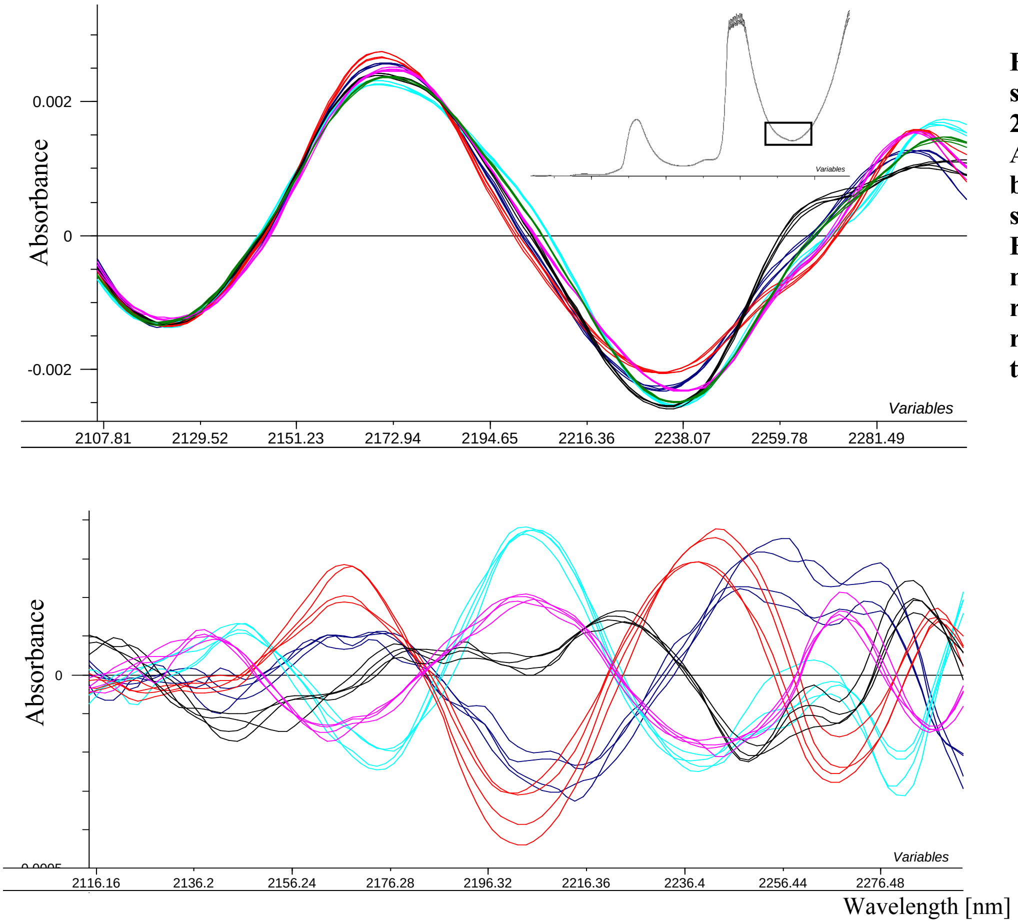


Fig. 3.2. NIR spectra of standard proteins in the 2100-2300 nm region. A) Protein spectra after buffer subtraction and subsequent EMSC. B) The same spectra after mean centering. Insert: raw spectra, showing the region as a valley between two intense water bands.

It is evident that each standard protein has its own distinct protein spectrum, although the differences are small. The differences become easier to see in the mean centred data (Fig. 3.2B). After 2nd derivative transformation of these spectra, the broad bands are seen to consist of some smaller peaks (Fig. 3.3).The large peak at 2263 nm for CAS is in agreement with the previous assignment of random coil to 2265 nm. In addition, the increase at 2209 nm and the decrease at 2166 nm upon denaturation of BSA is in agreement with the study by Izutsu et al (2006), indicating either intra- or intermolecular β-sheet formation.

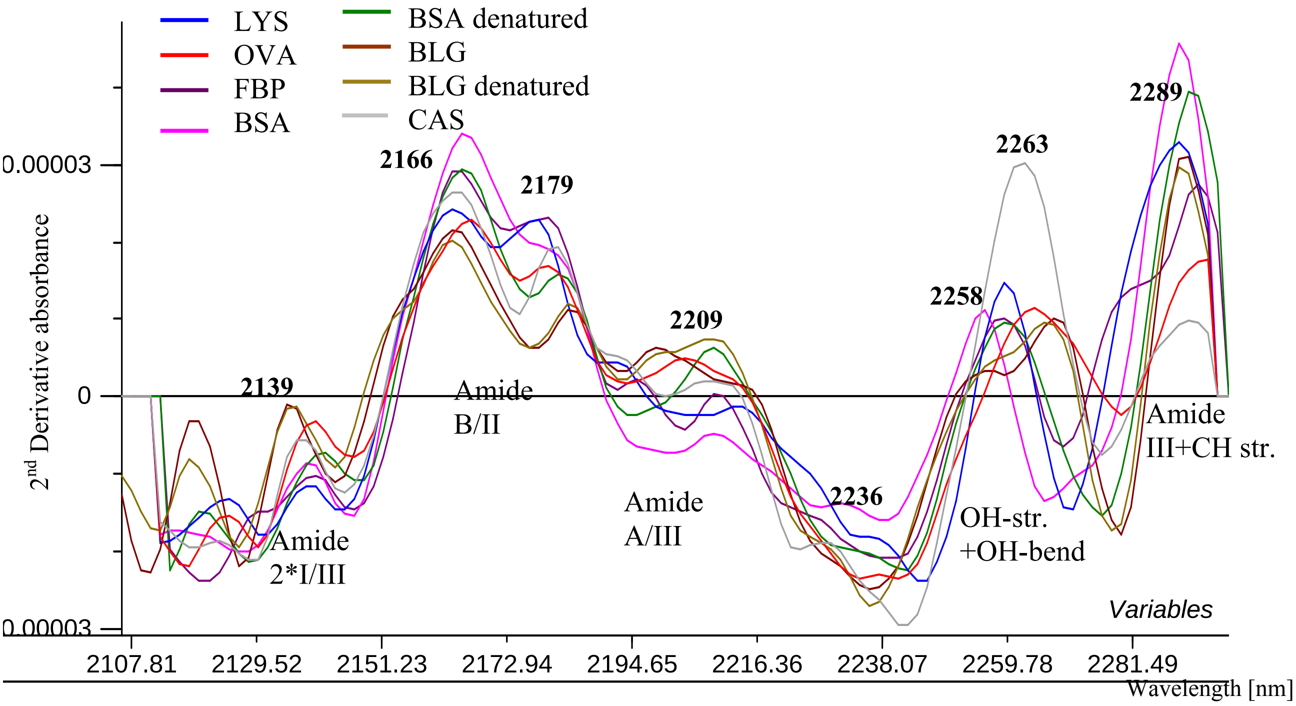


Fig. 3.3. Average NIR spectra of standard proteins and denatured proteins in the 2100-2300 nm region after preprocessing. Possible vibrational origins of the bands are shown.

In order to relate the spectral differences to secondary structure variations, the reference values in Table 3.2 were used as **Y** in two PLSR2 models, in which **X** was either the 2nd derivative transformed or the untransformed spectra. All variables were standardized, and the two models were validated by segmented CV. Only one component was found useful in the models as determined from the RMSECV(Y) plots.

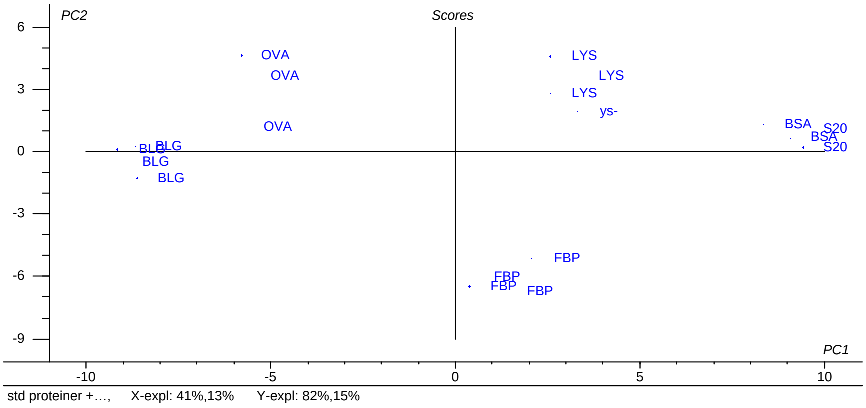


Fig 3.4. Score plot from a PLSR2 analysis (X= preprocessed 2100-2300 nm region. Y=α-helix, β-sheet contents). α-helix (bold) and β-sheet contens (normal) are shown.

The score plot from the 2nd derivative model is shown in Fig. 3.4. The use of 2nd derivative spectra resulted in a higher explanation of **Y** in PC1 (47 %) than by use of EMSC data (22 %), but neither of the models showed any predictive ability (R<0.13) for the α-helix content. However, the score plot showed a trend in PC1 that seemed related to the α-helix/β-sheet contents (Fig. 3.4) and the same was seen in a PCA score plot. Likewise, the β-sheet content was better explained (R=0.729) than the α-helix content and resulted in RMSECV(Y) of 13-15 % in the two models (compared to an original SD of 20 %). Jack-knifing was applied for identification of the significant NIR-wavelengths in the model with **X**=2nd derivative spectra. The hereby identified NIR-wavelengths with positive b-coefficients for the prediction of α-helix were the 2229-2239 and 2284-2286 nm regions, while wavelength with positive b-coefficients for β-sheet were the 2142-2146, 2206-2213 and 2268-2271 nm regions.

The 1600-1800 nm region, which contains the 1st CH-str. overtones, was examined as well by PLSR. The averaged 2nd derivative spectra are shown in Fig. 3.5. The CV results showed that spectra with and without 2nd derivative calculation resulted in models of similar predictive ability. Again, the β-sheet content was better predicted than the α-helix content as RMSECV(Y) was 14.1 % for α-helix (R= 0.48) by use of 1 PC and 12.8 % for β-sheet (R= 0.95) by use of 2 PCs. From these results it is considered that the 1600-1800 nm region also shows some structural information as regards α-helix and β-sheet contents. Only the wavelength 1785 nm had a significant positive correlation to α-helix, while the regions 1713-1728 nm and 1755-1765 nm correlated positively with β-sheet contents, as seen from jack-knifing. However, these assignments seems in contradiction with the assignments by Izutsu et al (2006): 1690 nm to β-sheet and 1738 nm to α-helix.

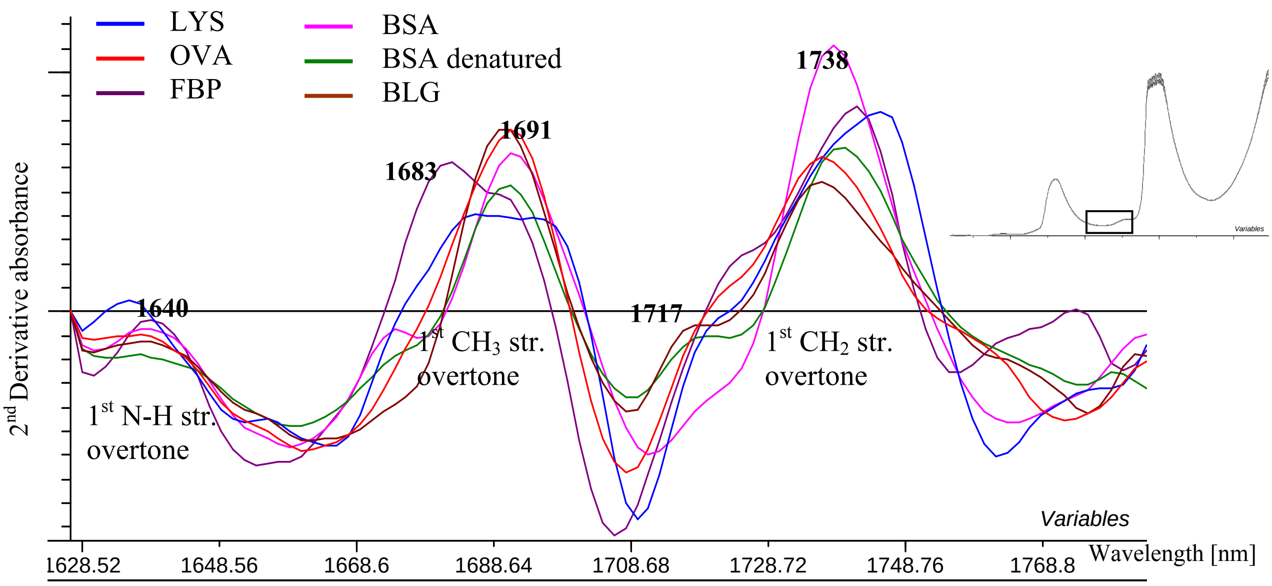


Fig. 3.5. NIR spectra of standard proteins in the 1625-1800 nm region after preprocessing. Insert: raw spectra, showing the position of the region between two water bands. Possible vibrational origins of the bands are shown.

Use of only the significant regions in the PLSR models resulted in improvement of the models based on both regions (1600-1800 and 2100-2300 nm). However, the α-helix content of OVA was predicted much lower than the reference value in both cases.

Results: MIR analysis

The amide I band was examined for its correlation to the reference values of secondary structure content.The preprocessed amide I spectra are shown in Fig. 3.6. A PLSR2 model for prediction of α-helix and β-sheet contents was made based on the preprocessed amide I band. The model was validated by segmented CV. All variables were standardized. The PLSR analysis resulted in an RMSECV of 8 % for prediction of α-helix content and an RMSECV of 5% for prediction of β-sheet content, which is better than results obtained from the NIR analyses. The regression coefficients for prediction of α-helix are shown in Fig. 3.6. The model used one PC for both predictions. A satisfactory model for prediction of turn and random structure contents could not be made. The sample set probably had too little variation in β-turn content for making a calibration for this structure. Also, the correlation between random structure and α-helix contents in the sample set made it difficult to discriminate these two structures, as there is high overlap of the two sub bands.

The intermolecular β-sheet is seen for the denatured proteins at 1620 cm^{-1}.

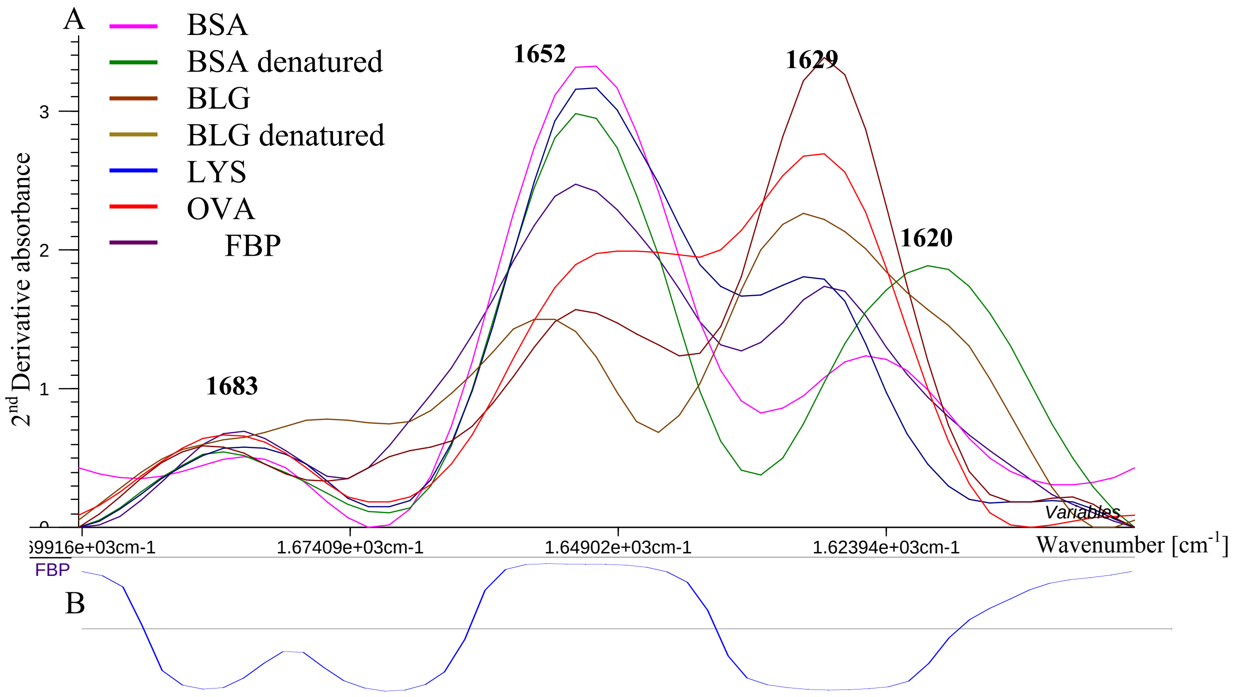


Fig. 3.6. A) Preprocessed amide I bands from ATR-FTIR measurements of standard proteins in solution. B) The b-coefficient plot for prediction of α-helix is shown below.

Correlations between MIR and NIR

The NIR spectra were correlated to the amide I spectra by PLSR2 regression. In this analysis, the spectra of denatured proteins, which have unknown secondary structure contents and also contain intermolecular β-sheet, could be included. A further advantage of this analysis was the independence of the reference data, which may contain errors (e.g. the α-helix content of OVA is predicted too low in both analyses).

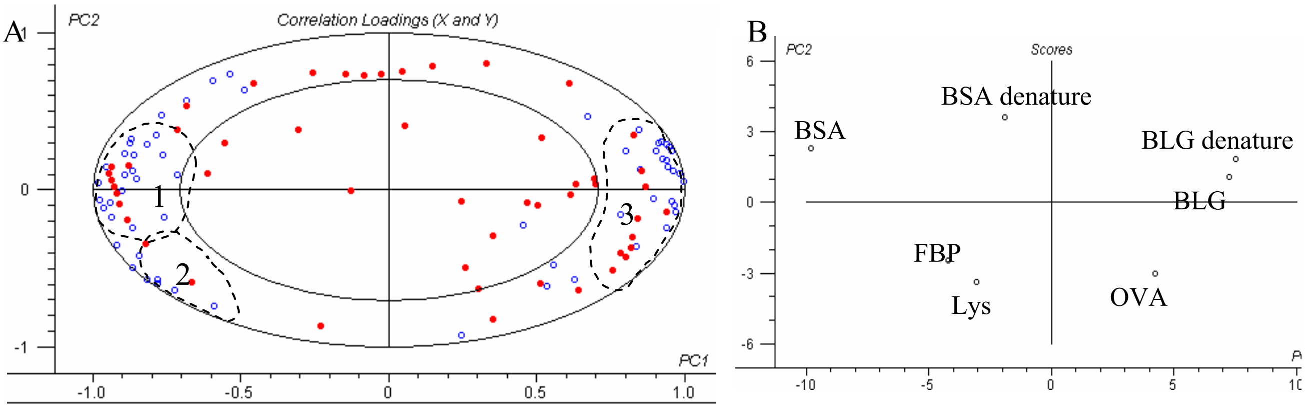


Fig. 3.7. PLSR results from a combined NIR-amide I analysis of standard proteins: X=inverted 2^{nd} derivative NIR spectra (2100-2300 nm), Y=inverted 2^{nd} derivative amide I spectra. A) Correlation loading plot of PC1 vs. PC2. The inner circle shows 50 % explanation and the outer circle shows 100 % explanation. Blue=NIR variables. Red= amide I variables. The three encircled regions (1-3) show the NIR and amide I-variables that are assigned to different secondary structures in Table 3.3. B) Score plot of PC1 vs. PC2. The two PCs explain together 21.5 % of the Y-variance.

The preprocessed and replicate-averaged NIR and amide I spectra were combined in an augmented data matrix (containing 7 samples), and a PLSR model was made for prediction of the amide I spectra on the basis of the 2100-2300 nm NIR spectra. All variables were standardized, the model was validated by use of full CV, and jack-knifing was applied for determination of the significant NIR variables. Two PCs were useful in the final model. PC1 explained 18 % of the variance in the amide I spectra (Y) and 50.84 % of the variance in the NIR spectra (X), whereas PC2 only explained 3.5 % and 9.5 % of the amide I- and NIR variance, respectively. The score plot in Fig. 3.7B indicates that PC1 is related to the α-helix and β-sheet contents, whereas PC2 may explain in some part the native vs. denatured state. The correlations between NIR and amide I variables were inspected from the correlation loading plot (PC1 vs. PC2) shown in Fig. 3.7A. The amide I and NIR variables that are close together in the correlation-loading plot are positively correlated. Thus, NIR-variables in the correlation-loading plot were classified into α-helix and β-sheet regions, based on the score plot and the known amide I assignments. The results are shown in Table 3.3 and Fig. 3.8.

Amide I wavenumbers (cm^{-1})	Assignment	NIR wavelengths (nm)	Literature [Robert, 1999]
1) 1660-1649	α-helix+random	2161-2172 2231-2239 2283-2286 (2256-2258)	2172 2239 2289

2) 1645-1643	random	2174-2182, 2278-2279	2268
3) 1691-1687 1635-1623	β-sheet	2144-2151 2196-2216 2264-2271	 2205, 2209 2264

Table 3.3. Assignments based on the PLSR model in Fig. 3.5. The numbers 1-3 correspond to the regions in the correlation loading plot in this figure. The assignments are compared to those from the literature.

Amide I wavenumbers corresponding to α-helix (+random) were correlated to three larger NIR-regions: 2161-2172, 2231-2239 and 2283-2286 nm. Only two of these NIR-regions were found significant for α-helix content in the former PLSR analysis, but all three regions agreed well with the literature values. From the present analysis, also the 2256 and 2258 nm variables could be ascribed to α-helix. The amide I frequencies at 1645-1643 cm^{-1} are representative of random structure, but due to the overlapping with the α-helix sub band, the assignment of the NIR wavelength 2174-2182 and 2278-2279 nm to random structure instead of α-helix is uncertain and do not agree with the literature values. Correlations to β-sheet were found in three NIR regions, of which only two (2196-2216 and 2264-2271 nm) are ascribed to β-sheet in the literature. The third 2144-2151 nm region seems more influenced by noise and the assignment in this region may be spurious.

The satisfactory agreement with literature values indicates that the NIR-amide I PLSR analysis is suitable for making assignments in the NIR region and can be applied for identification of possible intermolecular β-sheet absorptions in the NIR region. From Fig. 3.7, PC2 was found to explain the 1608-1620 cm^{-1} band, resulting from intermolecular β-sheet. However, this component explained very little of the NIR variance. The nearest NIR variables in the correlation-loading plot were the 2151-2152 and 2252 nm variables (Fig. 3.6A). Visual inspection of the spectra did not reveal any increase in these regions upon denaturation of BSA. Therefore, it seems difficult to use NIR for identification of intermolecular β-sheet.

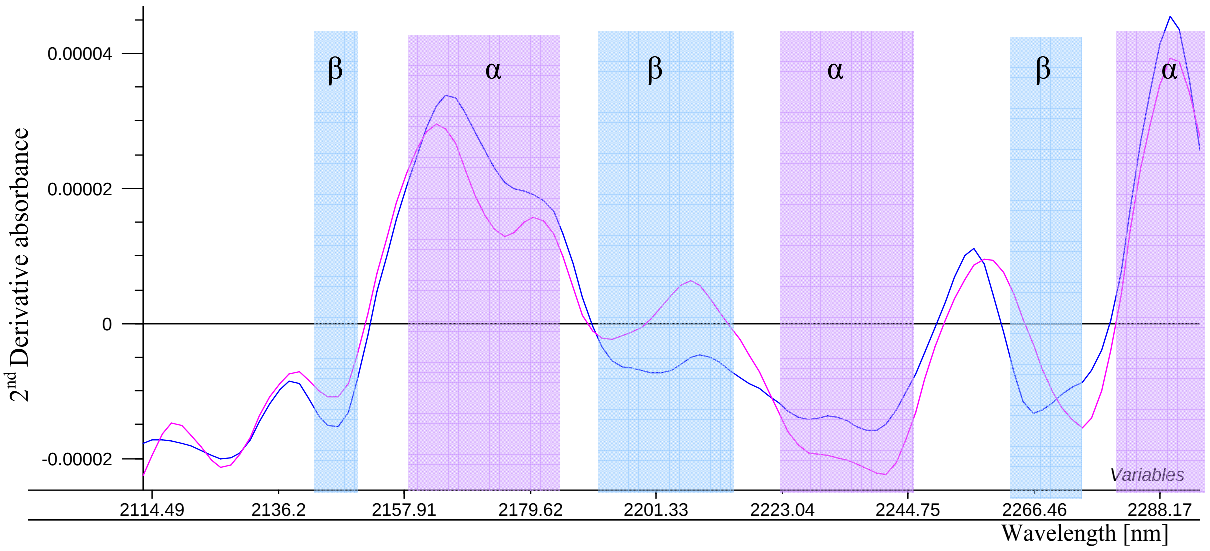


Fig. 3.8. Assignments of the NIR wavelengths in the 2100-2300 nm range to secondary structures, according to Table 3.3. Blue spectrum: BSA. Pink spectrum: BLG. Light blue=β-sheet. Purple=α-helix

The PLSR analysis was also carried out with the purpose of correlating the 1st CH-str. overtone region (1650-1750 nm) to the amide I-spectra. The explained validation variance showed that only one PC, explaining 14 % of the amide I variance, could be used, and the jack-knifing procedure found no significant NIR variables. Therefore, no assignments in this region were done.

Outer product analysis (OPA)

Another method that can combine the information in the two spectral regions (NIR and amide I) is OPA, which is described in [Maalouly, 2004]. OPA may reveal regions in NIR and MIR that increase simultaneously as a function of a property (here content of a secondary structure) and thereby highlight interesting regions in both spectra. An outer product (OP) matrix, which is calculated for every sample as the outer product of the NIR and MIR spectrum, contains all possible combinations of intensities in both the NIR and MIR signal. Values in the OP matrix are mutually weighted from NIR and MIR, meaning that two high signals reinforce each other, and a high and a low signal offset each other. The three-dimensional OP matrix is unfolded to a two-dimensional matrix, in which each row represents a sample, and this matrix is the starting point for the further analysis (e.g. PCA, PLS).

The NIR region 2100-2300 nm (109 variables) was combined with the amide I band (52 variables) in an OP matrix. The NIR and amide I spectra were preprocessed as described earlier, and before the outer product calculation, the amide I matrix was mean-centred but not the NIR matrix as this step resulted in a worse model. A PLSR analysis of the OP matrix was carried out (Y=α-helix content). No standardization was used. Three PLS components were considered useful based on the RMSECV(Y) plot (full CV was used for validation). The b-vectors were refolded to 109*52 matrices, and the NIR-profiles at the wavenumbers characteristic of β-sheet (1635 cm^{-1}) and α-helix (1654 cm^{-1}) were inspected (Fig. 3.9).

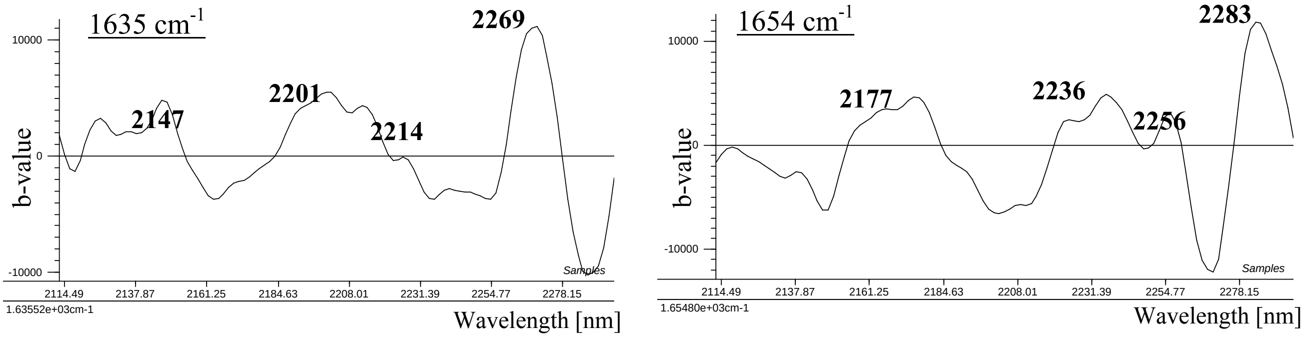


Fig. 3.9. OPA-results. NIR b-coefficient profiles at two different wavenumbers. The profile at 1635 cm^{-1} indicates β-sheet wavelengths, whereas the profile at 1654 cm^{-1} indicates α-helix wavelengths.

The two NIR-profiles in Fig. 3.9 show the peaks that can be related to α-helix and β-sheet, respectively, and the results agree with the previous findings (Table 3.3). The high values at 2269 and 2283 nm for β-sheet and α-helix, respectively, indicate that these wavelengths may be the most important for detection of β-sheet and α-helix. The NIR profiles at lower wavenumbers (1616-1620 cm^{-1}) were inspected (Fig. 3.10).The 1616 cm^{-1}

NIR-profile shows a low value at 2269 nm but a high value at 2249 nm, compared to the 1629 cm^{-1} NIR-profile. It could suggest that intermolecular β-sheet has an influence in this area, as was suggested in the PLSR analysis.

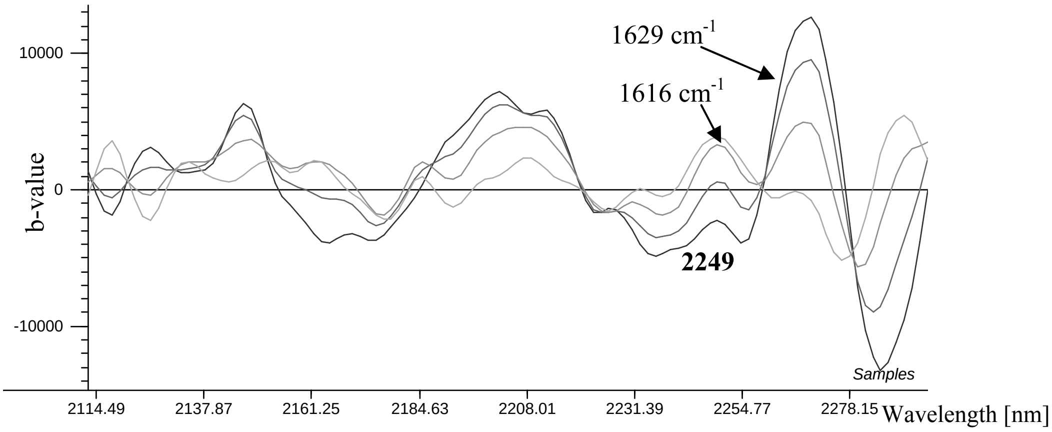


Fig. 3.10. OPA results. NIR b-coefficient profiles at different wavenumbers in the range 1616-1629 cm^{-1}.

Discussion and conclusion

From the above analyses it is possible to answer some of the previously asked questions, regarding NIR studies of protein solutions in order to obtain qualitative and quantitative secondary structure information, as well as the question of intermolecular β-sheet absorptions in NIR.

This study was limited to the two wavelength regions 1600-1800 and 2100-2300 nm, selected by the criteria of low noise and little influence from the water bands, as temperature variations may cause shift of the water bands. This prevented the analysis of the 1st overtone of N-H stretching and the amide A/II at 2056 nm but included e.g. the important amide B/II combination band at 2170 nm.

A protein concentration of 10 mg/ml was used in order to evaluate the performance of NIR at a rather low concentration, and this is the lower limit used in other NIR studies of protein structure [Wu, 2000; Murayama, 2002]. A detailed analysis of random structure, α-helix, β-turns, and parallel/antiparallel β-sheet contents, was not possible due to the small sample set, and only β-sheet and α-helix contents were considered. A sensitivity in selected NIR regions to the α-helix and β-sheet ratios was indicated from PLSR and PCA plots, and it was found that some information of secondary structure changes could be provided from inspection of the 2nd derivative spectra without application of a chemometric method. For example, the denaturation of BSA and BLG caused visible spectral changes.

The PLSR analysis based on the preprocessed NIR spectra did not show good correlation to α-helix and β-sheet contents, and this suggests that NIR spectra are not very suitable for quantitative analyses of secondary structures (at the applied protein concentration). However, the poor results may also be due to the small sample set (with too little variation) and some mistaken reference values. Instead, a combined NIR-amide I PLSR analysis was applied for making assignments of β-sheet and α-helix in the NIR spectra, and the found assignments agreed well with literature values. Therefore, a capacity of NIR for secondary structure analysis of proteins also in solution seems to be established.

However, the amide I analysis seemed to work better than NIR for a quantitative analysis of secondary structures, even though the bands ascribed to α-helix and β-sheet appear to be better resolved in NIR than in the amide I band. The NIR sub bands are not attributed to a single structure, and the bands ascribed to e.g. α-helix appear also for proteins with no α-helix. Also, in the side chain combination band region, the different amino acid compositions of different proteins may contribute significantly to the variation. These factors hampers the quantitative NIR analysis. Therefore, it seems that a multivariate calibration model with many more samples is needed for a proper quantitative prediction of secondary structures in different proteins, if possible.

No obvious signature of intermolecular β-sheet could be found in the analysed NIR regions from this method, but some variables (around 2250 nm) of low correlation to this structure were high-lighted and could be subject for further investigation. Thus, NIR seems to have much less potential for detecting and characterising intermolecular β-sheets than amide I analyses.

The performance of an NIR calibration and the detection limit for secondary structure changes both depend on the SNR. The applied instrument was equipped with a DTGS detector, which does not provide NIR with the optimal sensitivity, and an instrument equipped with another detector may therefore lead to an improved analysis. Also, the use of a higher protein concentration could increase the SNR. As the SNR is influenced by the absorbance level of the water, it can as well be optimised by adjusting the path length to give the optimal absorbance level of ~0.4 AU [Jensen, 2002; Isaksson, 2002]. In the 2100-2300 nm region, this can be achieved by using a cuvette with at path length of 0.4-0.5 mm. However, problems with temperature variations usually become more significant when using the very thin cuvettes. Another approach is to choose a wavelength region of the optimal absorbance level. Thus, it is advantageous to be able to use the CH-str. overtone region in the structure analysis, since the absorbencies in the 1600-1800 nm region are between 0.25 and 0.44 AU using the 1 mm cuvette. Although the 1600-1800 nm region was found somewhat sensitive to secondary structures, a low correlation to the amide I variations indicated that the use of this region only would not result in a proper structural analysis.

The above discussion on the SNR is relevant for transmission measurements, whereas the food samples often are measured in reflectance mode. For these samples, the light scattering governs the absorbance level. In addition, the presence of other absorbing constituents influence the spectra and complicate their analyses.

3.6. References

Alexander, M., Corredig, M., Dalgleish, D.G. (2006). Diffusing wave spectroscopy of gelling food systems: The importance of the photon transport mean free path (l*) parameter. *Food Hydrocolloid.* 20, 325-331.

Allain, A.F., Paquin, P., Subirade, M. (1999). Relationships between conformation of beta-lactoglobulin in solution and gel states as revealed by attenuated total reflection Fourier transform infrared spectroscopy. *Int. J. Biol. Macromol.* 26, 337-344.

Andersen, C.A.F. (2001). Protein structure and the diversity of hydrogen bonds. PhD-thesis. Technical University of Denmark, Kgs. Lyngby.

Bai, S.J., Nayar, R., Carpenter, J.F., Manning, M.C. (2005). Noninvasive determination of protein conformation in the solid state using near infrared (NIR) spectroscopy. *J. Pharm. Sci.* 94, 2030-2038.

Barros, A.S., Safar, M., Devaux, M.F., Robert, P., Bertrand, D., Rutledge, D.N. (1997). Relations between mid-infrared and near-infrared spectra detected by analysis of variance of an intervariable data matrix. *Appl. Spectrosc.* 51, 1384-1393.

Barth, A. (2000). The infrared absorption of amino acid side chains. *Prog. Biophys. Mol. Bio.* 74, 141-173.

Barth, A., Zscherp, C. (2002). What vibrations tell us about proteins. *Q. Rev. Biophys.* 35, 369-430.

Barton II, F. E., Himmelsbach, D. S., Archibald, D.D. (1996). Two-dimensional vibration spectroscopy. V: Correlation of mid- and near infrared of hard red winter and spring wheats. *J. Near infared. Spec.* 4, 139-152.

Brandl, M., Weiss, M.S., Jabs, A., Sühnel, Hilgenfeld, R. (2001). CH…π-interactions in Proteins. *J. Mol. Biol.* 307, 357-377

Brauner, J.W., Flach, C.R., Mendelsohn, R., J. (2005). A quantitative reconstruction of the amide I contour in the IR spectra of globular proteins: from structure to spectrum. *J. Am. Chem. Soc.* 127, 100-109.

Borgrah, A., Carpentier, R., Tajmir-Riahi, H.A. (1999). The effect of cholesterol on the solution structure of proteins of photosystem II. Protein secondary structure and photosynthetic oxygen evolution. *J. Colloid. Interf. Sci.* 210, 118-122.

Cai, S.W., Singh, B.R. (1999). Identification of beta-turn and random coil amide III infrared bands for secondary structure estimation of proteins. *Biophys. Chem.* 80, 7-20.

Cheatum, C.M., Tokmakoff, A., Knoester, J. (2004). Signatures of beta-sheet secondary structures in linear and two-dimensional infrared spectroscopy. *J. Chem. Phys.* 120, 8201-8215.

Creamer, L.K., Parry, D.A.D., Malcolm, G.N. (1983). Secondary structure of bovine β-lactoglobulin-B. *Arch. Biochem. Biophys.* 227, 98-105.

Czarnik-Matusewicz, B., Murayama, K., Tsenkova, R., Ozaki, Y. (1999). Analysis of near-infrared spectra of complicated biological fluids by two-dimensional correlation spectroscopy: Protein and fat concentration-dependent spectral changes of milk. *Appl. Spectrosc.* 53, 1582-1594.

Darby, N. J., Creighton, T.E. (1993). Protein structure. IRL press, Oxford university Press. Oxford.

Domenek, S., Morel, M.H., Bonicel, J., Guilbert, S. (2002). Polymerization kinetics of wheat gluten upon thermosetting. A mechanistic model. *J. Agric. Food Chem.* 50, 5947-5954.

Dong, A.C., Kendrick, B., Kreilgard, L., Matsuura, J., Manning, M.C., Carpenter, J.F. (1997). Spectroscopic study of secondary structure and thermal denaturation of recombinant human factor XIII in aqueous solution. *Arch. Biochem. Biophys.* 347, 213-220.

Dong, A.C., Meyer, J.D., Brown, J.L., Manning, M.C., Carpenter, J.F. (2000a). Comparative Fourier transform infrared and circular dichroism spectroscopic analysis of alpha(1)-proteinase inhibitor and ovalbumin in aqueous solution. *Arch. Biochem. Biophys.* 383, 148-155.

Dong, A.C., Randolph, T.W., Carpenter, J.F. (2000b). Entrapping intermediates of thermal aggregation in alpha-helical proteins with low concentration of guanidine hydrochloride. *J. Biol. Chem.* 275, 27689-27693.

Dou, Y.M., Baisnee, P.F., Pecout, Y., Nowick, J., Baldi, P. (2004). ICBS: a database of interactions between protein chains mediated by β-sheet formation. *Bioinformatics*. 20, 2767-2777.

Elliott, A., Ambrose, E.J. (1950). Structure of synthetic polypeptides. *Nature*. 165, 921-922.

Fabian, H., Chapman, D., Mantsch, H.H. (1996). New trends in isotope-edited infrared spectroscopy. In: *Infrared Spectroscopy of Biomolecules,* (Mantsch, H.H., Chapman, D., eds.). Wiley-Liss. pp. 341-351.

Fang, Y., Dalgleish, D.G. (1997). Conformation of beta-lactoglobulin studied by FTIR: Effect of pH, temperature, and adsorption to the oil-water interface. *J. Colloid. Interf. Sci.* 196, 292-298.

Fu, F.N., Deoliveira, D.B., Trumble, W.R., Sarkar, H.K., Singh, B.R. (1994). Secondary structure estimation of proteins using the amide III region of fourier-transform infrared-spectroscopy – application to analyze calcium binding-induced structural-changes in calsequestrin. *Appl. Spectrosc*. 48, 1432-1441.

Fändrich, M. (2003). Structure and formation of amyloid fibrils. *Acta. Histochem*. 105, 379-379.

Gezimati, J., Sing, H., Creamer, L. K. (1996). Aggregation and gelation of bovine β-lactoglobulin, α-lactalbumin, and serum albumin. In: *Macromolecular Interactions in Food Technology*, (Parris, N., Kato, A., Creamer, L.K., Pearce, J., eds.). Am. Chem. Soc., Washington DC. pp. 113-123.

Gouti, N., Rutledge, D.N., Feinberg, M.H. (1998). Factorial correspondence regression applied to multi-way spectral data. *Analyst*. 123, 1783-1790.

Griebenow, K., Klibanov, A.M. (1995). Lyophilization-induced reversible changes in the secondary structure of proteins. *Proc. Natl. Acad. Sci. USA*. 92, 10969-10976.

Haris, P.I., Seveercan, F. (1999). FTIR spectroscopic characterization of protein structure in aqueous and non-aqueous media. J. *Mol. Catal. B: Enzym*. 7, 207-221.

Holly, S., Egyed, O., Jalsovszky, G. (1992). Assignment problems of amino acids. Dipeptides and tripeptides and proteins in the near-infrared region. *Spectrochim. Acta A*. 48, 101-109.

Horton, H.R., Moran, L.A., Ochs, R.S., Rawn, D.J., Gray, K., Scrimgeour, K.G. (2002). Three-dimensional structure of proteins. In: *Principles of Biochemistry,* (Horton, H.R., ed). 3rd ed, Prentice Hall, New Jersey, chapter 4.

Isaksson, T., Griffith, P.R. (2002). Optimal absorbance for transmission or reflection spectra measured under conditions of constant detector noise in the presence of stray radiation. *Appl. Spectrosc*. 56, 916-919.

Izutsu, K-I., Fujimaki, Y., Kuwabara, A., Hiyama, Y., Yomota, C., Aoyagi, N. (2006). Near-infrared analysis of protein secondary structure in aqueous solutions and freeze-dried solids. J. Pharm. Sci. 95, 781-789.

Jahn, T.R., Radford, S.E. (2005). The Yin and Yang of protein folding (mini review). *FEBS*. 272, 5962-5970.

Jackson, M., Mantsch, H.H. (1995). The use and misuse of FTIR Spectroscopy in the determination of protein-structure. *Crit. Rev. Biochem. Mol. Biol.* 30, 95-120.

Jensen, P.S., Bak, J. (2002). Near-Infrared Transmission Spectroscopy of Aqueous Solutions: The influence of optical pathlength on signal-to-noise ratio. *Appl. Spectrosc*. 56,1600-1606.

Kaarsholm, N.C., Kolstrup, A.M., Danielsen, S.E., Holm, J., Hansen, S.I. (1993). Ligand-induced conformation change in folate binding protein. *Biochem J.* 292, 921-925.

Kamishikiryo-yamashita, H., Tatara, M., Takamura, H., Matoba, T. (1994). Effect of secondary structures of protein on determination of protein-content by near-infrared spectroscopy. *J. Jap. Soc. Food Sci. Technol.* 41, 65-69.

Katsumoto, Y., Adachi, D., Sato, H., Ozaki, Y. (2002). Usefulness of a curve fitting method in the analysis of overlapping overtones and combinations of CH stretching modes. *J. Near infrared spec.* 10, 85-91.

Khurana, R., Fink, A.L. (2000). Do parallel beta-helix proteins have a unique Fourier transform infrared spectrum? *Biophys. J.*, 78, 994-1000.

Kimura, S.R, Brower, R.C., Vajda, S., Camacho, C.J. (2001). Dynamical view of the positions of key side chains in protein-protein binding. *Biophys. J.* 80, 635-642.

Krimm, S., Bandekar, J. (1986). Vibrational spectroscopy and conformation of peptides, polypeptides, and proteins. *Adv. Prot. Chem.* 38, 181-364.

Lacey, D.J., Wellner, N., Beaudoin, F., Napier, J.A., Shewry, P. R. (1998). Secondary structure of oleosins in oil bodies isolated from seeds of safflower (Carthamus tinctorius L.) and sunflower (Helianthus annuus L.). *Biochem. J.* 334. 469-477.

Lazarev, Y.A. (1974). Action of fermi resonance on position and intensity of bands of amide-A and Amide-B in infrared-spectra of polypeptides and proteins. *Biofizika.* 19, 622-625.

Lefevre, T., Subirade, M. (2001). Molecular structure and interaction of biopolymers as viewed by Fourier transform infrared spectroscopy: model studies on beta-lactoglobulin. *Food Hydrocolloid.* 15, 365-376.

Lefevre, T., Subirade, M. (2003). Formation of intermolecular beta-sheet structures: a phenomenon relevant to protein film structure at oil-water interfaces of emulsions. *J. Colloid Interf. Sci.* 263, 59-67.

Li-Chan, E.C.Y. (1996). Macromolecular interactions of food proteins studied by Raman spectroscopy. In: *Macromolecular Interactions in Food Technology*, (Parris, N., Kato, A., Creamer, L.K., Pearce, J., eds.). Am. Chem. Soc., Washington DC. pp. 15-36.

Lin, S.Y., Li, R.J., Ho, C.J. (1999). pH-dependent secondary conformation of bovine lens alpha-crystallin: ATR infrared spectroscopic study with second-derivative analysis. *Graefes Arch. Clin. Exp.* 237, 157-160.

Liu, K.Z., Mantsch, H.H. (2001). Apoptosis-induced structural changes in leukemia cells identified by IR spectroscopy. *J. Mol. Struct.* 565, 299-304.

Liu, Y.L., Cho, R.K., Sakurai, K., Miura, T., Ozaki, Y. (1994). Studies on Spectra-Structure Correlations in Near-Infrared Spectra of Proteins and Polypeptides. 1. A Marker Band for Hydrogen-Bonds. *Appl. Spectrosc.* 48, 1249-1254.

Lorenz-Fonfria, V.A., Villaverde, J., Trezeguet, W., Lauquin, G.J.M., Brandolin, G., Padros, E. (2003) Structural and functional implications of the instability of the ADP/ATP transporter purified from mitochondria as revealed by FTIR spectroscopy. *Biophys. J.* 85, 255-266.

Maeda, Y., Hiuchi, T., Ikeda, I. (2000) Change in hydration state during the coil-globule transition of aqueous solutions of poly(N-isopropylacrylamide) as evidenced by FTIR spectroscopy. *Langmuir.* 16, 7503-7509.

Maalouly, J., Eveleigh, L., Rutledge, D.N., Ducauze, C.J. (2004). Application of 2D correlation spectroscopy and outer product analysis to infrared spectra of sugar beets. *Vib. Spectrosc.* 36, 279-285.

Mantsch, H.H., Perczel, A., Hollosi, M., Fasman, G.D. (1993), Characterization of beta-turns in cyclic hexapeptides in solution by fourier-transform IR spectroscopy. *Bioplymers.* 33, 201-207.

Marechal, Y., Interaction configurations of H_2O molecules in a protein (Stratum Corneum) by infrared spectrometry (1997). *J. Mol. Struct.* 416, 133-143.

Miyazawa, M. (1998). Second derivative near infrared studies on the structural characterisation of proteins. *J. Near Infrared Spec.* 6, 253-257.

Matsuura, J.E., Morris, A.E., Ketchem, R.R., Braswell, E.H., Klinke, R., Gombotz, W.R., Remmele, R.L. (2001). Biophysical characterization of a soluble CD40 ligand (CD154) coiled-coil trimer: Evidence of a reversible acid-denatured molten globule. *Arch. Biochem. Biophys.* 392, 208-218.

Mohney, B.K., Petri, E.T., Uvarova, V., Walker, G.C. (2000). Infrared absorption and ultraviolet-circular dichroism spectral studies of thermally induced unfolding of apomyoglobin. *Appl. Spectrosc.* 54,9-14.

Moran, A., Mukamel, S. (2004). The origin of vibrational mode couplings in various secondary structural motifs of polypeptides. *Proc. Natl. Acad. Sci. USA.* 101, 506-510.

Murayama, K., Yamada, K., Tsenkova, R., Wang, Y., Ozaki, Y. (1998). Determination of human serum albumin and γ-globulin in a control serum solution by near-infrared spectroscopy and partial least squares regression. *Fresen. J. Anal. Chem.* 362, 155-161.

Murayama, K., Czarnik-Matusewicz, B., Wu, Y.Q., Tsenkova, R., Ozaki, Y. (2000). Comparison between conventional spectral analysis methods, chemometrics, and two-dimensional correlation spectroscopy in the analysis of near-infrared spectra of protein. *Appl. Spectrosc.* 54, 978-985.

Murayama, K., Wu, Y.Q., Czarnik-Matusewicz, B., Ozaki, Y. (2001). Two-dimensional/attenuated total reflection infrared correlation spectroscopy studies on secondary structural changes in human serum albumin in aqueous solutions: pH-dependent structural changes in the secondary structures and in the hydrogen bondings of side chains. *J. Phys.Chem. B.* 105, 4763-4769.

Murayama, K., Ozaki, Y. (2002). Two-dimensional near-IR correlation spectroscopy study of molten globule-like state of ovalbumin in acidic pH region: Simultaneous changes in hydration and secondary structure. *Biopolymers.* 67, 394-405.

Navea, S., de Juan, A., Tauler, R. (2003). Modeling temperature-dependent protein structural transitions by combined near-IR and mid-IR spectroscopies and multivariate curve resolution. *Anal.Chem.* 75, 5592-5601.

Neault, J.F., Tajmir-Riahi, H.A. (1998). Interaction of cisplatin with human serum albumin. Drug binding mode and protein secondary structure. *BBA-Protein Struct. M.* 1384, 153-159.

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. *Appl. Spectrosc.* 54, 413-419.

Oberg, K.A., Ruysschaert, J.M., Goormaghtigh, E. (2004). The optimization of protein secondary structure determination with infrared and circular dichroism spectra. *Eur. J. Biochem.* 271, 2937-2948.

Ofran, Y., Rost, B. (2003). Analysing six types of protein-protein interfaces. *J. Mol. Biol.* 325, 377-387.

Ozaki, Y., Murayama, K., Wang, Y. (1999). Application of two-dimensional near-infrared correlation spectroscopy to protein research. *Vib. Spectrosc.* 20, 127-132.

Perez, C., Griebenow, K. (2000). Fourier-transform infrared spectroscopic investigation of the thermal denaturation of hen egg-white lysozyme dissolved in aqueous buffer and glycerol. *Biotechnol. Lett.* 22, 1899-1905.

Pedone, E., Bartolucci, S., Rossi, M., Pierfederici, F.M., Scire, A., Cacciamani, T., Tanfani, F. (2003). Structural and thermal stability analysis of *Escherichia coli* and *Alicyclobacillus acidocaldarius* thioredoxin revealed a molten globule-like state in thermal denaturation pathway of the proteins: an infrared spectroscopic study. *Biochem. J.* 373, 875-883.

Petsev, D.N., Thomas, B.R., Yau, S.T., Vekilov, P. G. (2000). Interactions and aggregation of apoferritin molecules in solution: effect of added electrolytes. *Biophys. J.* 78, 2060-2069.

Riley, D.P., Arndt, U.W. (1953). X-ray scattering by some native and denatured proteins in the solid state. *P. Roy. Soc. Lond. B Biol.* 141, 93-97

Robert, P., Devaux, M.F., Mouhous, N., Dufour, E. (1999). Monitoring the secondary structure of proteins by near-infrared spectroscopy. *Appl. Spectrosc.* 53, 226-232.

Sadler, A.J., Horsch, J.G., Lawson, E.Q., Harmatz, D., Brandau, D.T., Middaugh, C.R. (1984). Near-infrared photoacoustic-spectroscopy of proteins. *Anal. Biochem.* 138, 44-51.

Sakurai K, Goto, Y. (2002). Manipulating monomer-dimer equilibrium of bovine β-lactoglobulin by amino acid substitution. *J. Biol. Chem.* 277, 25735-25740.

Šašić, Š., Ozaki, Y. (2000). Band assignment of near-infrared spectra of milk by use of partial least-squares regression. *Appl. Spectrosc.* 54, 1327-1338.

Sefara, N.L., Magtoto, N.P., Richardson, H.H. (1997). Structural characterization of β-lactoglobulin in solution using two-dimensional FT mid-infrared and FT near-infrared correlation spectroscopy. *Appl. Spectrosc.* 51, 536-540.

Shukia, A., Guptasarma, P. (2004). Folding of β/α-unit scrambled forms of *S. cerevisiae* triosephosphate isomerase: evidence for autonomy of substructure formation and plasticity of hydrophobic and hydrogen bonding interactions n core of β/α-barrel. *Struct. Funct. Bioinform.* 55, 548-557.

Sokolowski, F., Modler, A.J., Masuch, R., Zirwer, D., Baier, M., Lutsch, G., Moss, D.A., Naumann, D. (2003). Formation of critical oligomers is a key event during conformational transition of recombinant Syrian hamster prion protein. *J. Biol. Chem.* 278, 40081-40492.

Srisailam, S., Kumar, T.K.S., Srimathi, T., Yu, C. (2002). Influence of backbone conformation on protein aggregation. *J. Am. Chem. Soc.* 124, 1884-1888.

Stein, P.E., Leslie, A.G.W., Finch, J.T., Carrell, R.W. (1991). Crystal-structure of uncleaved ovalbumin at 1.95 A resolution. *J. Mol. Biol.* 221, 941-959.

Torii, H., Tasumi, M. (1992). Model-calculatoins on the amide-I infrared bands of globular-proteins. *J. Chem. Phys.* 96, 3379-3387.

Tolstoguzov, V. (2003). Some thermodynamic considerations in food formulation. *Food Hydrocolloid.* 17, 1-23.

Walsh, S.T.R, Cheng, R.P., Wright, W.W., Alonso, D.O.V., Daggett, V., Vanderkooi, J.M., DeGrado, W.F. (2003). The hydration of amides in helices; a comprehensive picture from molecular dynamics, IR, and NMR. *Protein Sci.* 12, 520-531.

Wang, J., Sowa, M., Ahmad, K., Mantsch, H.H. (1994). Photoacoustic near-infrared investigation of homo-polypeptides. *J. Phys. Chem.* 98, 4748-4755.

Wang, Y., Murayama, K., Myojo, Y., Tsenkova, R., Hayashi, N., Ozaki, Y. (1998). Two-dimensional Fourier transform near-infrared spectroscopy study of heat denaturation of ovalbumin in aqueous solutions. *J. Phys. Chem. B.* 102, 6655-6662.

Wang, Y., Murayama, K., Myojo, Y., Tsenkova, R., Hayashi, N., Ozaki, Y. (1998). Two-dimensional Fourier transform near-infrared spectroscopy study of heat denaturation of ovalbumin in aqueous solutions. *J. Phys. Chem. B.* 102, 6655-6662.

Wellner, N., Mills, E.N.C., Browsney, G., Wilson, R.H., Brown, N., Freeman, J., Halford, N.G., Shewry, P. R., Belton, P.S. (2005). Changes in protein secondary structure during gluten deformation studied by dynamic Fourier transform infrared spectroscopy. *Biomacromol.* 6, 255-261.

Wu, Y.Q., Czarnik-Matusewicz, B., Murayama, K., Ozaki, Y. (2000). Two-dimensional near-infrared spectroscopy study of human serum albumin in aqueous solutions: Using overtones and combination modes to monitor temperature-dependent changes in the secondary structure. *J.Phys. Chem. B.* 104, 5840-5847.

Yamashita, H., Takamura, H., Matoba, T (1994). Effect of non-peptide and non-protein nitrogen compounds for the determination of protein content by near infrared spectroscopy. *J. Near Infrared Spec.* 2, 145–151.

Yuan, B., Muayama, K., Wu, Y.Q., Tsenkova, R., Dou, X.M., Era, S., Ozaki, Y. (2003). Temperature-dependent near-infrared spectra of bovine serum albumin in aqueous solutions: Spectral analysis by principal component analysis and evolving factor analysis. *Appl. Spectrosc.* 57, 1223-1229.

Chapter 4: Study of gluten interactions and functionality part 1: Hydration.

In this and the following chapter, the protein structural and interaction analysis by NIR is attempted for proteins in a complex matrix, which contains several constituents, and for which the particulate nature gives rise to light scattering. The system under consideration is the gluten-water system, which has a particular functional role in bread-making. Therefore, these analyses also serve to give insight into the performance of NIR in a functionality study.

This chapter has focus on the gluten-network development. In bread-making, the essential cohesive and viscoelastic properties of the dough are developed when the wheat gluten proteins hydrate and interact during the mixing. An interest in gluten protein structure and the physicochemical basis of gluten viscoelasticity has emerged in the past decade, and several models of the gluten network have been suggested e.g. with the purpose of explaining its viscoelastic behaviour [Schofield, 1996; Lindsay, 1999; Belton, 1999; Wellner, 2005]. This topic has interest not only in relation to the baking process but also for the processing of gluten into biomaterials/bioplastics [Irrisin-Mangata, 2001]. Still, the gluten structure and its relation to bread-making quality remain somewhat obscure.

4.1. Gluten proteins and relation to baking quality

The typical protein content in flour is ~10 %, and most of the proteins belong to the water insoluble gluten proteins [Schofield, 1996].

Protein composition

Gluten is comprised of more than 50 different proteins, which are mainly divided into gliadins and glutenins [Schofield, 1996]. Gliadins account for ~50 % of the gluten proteins and are divided into α-, β- γ and ω-gliadins of different structures and properties. Glutenins comprise the Low Molecular Weight glutenin subunits (LMW-GS), which constitute ~40 % of the gluten proteins, and the High Molecular Weight glutenin subunits (HMW-GS), which constitute the remaining 10 % [Schofield, 1996; Belderok, 2000a]. Gliadins are monomeric proteins and most of them have a globular conformation. On the other hand, glutenins exist as large polymers (formed by intermolecular S-S bridges) and they have a more extended conformation [Schofield, 1996]. A part of the glutenin fraction is called Glutenin Macropolymer (GMP) and it can be isolated as an SDS-insoluble layer, containing the largest glutenin polymers. A recently suggested model of GMP outlines the HMW-GS as assembled into linear polymers, on which the LMW-GS constitute the branches [Lindsay, 1999]. The HMW-GS also promote increased polymer size, and they are believed to be mainly responsible for the elasticity of dough [Hamer, 1998; Wrigley, 1988].

The amino acid sequences of gliadins and glutenins are closely related. Both gluten proteins are abundant in Pro and Gln and have therefore been named prolamins [Schofield, 1996]. A characteristic of the prolamins is the presence of repetitive domains, with repeat motives based mainly on Pro, Gln and hydrophobic amino acids in gliadins and LMW-GS and on Tyr, Gln, Gly, Tyr, Ser, Leu in the HMW-GS [Schofield, 1996]. In solution, the repetitive domains of some glutenins (especially HMW-GS) and gliadins have been suggested to adopt the structure of a loose β-spiral based on repeated β-reverse turns, whereas the flanking C- and N-terminal domains of both glutenins and gliadins have demonstrated random, β-sheet and α-helix structure [Veraverbeke, 2002; Schofield, 1996]. Cysteins are found in the non-repetitive domains and form either intramolecular or intermolecular S-S bridges in gliadins and glutenins, respectively (only the ω-gliadins lack cystein residues) [Schoefield 1996; Veraverbeke, 2002].

Gluten baking quality

Gluten protein structure and interactions are of concern in the baking process, since these factors can ultimately be correlated with the baking result. The baking result is evaluated by parameters such as bread volume and crumb texture: A large bread volume, and usually a soft, uniform and fine textured crumb with thin cell walls is demanded. However, as the crumb quality and bread volume are governed by different properties of the dough, the optimal bread may involve a compromise between the two [Alava, 2001]. Also, the properties of the unbaked dough can give an indication of the baking quality [Tipples, 1996] and may be determined from various rheological measurements and dough testing methods (e.g. mixograph, alveograph, extensograph tests). These tests provide measures of the viscosity, extensibility and elasticity etc. A common measure from the dough testing methods is the dough strength, which actually reflects both strength, extensibility and viscosity [Edwards, 2001]. The better the extensibility and elasticity, the better is the capacity of the dough to retain the carbon dioxide. However, for optimal baking quality, a balance between the extensibility and elasticity is required [Veraverbeke, 2002].

The baking property of a flour is affected mutually by the concentration and quality of the gluten proteins in the flour. The quality of gluten in a cultivar is reflected in the slope from the plot of bread volume against protein concentration [Schofield, 1996]. A ‘strong’ flour results in a high slope and produces optimal bread at medium protein concentrations, while ‘weak’ flours result in lower slopes and need high protein concentrations to obtain the same result [Seabourn, 2002].

That cultivars exhibit varying gluten qualities relates mainly to their different HMW-GS/LMW-GS and glutenin/gliadin ratios and to their different HMW-GS compositions, as some HMW-GS subunits have been associated with good baking performance and others with poor baking performance [Schofield, 1996; Belderok, 2000b; Tronsmo, 2003]. The subunit composition may ultimately determine the ability of glutenins to aggregate, and this ability correlates well with the bread-making quality of the flour [Veraverbeke, 2002]. Thus, the glutenin polymer size distribution is found to be an important factor to the functionality of gluten. Long glutenin polymers cause doughs of high strength and elasticity, and also, the amount of GMP in the dough is positively correlated to dough strength [Schofield, 1996; Tronsmo, 2003].

A model considering the physical entanglement of the glutenin polymers may explain much of the physical dough properties [Veraverbeke, 2002; Hamer, 1998]. According to this theory, the long polymers are hard to disentangle as they experience a high resistance to friction, leading to increased resistance to extension and longer mixing time [Hamer, 1998]. However, Lefebvre et al (2003) found that gluten could not be considered an entangled polymer system, but rather a particulate gel, in which a network is formed by aggregation of particles. The existence of glutenin particles (with diameters of 0.1-100 μm) has been confirmed in studies by Don et al (2003a, 2003b, 2005), who found the GMP to consist of glutenin particles above a certain size (below the size criterion, the particles were SDS-soluble). In addition to the amount of GMP, also several properties of the GMP particles were shown of importance for determining the rheological properties of doughs. These properties were considered more important than the quantity of GMP and included the size of the particles and their tendency to interact with each other [Don, 2005].

4.2. Interactions in the gluten matrix

A hyperaggregation model outlines three levels of aggregation for the glutenins: First, the covalent S-S bridges are involved in the formation of soluble glutenins. Second, interactions between the soluble glutenin polymers lead to the appearance of insoluble GMP particles, and last, the GMP particles participate in network formation by interactions with other particles [Don, 2005]. The aggregation is determined by various chemical interactions.

Types of chemical gluten-gluten interactions

Due to its role in the formation of glutenin polymers, the covalent S-S bridges play a major role in the gluten network stabilization, and thus modification with oxidizing or reducing agents greatly affects dough properties, e.g. the solubility and dough development time [Mejri, 2005; Rao, 2000]. Non-covalent interactions are important as well, although smaller effects on the dough are seen when interactions are modified by e.g. addition of salts, urea or ethanol [Hamer, 1998].

The prevalence of charged amino acids in gluten proteins is rather low (less than 10 %), but even a few salt bridges could play a role in the stabilisation of the gluten structure [Wrigley, 1988; Hamer, 1998]. On the other hand, gluten proteins contain a high amount of hydrophobic amino acids, and it has been established that hydrophobic interactions are essential to their ability to aggregate [Wrigley, 1988, Kinsella, 1984], so when stabilised by S-S bridges, the hydrophobic interactions may contribute significantly to the dough strength [Hamer, 1998]. In agreement herewith, the surface hydrophobicity of gliadins and glutenins has been correlated positively to the dough strength [Torres, 2000].

On the other hand, Grant et al (1999) found that gluten has a hydrophilic character. Likewise, NMR studies have indicated that hydrogen bonding may be the dominant factor compared to hydrophobic interactions with regard to their stabilizing effect on gluten protein conformations and gluten structure [Hargreaves, 1995]. The repeat sequences of glutenin and gliadin are very hydrophilic and possess a high capacity for inter- and intramolecular hydrogen bonding, caused by the high contents of Gln and Asn [Schofield, 1996]. However, there are many other possibilities for hydrogen bonding in gluten (between side chains, between peptide groups and between side chains and peptide groups) since many groups (unionised carboxyl groups, amide groups, phenolic/aliphatic hydroxyl groups and carbonyl/carboxyl groups) may participate in these interactions [Seabourn, 2002].The presence of intermolecular β-sheet in hydrated gluten proteins has been shown from FTIR studies [Pezolet, 1992], and the involvement of Gln side chains from the repetitive domains in the interchain hydrogen bonding has been indicated e.g. by NMR studies [Alberti, 2002] and found to be central in the stability of gluten conformations [Hamer, 1998].

Gluten network development

Dough mixing leads to the hydration of dough components, dispersion of the gluten phase, development of the three-dimensional network structure and the incorporation of gas bubbles [Hamer, 1998]. Based on the mobility of water, Chen et al (2002) established four stages of dough development: 1) Unfolding of the proteins into a random network. In this process, water acts as a plasticiser that increases the protein mobility. Water also becomes increasingly bound to the proteins in a hydration process, in the end of which all water is 'bound'. 2) The immobilised water rearranges and becomes more homogeneously distributed. 3) Realignment of the gluten takes place and some gluten-water interactions are replaced by protein-protein interactions. 4) The gluten starts to break down and more water is released.

The functional gluten matrix is a network of fibrillar strands, which interact with each other to form continuous sheets [Grant, 1999]. Development of this optimal structure involves rearrangements and alignment of the protein chains, promoted by disulphide interchange and exchange of intermolecular hydrogen bonds [Hamer, 1998]. Due to the chemical and physical alterations, the GMP particles that can be isolated from the dough have much distinct properties than those isolated from the flour. The initially spherical particles become smaller and obtain a more elliptic shape upon mixing, and at high shear rates, they form the continuous phase in the dough [Don, 2005]. A decrease in GMP particle voluminosity (correlating to particle size) may be fundamental in the dough development, since different doughs are found to contain GMP particles of nearly the same voluminosity at the optimal mixing time (MT), even though their initial particle voluminosity differ [Don, 2003b]. The decrease of GMP particle voluminosity is caused by dissociation of glutenin particles (involving disruption of hydrogen and hydrophobic interactions) and involves the conversion of GMP particles to smaller SDS-soluble glutenin particles [Don, 2005]. At MT, most of the GMP has disappeared from the dough, but the remaining GMP has a high capacity to undergo particle-particle interactions, probably due to these particles being small and irregular. During the subsequent resting period, the reassembly of soluble particles takes place and leads to fast formation of GMP particles in case of optimally mixed doughs [Don, 2005]. If MT is further increased, a depolymerisation of the soluble glutenin polymers (involving breakage of S-S bridges) takes place, and a low GMP voluminosity is seen after resting [Don, 2005]. The diminishing of GMP voluminosity in the overmixed doughs agrees with the end-blocker theory: During mixing, the LMW/HMW ratio is increased in GMP [Hayta, 2001] and this could block the formation of larger polymers [Don, 2005].

The importance of gliadins in the aggregation behaviour of gluten has been demonstrated by Bean et al (1998), as they found that removal of gliadins before the addition of salts counteracted the salt-induced aggregation.

Starch and lipid interactions with gluten proteins

Flour contains 1-2 % lipids. The polar lipids consist of phospholipids and glycolipids, and the non-polar lipids consist of tri-, di- and monoglycerides, free fatty acids, steroids, carotenoids, tocopherole etc. It has been established that lipids are not essential to the gluten network formation [Hamer, 1998]. However, the polar lipids (that are not bound to starch) exert a positive influence on the bread volume and bread crumb, since they stabilise the dough network during proofing and baking [Hamer, 1998; 2003]. They may act as foaming agents to stabilise the thin aqueous films that separate the air bobbles in the dough foam. These effects are not the result of specific lipid-gluten protein interactions [Hamer, 1998]. Instead lipids are organized into liposomes and vesicles etc, which are either physically entrapped in the network or bound by unspecific interactions between proteins and the lipid phase interface [Hamer, 1998]. However, it has been found that some specific polar lipids bind to the gluten proteins upon hydration and that nonpolar lipids may bind by hydrophobic interactions [Hamer, 1998; Alzagtat, 2002].

Flour contains 63-72 % starch. As mentioned, gluten fibrils adhere to the starch particles, which are embedded in the protein network, and the protein-starch interactions could thereby influence the rheological properties of the dough. There are indications that the protein-starch interactions increase upon mixing until optimal MT, where after they diminish [Hamer, 1998]. The non-starch polysaccharides in flour are mainly divided into water extractable pentosans (WEP) and water unextractable solids (WUS). These fractions constitute together 2-3 % of the flour, and both exert an effect on the bread-making quality, probably due to their interference on the gluten network formation [Goesaert, 2005]. First, they may compete with gluten for the water, and secondly, they may be able to bind to the glutenins (pherulic acid is suggested to be able to crosslink WEP to the gluten proteins). Wang et al (2002) showed that WUS significantly changed gluten and GMP compositions and the rheological properties of the dough.

4.3. ATR-FTIR studies: Protein secondary structure in relation to gluten functionality

The structures of the functional hydrated gluten proteins differ from those in solution and have been the topic for a number of studies, involving the relationship between molecular structure and gluten functionality [Wellner, 2005, 2003, 1996; Feeney, 2003; Popineau, 1994]. Investigation of hydrated gluten proteins became possible with the appearance of ATR-FTIR instruments, and Pezolet et al (1992) were the first to use this method to demonstrate that hydrated gluten proteins form intermolecular β-sheet not seen in solution. Later ATR-FTIR studies have shown that the secondary structures of gluten proteins depend much on the water content [Belton, 1995; Feeney, 2003; Wellner 1996; Mangavel, 2001], and based on ATR-FTIR studies of different gluten fractions, Popineau et al (1994) suggested that the secondary structure is influenced by the aggregation of the gluten proteins.

Studies of the hydration-induced structure changes in glutenin and gliadin subunits have shown an initial increase of intra- and intermolecular β-sheet simultaneously with a decrease of random structure upon moistening of the dry proteins [Belton, 1995; Feeney, 2003]. The studies showed that at a certain degree of hydration, the β-sheet structure content was lowered again (probably due to water competing with peptide groups for binding), and extensively hydrated structures containing Pro appeared [Wellner, 1996; Belton. 1995]. Also, the spectral changes have pointed to an increase of (hydrated) β-turns and extended structures at increasing hydration [Belton, 1995; Feeney, 2003]. At full hydration, intermolecular β-sheets and hydrated structures have been found to co-exist in gluten. A 'loop and train' model has been proposed on the basis of these findings and suggests that the balance of the two structures (intermolecular β-sheets and hydrated structures) is important to the functional properties of gluten [Belton, 1999]. According to the model, the repetitive domains of glutenins exist either as hydrated chains i.e. the 'loops' or as 'trains', at which intermolecular β-sheets and hydrophobic interaction establish interactions between two or more chains (see Fig. 4.1). At increasing hydration, the loop regions, consisting of hydrated extended structures and β-turns, increase at the expense of 'train' regions [Belton, 1999]. It has been suggested that the final loop/train ratio affects the viscoelastic properties of gluten and is determined e.g. from the length of the repeat units, in agreement with the hypothesis that the intermolecular β-sheet interaction takes place between the repetitive domains [Belton, 1999]. Also, Feeney et al (2003) showed that imperfections in the repeat sequences lead to less intermolecular β-sheet interaction, and Popineau et al (1994) found that the long glutenin polymers would cause increased intermolecular β-sheet interactions.

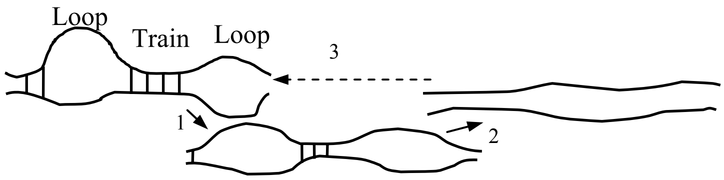


Fig. 4.1. Sketch of the 'loop and train' model of glutenin. Extension causes deformation of the loops (1). Further extension causes a disruption of the intermolecular hydrogen bonds (2). The initial structure is regained upon stress release (3). [Belton, 1999].

The 'loop and train' model can explain the elasticity of gluten, which has not been accounted for by other models. Wellner et al outlined in an 'extended loop and train' model how some of the HMW subunits may form stable long-lived polymers with high amounts of intermolecular β-sheets, due to the favourable alignment of their loops upon extension, i.e. the extended loops interact with other chains and form interchain β-sheet structure. Only the subunits that do not align favourably revert to the original loop conformation, and thus, there is a build up of intermolecular β-sheet during mixing [Wellner, 2005]. Seabourn (2002) showed in agreement herewith that the β-sheet/α-helix ratio was maximal at mixograph MT, where after the intermolecular β-sheet began to decrease.

4.4. Experiment IV: NIR analysis of gluten protein hydration and denaturation

In this study, the NIR spectra of dry and moistened gluten are compared with the purpose of identifying protein bands that are sensitive to the structural alterations, occurring during hydration of gluten proteins. Interpretation of the qualitative spectral changes is done with the help of ATR-FTIR spectra. The task is to obtain more knowledge about the influence of protein structure and interactions on NIR spectra, obtained from a complex matrix, and thereby evaluate the capacity of NIR for application in structure-function studies of food systems. The study will also shed light on the influence of water-protein interactions in NIR. (The experiment is also shortly described in Paper I).

Method

Gluten powder was prepared by hand washing from a wheat dough. One hundred gram commercial wheat flour was mixed with 50 ml distilled water and kneaded for 2 min. The starch, pentosans and soluble proteins were washed out by hand washing in 5*1 L distilled water or 2 % NaCl solution. In each case, the last wash was carried out in distilled water. The gluten preparations were freeze-dried, pulverized, and later kept at room temperature in an exicator with drying agent. The two gluten preparations are termed gluten A (prepared without salt) and gluten B (prepared with salt). Starch fractions were prepared from centrifugation of a batter: 250 g commercial wheat flour and 180 ml water was mixed and kneaded by hand before additionally 220 ml water was added. The mixture was blended in a blender for 1 min and the batter was centrifuged at 2500 rpm for 15 min. Two separate layers of starch (different particle sizes) were recovered and freeze-dried.

Experiment 1: Assignments

Three different gluten preparations: gluten A, gluten B and Gluten S (commercial gluten from Sigma) were measured by use of NIR. Only gluten A and gluten S were measured by use of FTIR. Initial water contents of the preparations were determined from oven-drying: Gluten samples of 1 g were dried in the oven at 120^{o}C until constant weight was obtained, and the water content was calculated as (1000 mg-final weight)/1000 mg and determined as ~4.3 % (wet basis) for gluten A and gluten B and ~6 % for gluten S. Also starch fractions and gluten-starch mixtures were measured by NIR and FTIR. A lipid spectrum was obtained by transmission-NIR measurements.

Experiment 2: Hydration

a) Gluten A powder samples of 700 mg were spread on petri-dishes, inserted into a humid closed container and kept at either room temperature or at 5^{o}C. After different time-intervals (t= 1, 2, 3, 4, 23, 27, 45 and 53 hrs) the samples were weighted and the NIR spectra obtained. After the measurements at t=4 hrs, two of the samples were transferred to a closed container with drying agent. Starch was moistened for 20, 24, 42 and 50 hrs and measured by NIR. All gluten treatments were carried out in replicates.

b) Samples of gluten A and B were placed in closed containers with either saturated $MgCl_2$ solution (1), saturated NaCl solution (2) or distilled water (3 and 4) in the bottom. NIR spectra were recorded after ~40 hrs.

c) Samples of gluten A and of gluten S were kept in humid closed containers at room temperature or 5°C for different time intervals until the FTIR spectra were obtained.

Experiment 3: Denaturation

Gluten S was moistened in a humid chamber for 65 hrs until water contents of ~20-25 %. The samples were filled in Eppendorph tubes and heated in water bath for different time intervals and different temperatures: at 65°C for ½ hr, or at 85°C for ½, 1 hr or 1.5 hrs. Some control samples were moistened and then dried at room temperature. Other control samples were not moistened but heated for ½ hr at 85°C. All samples were measured by NIR before and after the different treatments.

For NIR measurements, the powders were filled in a sample cup and measured in reflectance mode from 790 to 2500 nm on a Perkin-Elmer, Spectrum One, FT-NIR instrument, equipped with a reflectance assessory and an InGaAs detector, with a resolution of 8 cm^{-1} and use of 50 scans. The data interval was 1.67 nm. NIR spectra of lipid were obtained in transmission mode. Replicate measurements were obtained for the dry samples with different fillings of the sample cup. The pretreatments of the NIR spectra are described in the result section. FTIR-ATR measurements were carried out on a Bomen spectrometer equipped with a horizontal ATR-crystal (ZnSe). The FTIR spectra from 4000 cm^{-1} to 748 cm^{-1} were recorded at a resolution of 8 cm^{-1} and with co-addition of 50 scans. The data interval was 1.93 cm^{-1}. ATR-FTIR spectra were corrected by extended multiplicative scatter correction (EMSC) and 2nd derivative spectra were calculated by use of Savitzky Golay transformation. The inverted 2nd derivative FTIR spectra were examined.

Results: Assignments in the NIR spectra

The NIR spectra of the dry gluten samples were compared to the spectra of lipid and starch, which are contaminants in the gluten fraction (the gluten preparation produces a gluten fraction of ~80 % protein) (Fig. 4.2). The contaminants and their NIR absorptions should optimally be identified in order to assign any spectral changes to protein structural alterations.

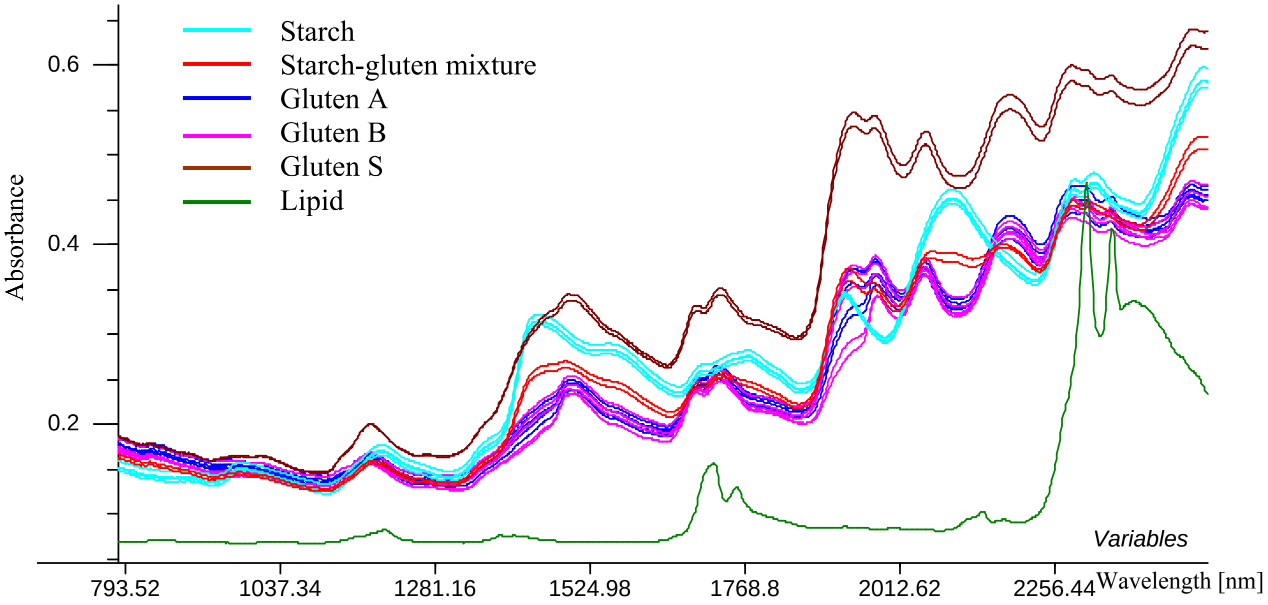


Fig. 4.2. NIR spectra of gluten A, gluten B, gluten S, starch and lipid. All are obtained in reflectance mode except for the lipid spectrum, which is obtained in transmission mode. Gluten powders were measured either after freeze-drying or oven-drying.

From Fig. 4.2, profound overlapping of the broad gluten and starch absorptions is seen throughout the spectrum. Light scattering effects are found to cause baseline differences between measurement replicates and between different gluten preparations. These effects are due to different sample packing and particle sizes (gluten S probably has the largest particles).

Lipid absorptions: The lipids used for obtaining the NIR reference spectrum originated from rapeseed, and therefore may differ from the actual gluten lipid spectrum. However, the confirmation of the lipid assignments is here done by use of the spectral differences between gluten A and B. Results from a PCA including the two preparations are shown in Fig. 4.3.

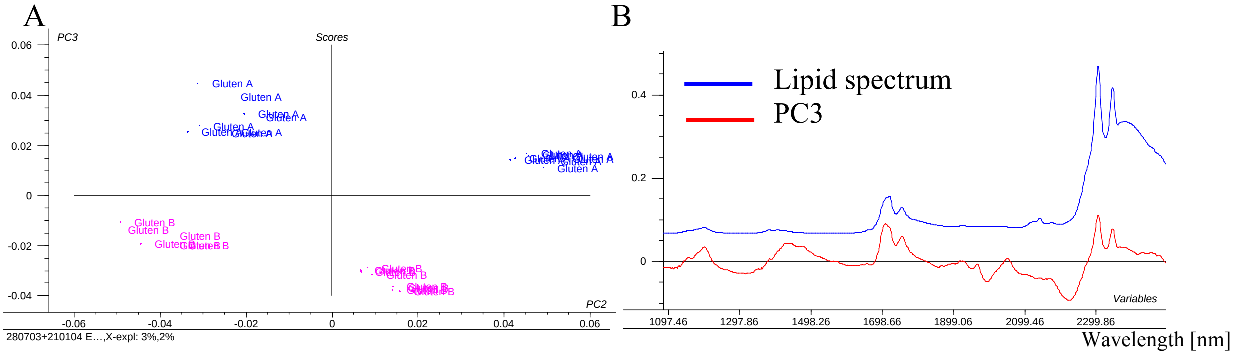


Fig. 4.3. A) PCA score plot (PC2 vs. PC3), discriminating gluten A (blue) and gluten B (pink). The PCA was based on the EMSC corrected NIR spectra of gluten A and B preparations. B) PC3 loading, compared to the NIR transmission spectrum of rapeseed lipids.

In Fig. 4.3, PC2 describes the batch to batch variation, whereas PC3 is able to discriminate between gluten A and gluten B. The PC3 loading plot in Fig. 4.3B shows similarity with the lipid reference spectrum, and the results imply that gluten A contains more lipid than gluten B. This is in accordance with the report that addition of salts, as in the preparation of gluten B, leads to greater washout of lipids [Sapirstein, 2002]. The lipid absorptions mostly result from vibrations of the CH groups, which are also abundant in proteins, and thus the lipid peaks are greatly overlapped by protein side chain absorptions.

In order to show in more details the contributions from protein and lipid absorptions to the gluten spectrum, the spectra of ‘lipid-deficient’ and ‘lipid-rich’ gluten were calculated by use of the PCA loadings: S^{lipid}= mean+0.1*p3 and S^{gluten}= mean-0.1*p3. As the following analyses will consider 2nd derivative spectra, the calculated spectra were also 2nd derivative transformed, and some interesting regions are shown in Fig. 4.4.

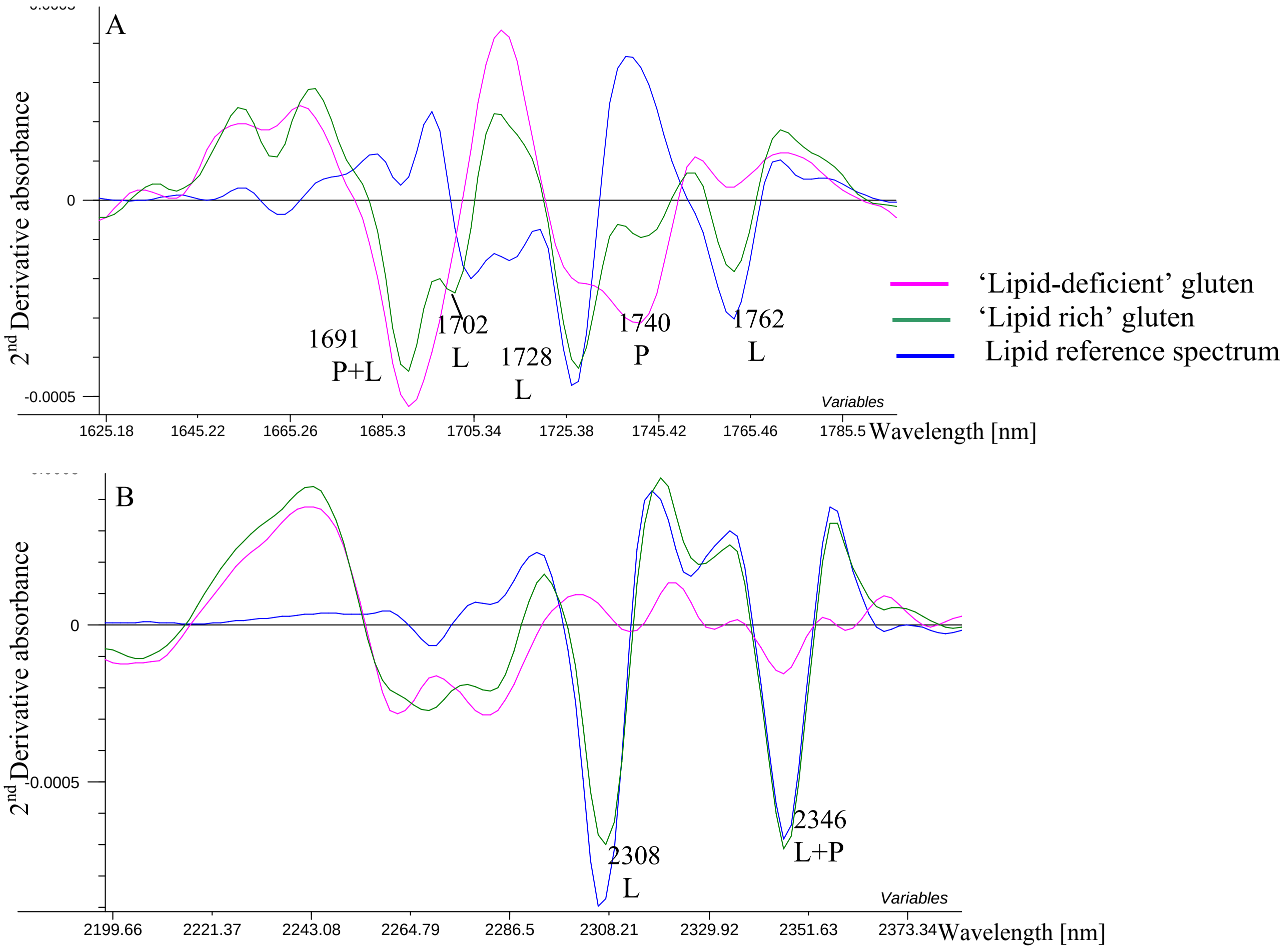


Fig. 4.4. Calculated 2nd derivative NIR spectra of ‘lipid-deficient’ and ‘lipid-rich’ gluten, compared to the lipid reference spectrum. Absorption peaks are downward-pointing. A) 1st CH-str. overtone region. B) Combinations band region. L=assigned to lipid, P=assigned to protein (or starch).

Based on Fig. 4.4, the most prominent peaks from lipids are seen at 1728, 1762, 2308 and 2346 nm. Small contributions are possible at 1691, 2269 and at 2283nm. (Compare to the lipid assignments in Appendix IV-1). Very small variations were also seen in the amide combination band region (not shown) for the two calculated spectra and may result from the weak lipid absorptions in this region. Absorptions due to cis-unsaturated bonds are seen at 2140 nm and 2170 nm in spectra of wheat [Law, 1977].

Starch absorptions: Starch contributions to the gluten spectrum were examined as well. The contaminations of the starch and gluten fractions are seen from the ATR-FTIR spectra (Fig. 4.5).

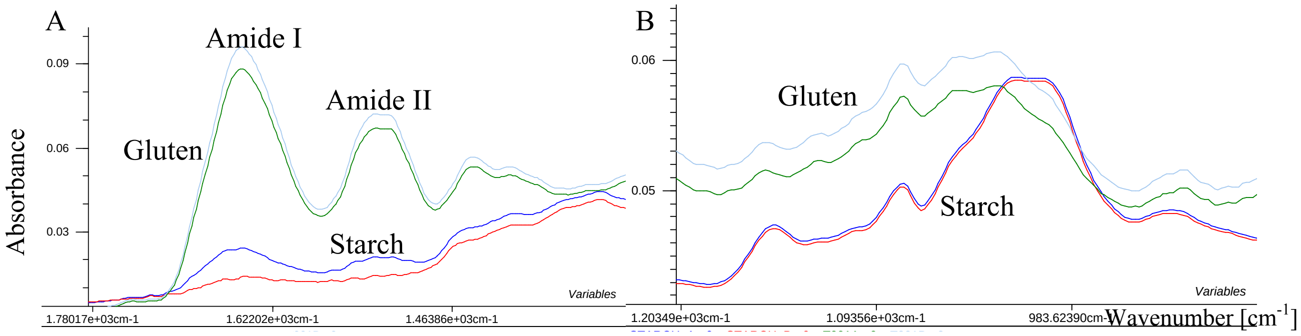


Fig. 4.5. ATR-FTIR spectra. Comparison of gluten and starch fractions. A) The amide I and II regions. B) The MIR fingerprint region. Blue and red: starch fraction. Green and light blue: gluten fraction.

In Fig. 4.5A, the amide I and II bands are observed in the starch spectrum, indicating that the starch fraction is not free of proteins, and likewise, small peaks at 1149 and 1078 cm^{-1} show that some residual starch is left in the gluten preparation. The starch absorptions in the gluten NIR-spectra were identified from a PCA, using spectra of gluten-starch mixtures (Fig. 4.6).

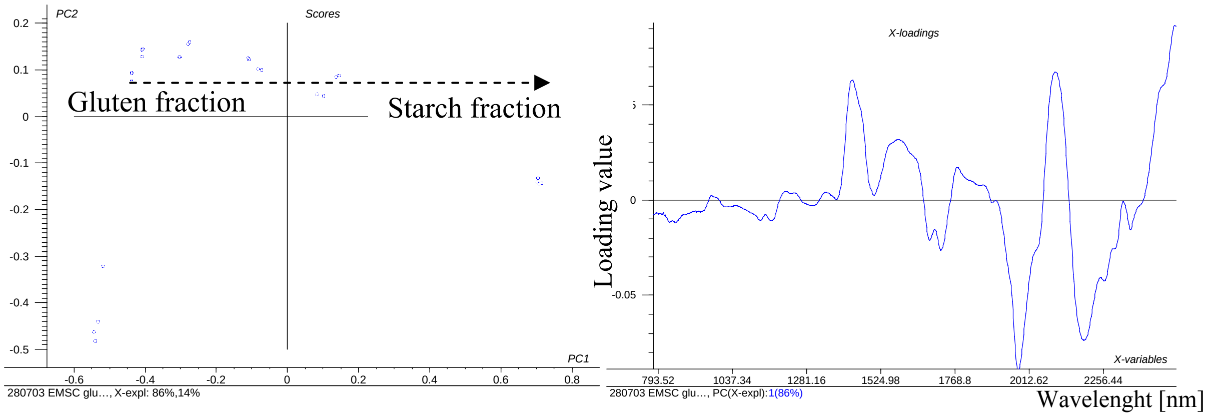


Fig. 4.6. A) PCA score plot (PC1 vs. PC2), describing the gluten-starch ratio. The PCA was based on the EMSC processed NIR spectra of gluten-starch mixtures. B) PC1 loading.

PC1 describes the gluten-starch ratio, and thus, to emphasise the differences between starch and protein spectra, the ‘starch-deficient’ gluten spectrum was calculated from the PC1 loading (Fig. 4.7). The small peaks in the gluten spectrum at 1433, 1696, 1778 and 2283 nm can from Fig. 4.6 be assigned to starch. The prominent and broad starch bands at 1927 and 2095 nm are also seen in the 2nd derivative gluten spectrum as weak and blurred bands (2nd derivative transformation emphasises the narrow peaks). Likewise, contributions of starch to the bands at 2263 nm (very weak) and 2319 nm may be observed.

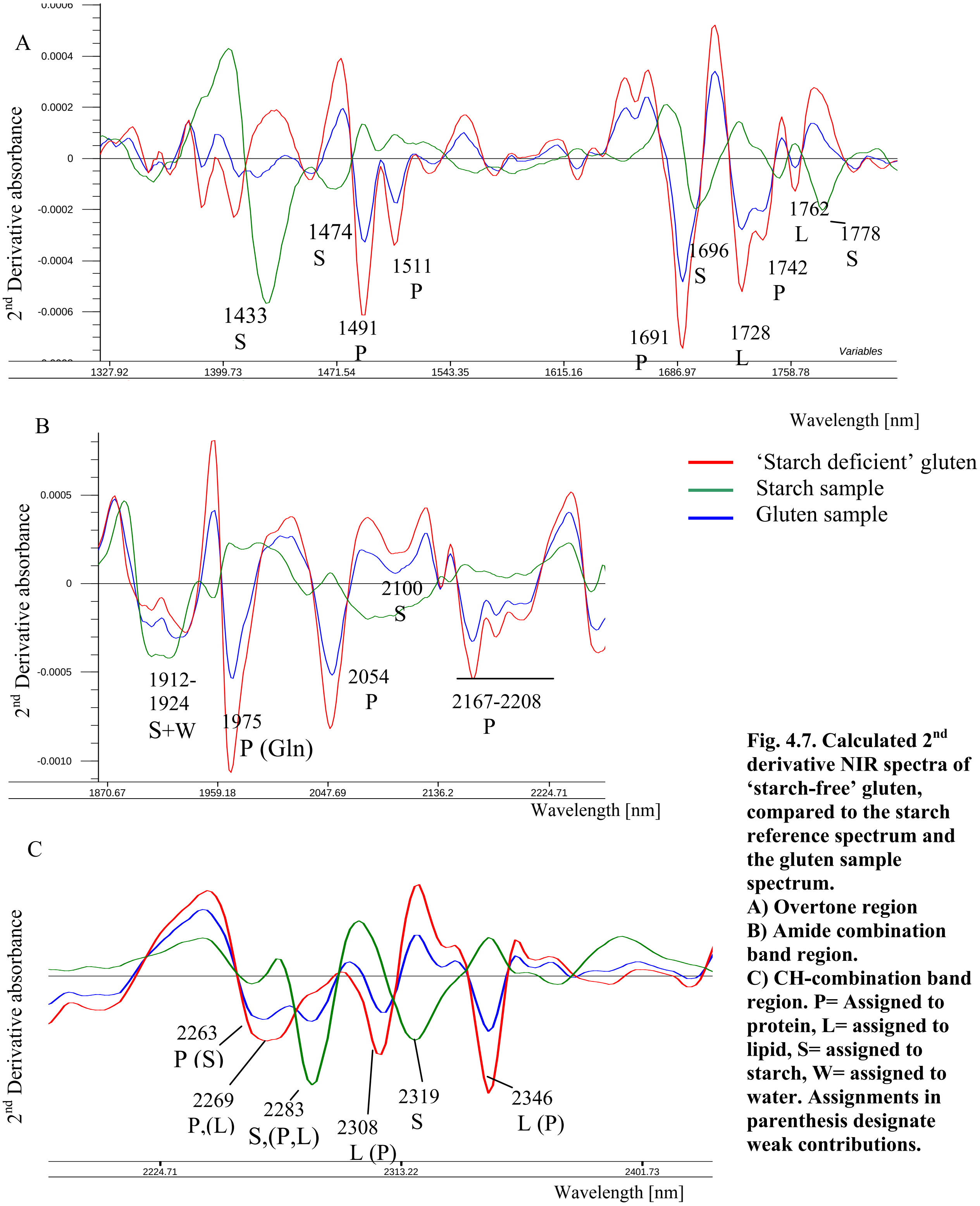


Fig. 4.7. Calculated 2nd derivative NIR spectra of 'starch-free' gluten, compared to the starch reference spectrum and the gluten sample spectrum.
A) Overtone region
B) Amide combination band region.
C) CH-combination band region. P= Assigned to protein, L= assigned to lipid, S= assigned to starch, W= assigned to water. Assignments in parenthesis designate weak contributions.

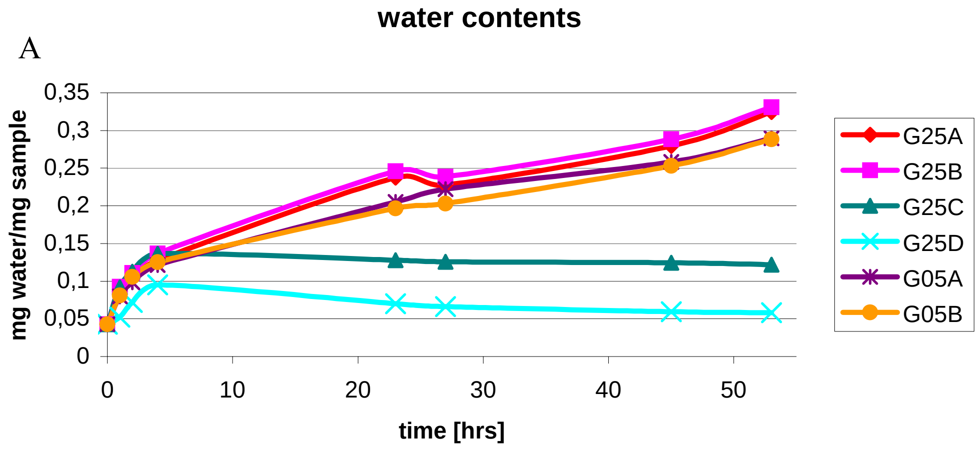
Assignments of bands to protein, lipid and starch based on the above analyses are also shown in Fig. 4.7 and are summarised in Table 4.1. In conclusion, many peaks (even in the resolved spectra) may be composed of absorptions from all three constituents (protein, lipid, starch). However, some unique peaks are also shown in the table.

	1400-1800 nm	1900-2220 nm	2220-2400 nm
Protein	1491, 1511, 1691, 1742	1975, 2054, 2167-2208	2263, 2269, (2283), (2308), (2346)
Lipid	(1691), 1702, 1728, 1762	Very weak bands	(2269), (2283), 2308, 2346
Starch	1433, 1474, 1696, 1778	1912-1924, 2100 (broad),	(2263), 2283, 2319

Table 4.1. Some assignments made in the 2nd derivative gluten spectrum. Compare to Appendix IV-1.

Results: Gluten hydration

Water contents for the moist gluten samples in Experiment 2a and 2b are shown in Fig. 4.8 and ranged from ~4 to ~33 % (calculated on a wet basis).



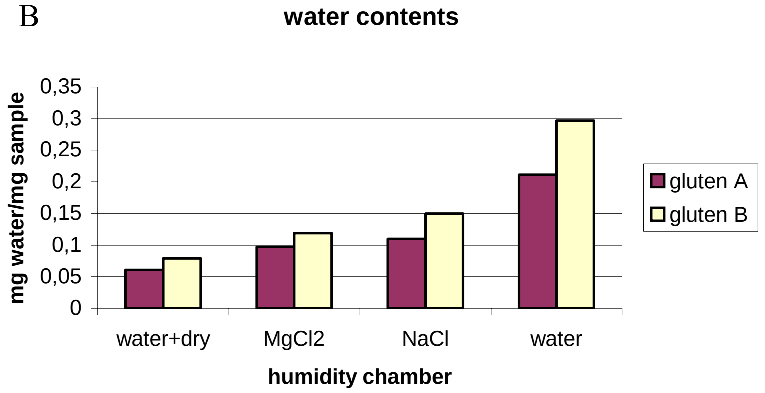


Fig 4.8. Water contents in gluten.

**A) Water contents in experiment 2a.
G25: stored at 25°C in humid chamber.
G05: stored at 5°C in humid chamber.
G25C and G25D: dried after 4 hrs of moistening at 25°C.
B) Water contents in experiment 2b.**

Samples stored at 5°C (G05A,B) absorbed less water than samples stored at 25°C (G25A,B), and these samples also had different colour and textures: the particles in the moist G25 samples were larger and more yellowish than the particles in the equally moist G05 samples. These differences may be explained from 5°C and 25°C being below and above the glass transition temperature (T_g) of moist gluten, respectively [Lens, 2003]. Storage of gluten above its T_g increases water absorption due to increased exposure of polar groups [Elizalde, 1999], and therefore, the G25 samples also has increased particle sizes (the coherency increased with increasing water content). The colour differences may be explained from different densities, as the density may be increased for gluten samples stored above T_g. At this condition the porous structure is lost due to rearrangement of the proteins, and this affects the colour [Elizalde, 1999]. Also, lipid oxidation may contribute to the colour changes. Between gluten A and B was noticed the difference that gluten B absorbed more water than gluten A (Fig. 4.7B). This could stem from a higher protein content in gluten B compared to gluten A. Alternatively, the salt has caused increased exposure of polar and ionic groups on the gluten surface, which could have increased the interaction with water [Mejri, 2005].

FTIR-results: ATR-FTIR spectra reveal the structural changes that underlie the physical changes observed upon increasing water content. For these measurements, the water contents of the moistened gluten samples only ranged from 15-20 % (on a dry basis). Raw ATR-FTIR spectra, resulting from both gliadins and glutenins, are shown in Fig. 4.9.

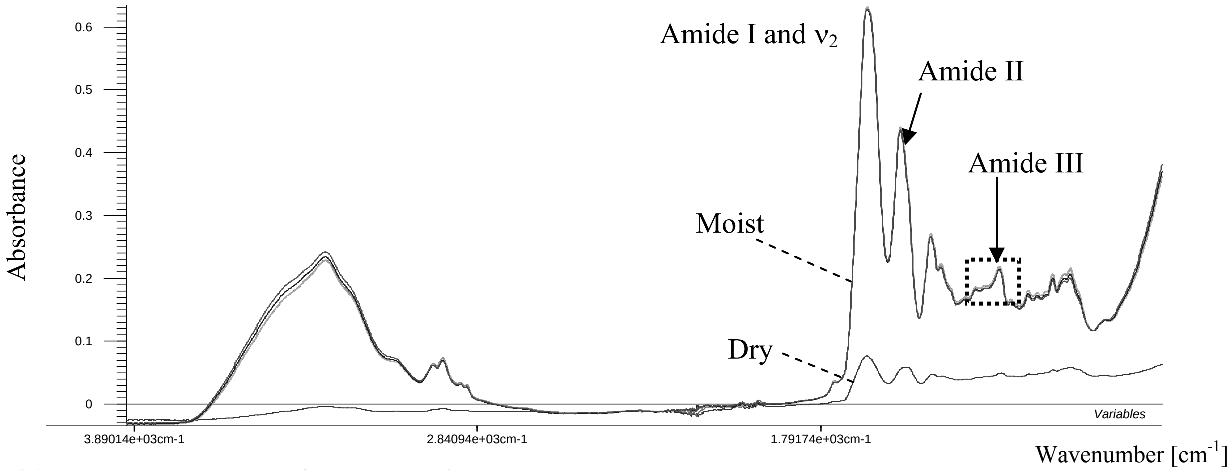


Fig. 4.9. ATR-FTIR spectra of dry and moist gluten.

The low intensity of the spectra of the dry gluten powder is the result of a low contact of these samples to the ATR-crystal. The different intensities of dry and moist gluten spectra may cause different noise levels and cause the bands to appear at slightly different wavenumbers due to the anomalous dispersion effect.

The inverted 2nd derivative spectra in the amide I and amide II region of dry and moist gluten are shown in Fig 4.10. Water usually absorb with a maximum at 1635 cm^{-1}, but it appears that water may cause only very absorption at this position. (The binding of water to gluten is considered weak and is not thought to cause shift of the water band).

Hydration may give rise to spectral changes that are not related to secondary structure changes. For example, it is still under debate whether spectral changes introduced by freeze drying of proteins reflect the removal of water *per se* or changes in secondary structure upon dehydration [Griebenow, 1995; van de Weert, 2001; Al-Azzam, 2002]. First of all, the hydrogen bonding between water molecules and the protein amide groups may change the amide frequencies as described in chapter 3. In addition, an intensity-increase and narrowing of the amide I band is generally observed when proteins are hydrated. Pevnser et al (2001) suggested that this was related to increasing dielectric properties of the protein environment upon hydration.

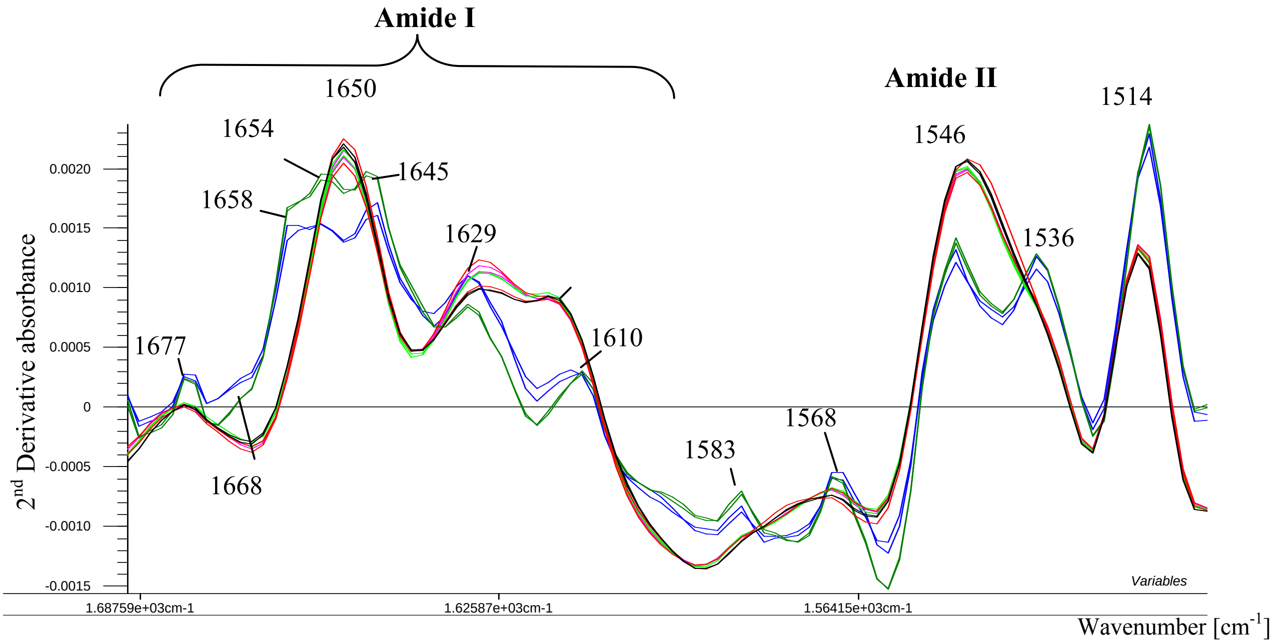


Fig. 4.10. Inverted 2nd derivative ATR-FTIR spectra of the amide I and amide II region (1700-1500 cm-1). The dry gluten spectra have been scaled for comparison. Dry samples: gluten S=Blue, gluten A=green. Moist samples: gluten S= light green, pink. Gluten A= red, black (15-20 % water contents on a dry basis).

On the contrary, Al-Azzam et al (2002) concluded that the amide I band is relatively insensitive to the physicochemical environments and supported the hypothesis that the spectral changes solely reflect the structural changes (β-sheet content usually increases on the expense of α-helix upon dehydration) [Griebenow, 1995].

The peak assignments in the amide I region are done according to the references: Belton, (1995), Feeney et al (2003) and Wellner et al (1996). In the dry gluten spectra, the peaks at 1610 cm^{-1} and 1629 cm^{-1} result from intermolecular β-sheet/Gln side chains and intramolecular β-sheet, respectively. The broad band around 1650 cm^{-1} contains contributions from α-helix (1654 cm^{-1}), β-turns (1640 cm^{-1}), random coil (1645 cm^{-1}), Gln side chains (1658 cm^{-1}) and perhaps 3_{10}-helices (1660 cm^{-1}). The band at 1660 cm^{-1} has also been ascribed to a distortion caused by the intermolecular hydrogen bonding between Gln side chains and the peptide backbone in the dry gluten [Wellner, 1996].

Upon hydration the broad band at 1650 cm^{-1} becomes narrower, and perhaps reflects the unfolding of the protein chains and thereby decreased distortion. Also a decrease of random structure is indicated and is in agreement with other FTIR studies, in which hydration has been found to increase the secondary structures in gluten proteins [Feeney, 2003]. The small intensity-increase at 1650 cm^{-1} suggests a somewhat increased α-helix content. However, some of the above changes could stem from the altered dielectric properties, as described.

Another obvious change upon hydration is the intensity-increase in the region between 1610 cm^{-1} and 1630 cm^{-1}. This region can be ascribed to intermolecular β-sheets and to extended hydrated chains (both at 1612-1620 cm^{-1}). The development of extended hydrated structures, which give rise to intermolecular β-sheet, is in agreement with other FTIR studies of glutenin/gliadin hydration [Belton, 1995; Wellner, 1996; Feeney, 2003]. Furthermore, absorptions from extensively hydrated Pro have been ascribed to this low frequency region. The structural changes seen from the amide I band are not prominent. This is in agreement with a hydration study of gliadin, which indicated minor structure changes to take place at moisture contents above 15 %, while larger structural changes required water contents above 38 % [Wellner, 1996].

In the amide II band, the dry protein absorb at 1546 and 1536 cm^{-1}. After hydration, the peak intensity increases at 1546 cm^{-1}, whereas the 1536 cm^{-1} band is turned into a weaker shoulder. These changes are hard to interpret due to the ambiguous assignments in the amide II region. Pevsner et al (2001) observed increased amide II band intensity and a shift to higher frequencies upon hydration of several different proteins. A linear increase of amide II band area with increasing water content in gluten has also been reported by van Velzen et al (2003), and they explained this from e.g. a better contact of the sample to the ATR-crystal or an increased dipole moment of the NH amide groups upon water binding. Also, changes seen in amide II may reflect the hydration more than the structure changes [Wellner, 1996].

The peak at 1515 cm^{-1} has been ascribed to Tyr absorptions. The decreased intensity of this peak in the moist gluten spectrum may reflect a lower content of unordered structure, as this band in a study was shown to increase concomitantly with the unordered amide I component [Liu, 2001].

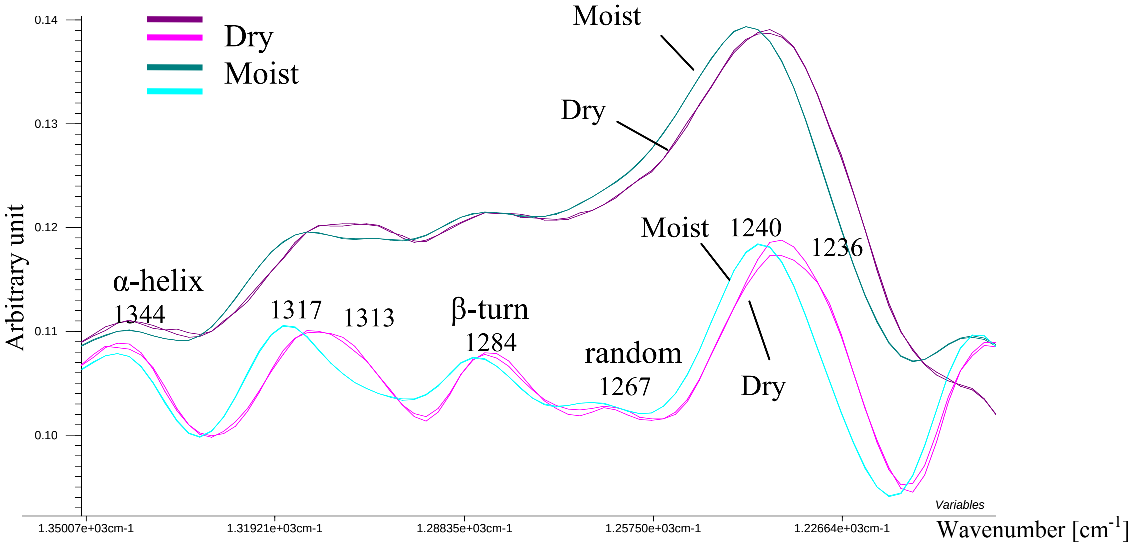


Fig. 4.11. Raw and inverted 2nd derivative spectra of dry and moist gluten in the amide III region. Assignments are: 1330-1295 cm^{-1}: α-helix, 1295-1270 cm^{-1}: β-turns, 1270-1250 cm^{-1}: random coil, 1250-1220 cm^{-1}: β-sheet. Spectra have been scaled to comparable intensities.

The raw and 2nd derivative amide III region of dry and moist gluten are shown in Fig. 4.11. The different secondary structure assignments shown in Fig. 10 are in accordance with the work of Seabourn (2002). The high resolution of peaks in the raw amide III band means that 2nd derivative transformation is unnecessary and also will remove some quantitative information. The shifts to lower frequencies of all bands in the amide III region could result from the hydration and the conversion of NH..CO hydrogen bonds to NH-water hydrogen bonds. Other changes related to the hydration include a narrowing of the 1317 cm^{-1} and the 1240 cm^{-1} band, but these variations do not reveal large structural changes.

In conclusion, hydration-induced changes in the MIR region are observed mostly in the amide I band, but not all of them reflect structural changes in gluten proteins. The most prominent structural change is the development of intermolecular β-sheet and hydrated structures, as seen from the amide I band.

Results: Hydration induced changes in NIR

Raw NIR spectra from Experiment 2 are shown in Fig. 4.12. The increases of the water bands at 1450 and 1935 nm are seen. The baseline variations are the result of the light scattering effect, probably due to the increased particle size of the moist samples and to the change of refractive index of the media (water has replaced air).

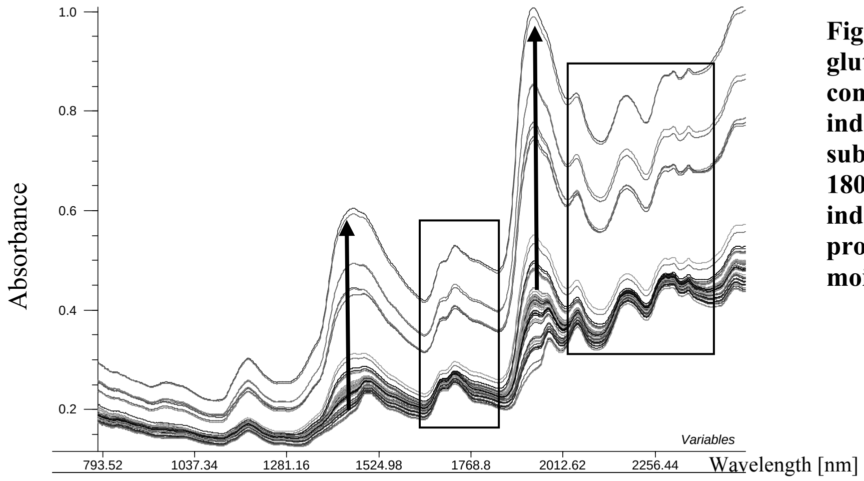


Fig. 4.12. Raw NIR spectra of gluten at different moisture contents (~0-33 %). The boxes indicate the regions that are subjected to further analyses (1600-1800, 1960-2380 nm). The arrows indicate the increases of the most prominent water peaks upon moistening.

Especially a large baseline increase was seen for samples moistened at 25°C to at least 20 % water content, whereas this was not seen for the samples of equal water contents but stored at 5°C. The difference most likely results from the different particle sizes obtained at the two temperatures. EMSC was performed on the whole spectrum in order to remove and analyse these physical effects. The EMSC coefficients all showed high correlation to the water content, but the slope-parameter (b) differed for the 5°C and 25°C-samples of moisture contents above 20 %, and probably, this coefficient is related to the particle sizes. In a PLSR analysis of the correlation between the EMSC parameters and the water content, two PCs were needed for obtaining an R or 0.96. Thus, the water content correlates well with physical information in the NIR spectra.

The physical effects seemed not possible to eliminate by use of 2nd derivative calculation only, since EMSC applied after the 2nd derivative transformation still yielded EMSC parameters with correlation to the water content (R=0.97 in a PLSR). For improving the correction, EMSC was performed on the 2nd derivative transformed spectra in two separate regions (1600-1800 nm and 1960-2380 nm) that were overlaid only partly by the water bands. This pretreatment seemed to eliminate most of the light scattering effects and the minor quantitative variations that result from the different water-gluten ratios (only the qualitative changes are of interest in the following analysis). Another preprocessing method, based on EMSC with the water spectrum used as "bad spectrum", is described in Appendix IV-2 and had the advantage that a larger part of the spectrum could be included in the analyses. This method was able to remove most of the water band variations at 1930 nm but not those at 1440 nm. As the pretreatment resulted in spectra similar to those obtained in the first procedure, the latter are used in the following analysis. The pretreated spectra of dry and moist gluten are compared in Fig. 4.13.

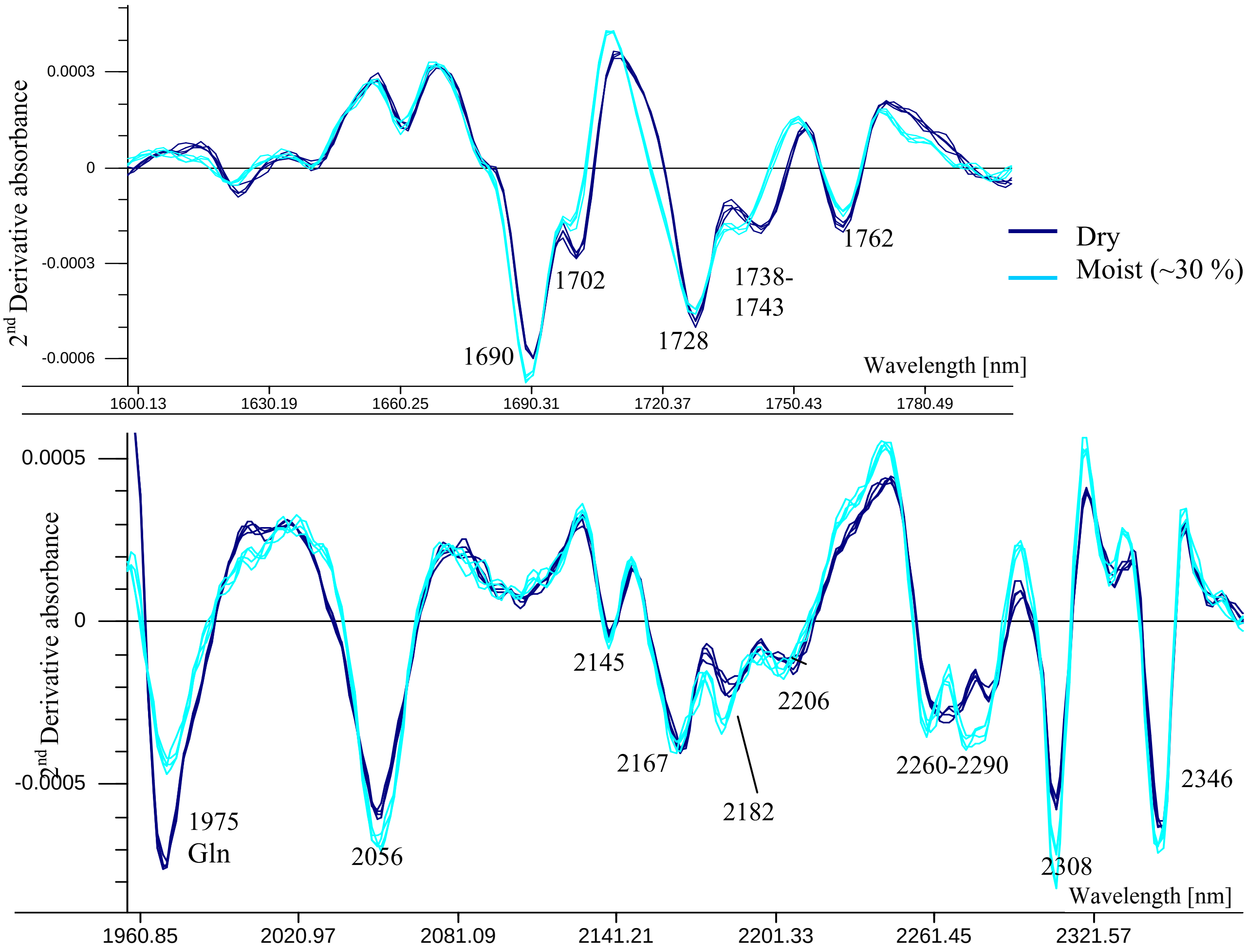


Fig. 4.13. Preprocessed 2nd derivative NIR spectra of dry and moist gluten A. Absorption peaks are downward-pointing. A) 1600-1800 nm, B) 1960-2380 nm. See assignments to protein, lipid and starch in Fig. 4.6.

In the two regions are found some spectral changes, which could reflect qualitative changes in the gluten samples upon hydration. One region showing an interesting spectral change is the 2260-2290 nm region, where similar changes are seen for both gluten A and B and for the samples stored at different temperatures (see Fig. 4.14).

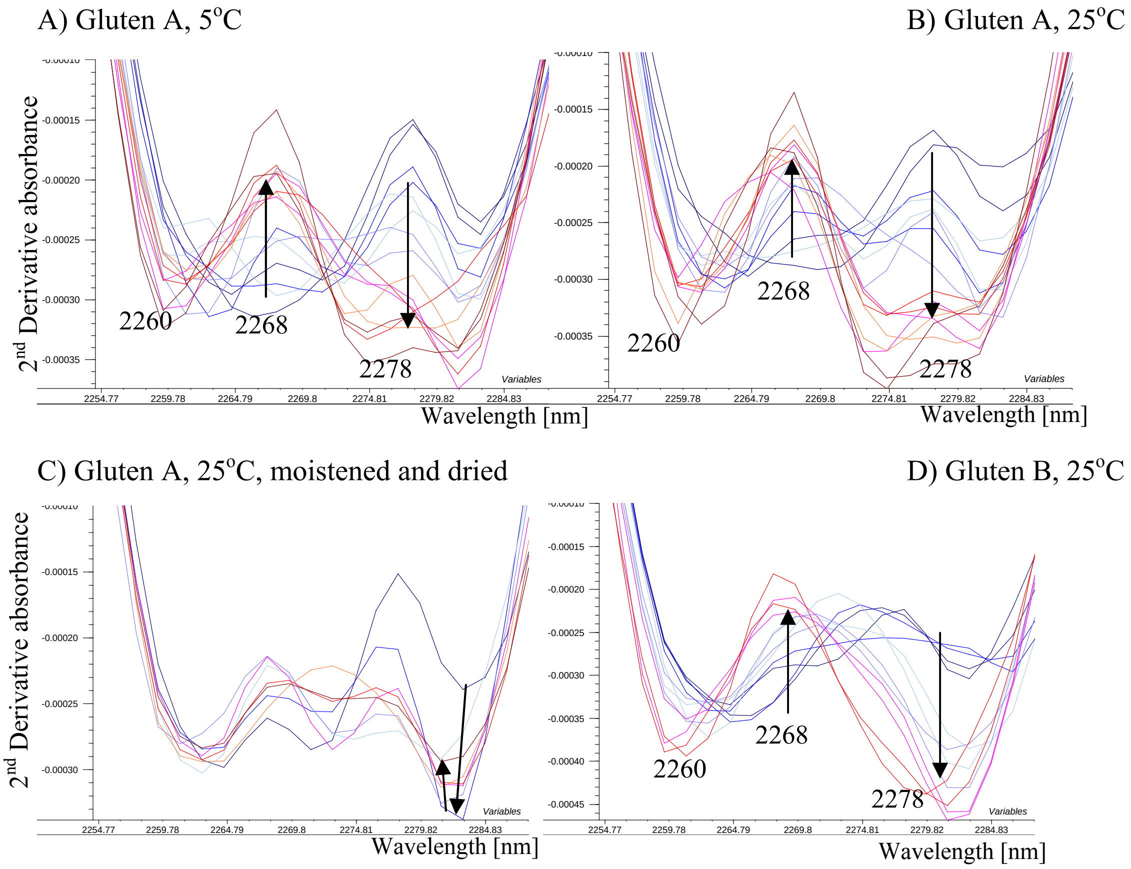


Fig. 4.14. Effect of gluten hydration in the 2260-2290 nm combination band region. The 2nd derivative absorption peaks are downward-pointing. The colours indicate the time-course of hydration (from blue to red). The ratio A_{2268nm}/A_{2278nm} has a correlation of 0.94 to the water content.

The fact that the changes are similar for gluten A and B, which contains different amounts of lipids, indicates that lipids are not the primary cause of the spectral change. Also, as the spectra of gluten at equal moisture content but stored at 5 and 25°C are similar, it is indicated that the spectral change is not related to the particle size or the lipid oxidation. The shift starts already at 9 % water content or lower, and drying of the moist samples is found to cause the reversion towards the initial spectral pattern (Fig. 4.14C).

A possible mechanism contributing to this spectral change is fermi resonance.

Overview of the spectral changes in NIR

A PLSR was performed for correlating the preprocessed spectra (**X)** to the water content (**Y)**. The resulting score plot is shown in Fig. 4.15. Three PLS components describe together 94 % of the Y-variance and 38 % of the X-variance. PC1 is correlated to the water content, and PC3 shows a parabolic shape with increasing water content. (PC2 discriminates gluten A and gluten B). Use of EMSC data without 2nd derivative calculation resulted in a PCA showing the same three phenomena. Whereas PC1 captures ~56 % of the X-variance and ~73 % of the Y-variance, PC3 can only explain ~7 % and ~8 % of the X- and Y-variance, respectively. PC3 scores decrease from the oven-dried samples to a minimum at the low water contents (~5-10 %) and then increase again for the higher water contents.

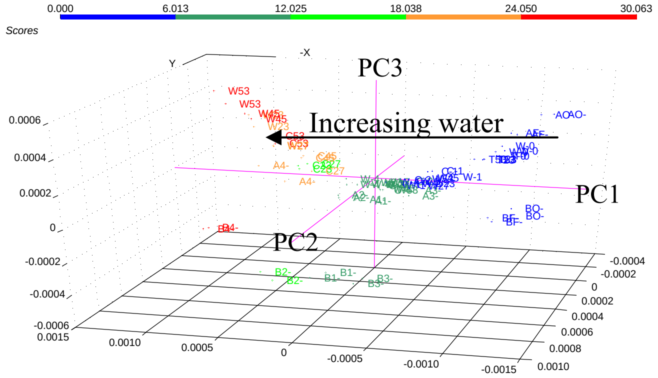


Fig. 4.15. Score plot (PC1, PC2,PC3) from at PLSR. X= NIR spectra in the 1600-1800 nm and 1960-2380 nm regions (after 2nd derivative+EMSC transformation). Y=Water content. Samples include gluten A and B at different moisture contents. Segmented CV with replicates kept together in the segments was applied.
Performance of the model with 3 components: R=0.968. Slope=0.93. RMSECV(Y)=2.22 %.

The PC3 variation could therefore possibly reflect the progress from the extensive protein-protein interactions in the dry protein to the protein unfolding and then the development of more protein-protein interactions again, when intermolecular β-sheets starts to form. This is only a hypothesis.

From the ATR-FTIR spectra it could be expected that some of the hydration-induced changes in NIR are related to quantitative variations of some constituents relative to others, as there is observed an increase of protein bands compared to starch bands upon hydration in the ATR-FTIR spectra (not shown). Van Velzen et al (2003), who observed the same, suggested that hydration-induced travelling of gluten to the surface layers could be the cause of this observation. However, the rearrangement of gluten and starch in different layers on hydration should not affect the NIR spectra, since NIR light can penetrate millimetres into the sample. Although, the relative contributions of protein, starch and lipids to the NIR spectrum is unaltered, the hydration and the thereby induced interactions may lead to qualitative changes in both protein, starch and lipid spectra, as will be examined further.

Possible starch and lipid conformation changes

Hydration of a flour leads to the swelling of starch granules and beginning leakage of starch. As NIR is sensitive to the intra- and intermolecular hydrogen bonding interactions in starch, these changes could possibly contribute to the observed spectral changes.

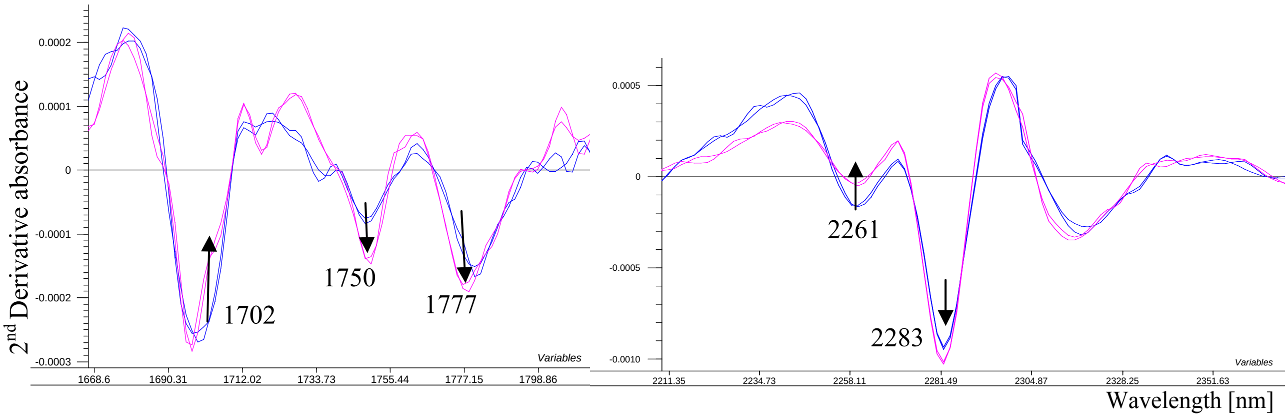


Fig. 4.16. Effect of the hydration of wheat starch on 2nd derivative NIR spectra in the regions 1670-1800 nm and 2211-2340 nm regions. Peaks are downward-pointing.

The OH-stretching 1st overtone bands have been used for detection of the water absorption capacity of wheat flour [Delwiche, 1994] and for measurement of the starch melting and degradation during extrusion processing [Millar, 1996]. The NIR spectra of dry and moistened starch are shown in Fig. 4.16. The OH-stretching is here represented by the 2283 nm absorption, which is the OH stretching combined with C-C stretching. Comparing to Fig. 4.13 and 4.14, it is found that starch hydration only may contribute to the decreasing absorption at 1702 nm and the increasing absorption at 2283 nm in the spectra of increasingly hydrated gluten.

Gluten washing leads to removal of non-polar lipids, and the remaining lipids consists mainly of polar lipids. The phase of these lipids may change during hydration of gluten and affect the NIR spectra. In dry flour, the polar lipids form aggregates with arrangements of tubules (hexagonal phase) or granules (cubic phase) [Hamer, 1998]. At low water content (<15 %) in the flour, only tubules exist, but at increasing water content, a lamellar phase is formed from the aggregates. The CH stretching and bending vibrations of lipid acyl chains are sensitive to the conformation of the hydrocarbon chains, which may vary in the different phases. For example, the fundamental CH_2 stretching bands shifted 3-4 cm^{-1} towards higher wavelength when the phosphatidylserine changed from a lamellar crystalline phase to a lamellar gel phase [Lewis, 2000]. The CH_2 deformation bands are also very sensitive to the packing of the hydrocarbon chains, and they may also undergo a splitting due to crystal field splitting [Lewis, 2000]. Thus, it is likely that some lipid conformation changes affect the NIR spectra of gluten during hydration (e.g. cause the increases at 2308 nm and 2346 nm in Fig. 4.13).

Possible protein conformation changes

In order to gain more information on the effect of gluten conformations on the NIR spectra, the heat denatured gluten was analysed. Spectra were preprocessed as described and a PCA was carried out. The score plot is shown in Fig. 4.17. PC1 explains 56 % of the spectral variation and seems related to the water content. (Some drying of the samples is noticed to take place during the heat-treatment).

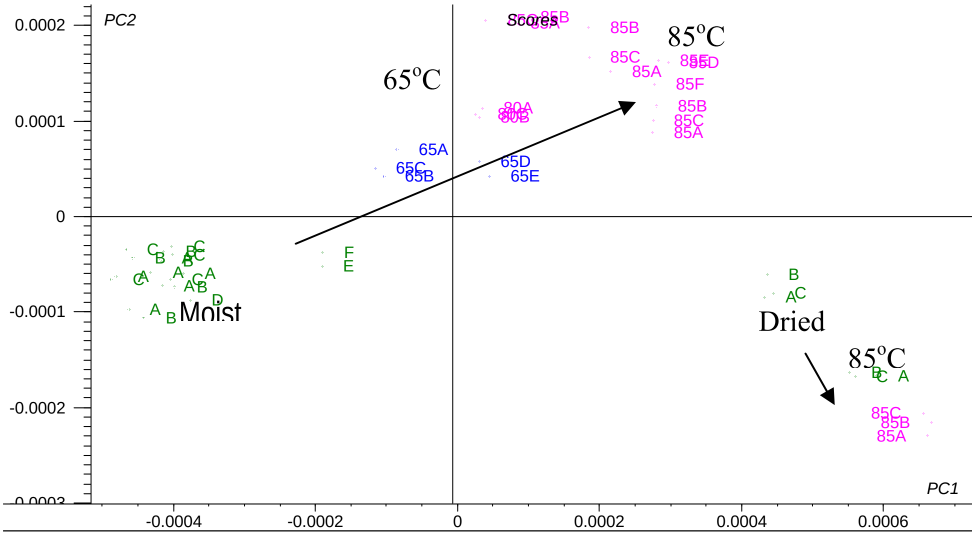


Fig. 4.17. Score plot (PC1 vs. PC2) from a PCA using the preprocessed spectra from the gluten denaturation Experiment 3.

Heat-treated (65°C)
Heat-treated (85°C)
No heat treatment

PC2 scores are increased for those moist samples that have been heat-treated but are decreased for the equally heat-treated dry samples. Therefore, the PC2 scores can be related to heat-induced conformational changes of the gluten proteins, which take place for moist gluten but not for dry gluten. Upon heating, the gluten proteins unfold and expose hydrophobic regions that may aggregate by hydrophobic interactions, causing increased random structure. The denaturation then become irreversible, as S-S interchange takes place and keeps the proteins in the denatured aggregated conformation [Domenek, 2002]. Thus, heating above ~50°C induces more cross linking of gluten proteins, increases the gluten strength (notably at 90°C) and alter the rheological properties of the dough [Hamer, 1998; Hayta, 2001; Micard, 2001]. Also at 60°C, starch gelation starts with the gelatinisation (breakdown of internal crystal structure).

In Fig. 4.18, the spectral changes caused by heating of gluten are compared to those caused by moistening. The latter changes are found the most prominent. In Fig. 4.18A and Fig. 4.19A, an increase at ~2184 nm with increasing water content is seen, suggesting an increase of α-helix content (see secondary structure assignments in Table 3.3).

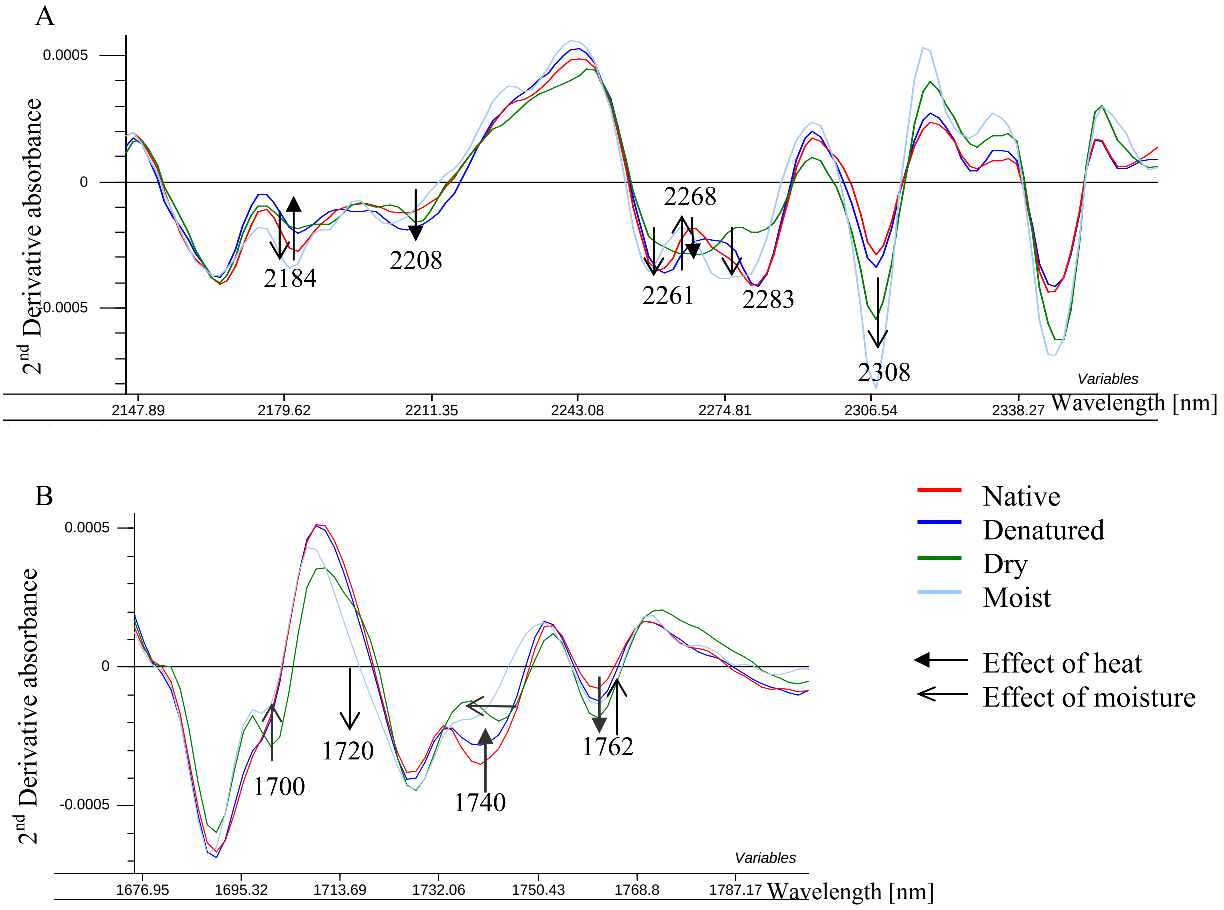


Fig. 4.18. Comparison of heat- and moistening-induced effects on 2nd derivative NIR spectra. Absorption peaks are downward-pointing. The spectra are calculated from loadings and scores from the PCA analyses shown in Fig. 4.15 and 4.17. A) Combination band region. B) Overtone region. Assignments to secondary structures are in accordance with Table 3.3 (in chapter 3).

Oppositely, a decrease at the same wavelength for the heat treated samples could stem from an expected decrease of α-helix upon heat-denaturation. However, the band at 2167 nm is not concurrently increased/decreased, and this could mean that the increase mostly reflects changes in the side chains from Gln/Asn. This is not unlikely, as these side chains take part in intermolecular β-sheet formation and may participate in water interactions. However, a simultaneous increase at ~2283 nm for the moistened sample supports an α-helix-increase (if not resulting from starch).

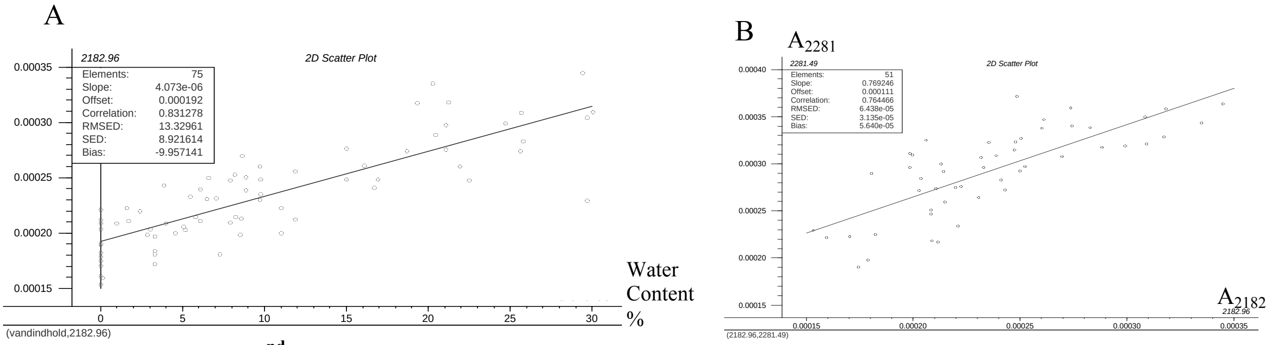


Fig. 4.19. A) Inverted 2nd derivative absorbance at 2182 nm as a function of water content. B) Plot of A_{2281nm} against A_{2182nm}. Both wavelengths are thought to represent α-helix.

At 2268 nm, an increase and decrease of unordered structure after heat treatment and moistening, respectively, is indicated and is in agreement with the expectations. In the 2205-2210 nm β-sheet region, an increase is seen for the heat treated sample and could reflect increased β-sheet. However, no increase is seen for the moistened samples in this region, even though an increase due to the formation of intermolecular β-sheet could have been expected. Instead, there is a small increase at 2258 nm (also β-sheet region). The indication of distinct intermolecular β-sheet absorptions at 2250 nm in previous experiments (Experiment III) is not supported in the present study. In the overtone region, an increase at 1720-1728 nm and a decrease at ~1700 nm as well as a shift to lower wavelength for the peak at 1740 nm (Fig. 4.18B) upon moistening could also indicate secondary structure changes, but could as well stem from lipid changes (see Table 4.1). In addition, the S-S interchange could cause some changes in this region, as the 1st overtone of SH-str. is found at 1735-1745 nm [Siesler, 2002].

Also, in the 2nd derivative spectra, changes in the amide A/II bands from the protein backbone (2056 nm) and Gln side chains (1975 nm) are seen when gluten is moistened (see Fig. 4.13). Whereas the 1975 nm band decreases with increasing water content, the 2056 nm band increases concomitantly. The large decrease of the 1975 nm band could again result from the involvement of Gln in intermolecular β-sheet formation or in interaction with water. Though, the band intensity at 1975 nm could be highly influenced by the nearby water band. On the other hand, the increase of the 2056 nm band could be related to an increased α-helix content (see Table 3.3).

Not all spectral changes in the protein spectrum may be explained solely from secondary structure changes, as also the increased aggregation and the interaction with water may influence the amide- and side chain absorptions (e.g. the amide II band is generally increased upon hydration). In an FTIR hydration study of poly-L-Pro, the hydration was shown to cause shift of the CH deformation bands to higher frequencies, and this was explained by the aggregation of Pro helices [Wellner, 1996]. As gluten is abundant in Pro, this could be relevant to the present experiment. Also a changed (hydrogen-bond like) interaction of CH from Trp, Tyr, Phe could underlie some of the spectral changes, as hydrogen bonding causes a high-frequency shift of the CH-str. bands. Also the S-H str. absorptions could change as disulfide interchanges take place.

Discussion and conclusion

In this experiment, gluten hydration is used as an example of a rather concentrated protein system (also containing starch and lipids), in which intermolecular interactions are developed during hydration. The experiment is useful for answering some of the questions posed in chapter 3, regarding NIR spectra of proteins in a complex matrix:

e. Are their any unique protein absorptions ? (distinguishable from starch and lipid absorptions)
f. Can protein conformation/interaction changes be detected ?
g. Is it possible to interpret the spectral changes and obtain structural information from NIR ?
h. Can information regarding protein-water interactions be obtained?

In the assignment of the peaks in 2nd derivative NIR spectra to protein, lipid and starch, some of the characteristic protein bands at 1975, 2054 and 2167-2208 nm were found only little overlapped by lipid and starch absorptions. Also some other bands in the overtone region (1400-1800 nm) were ascribed mostly to proteins, whereas the combination band region (2200-2400 nm), as expected, had no unique protein peaks. Therefore, only changes in a few bands may be indicative of protein conformation changes.

Upon gluten moistening, the increase of extended hydrated structures and intermolecular β-sheet (central to the gluten functionality) was evidenced from ATR-FTIR amide I spectra. In the 2nd derivative NIR spectra, some qualitative spectral changes were revealed upon moistening and were ascribed to qualitative changes in the constituents. The light scattering effects, which were influenced by the water content, were removed beforehand by use of 2nd derivative transformation and subsequent EMSC. Even though the more complex light scattering effects may not have been removed from this procedure, it seems unlikely that these effects could cause the very reproducible spectral changes upon moistening that were seen in the corrected spectra from several repeated experiments.

The protein conformation and interaction changes, provoked in this experiment, thus seemed to cause changes in some protein bands in the NIR spectra. However, the interpretation of the spectral changes solely as secondary structure changes was difficult, and some spectral changes rather suggested changes in the hydration of Gln side chains. The peaks at 1975 and 2182 nm in the 2nd derivative NIR spectra of gluten, which have been related to the primary amide groups from Gln, showed some significant changes upon moistening, in agreement with these side chains taking part in intermolecular β-sheet formation and water binding. Therefore, the NIR could possibly provide some information, which is more difficult to obtain from FTIR spectra, as the Gln absorptions in MIR spectra are overlaid by the amide I band. A wavelength region from 2250 to 2290 nm also showed several significant changes upon moistening of gluten. The starch and lipid contributions to these spectral changes could not be ruled out but also did not seem to explain all of the variation. As the spectral changes are not seen upon moistening of proteins in general, they could be related to the aggregation of gluten proteins and therefore be of importance to the gluten functionality. The differences from other proteins also emphasise that the results obtained in this study cannot be directly transferred to protein analyses of other food samples.

In conclusion, the above questions b-d are not finally answered from this experiment, which did not provide direct evidence but only indications. More experiments and analyses will be necessary, also in order to evaluate the NIR method for structure-functionality studies. The future analyses could include the measurements of NIR and MIR on the same samples with a range of moisture contents, whereby better correlations between the two spectral regions could be established. This was not attainable in the present experiment, as the moisture content could not be kept exactly constant for the two types of measurements. Other attempts to improve the interpretation of the NIR spectra could involve moistening of gluten preparations with a range of lipid and starch contents for better identification of these absorptions. Furthermore, it could be necessary to examine larger perturbations of the protein conformations (e.g. involving oxidation or reduction of disulfide groups).

4.5. References

Alava, J.M., Millar, S.J., Salmon, S.E. (2001). The determination of wheat breadmaking performance and bread dough mixing time by NIR spectroscopy for high speed mixers. *J. Cereal. Sci.* 33, 71-81.

Al-Azzam, W., Pastrana, E.A., Ferrer, Y., Huang, Q., Schweitzer-Stenner, R., Griebenow, K. (2002). Structure of poly(ethylene glycol)-modified horseradish peroxidase in organic solvents: Infrared amide I spectral changes upon protein dehydration are largely caused by protein structural changes and not by water removal per se. *Biophys. J.* 83, 36737-3651.

Alberti, E., Gilert, S.M., Tatham, A.S., Shewry, P.R., Gil, A.,M. (2002). Study of high molecular weight wheat glutenin subunit 1Dx5 by C-13 and H-1 solid-state NMR spectroscopy. I. Role of covalent crosslinking. *Biopolymers.* 67, 487-498.

Alzagtat, A. A., Alli, I. (2002). Protein –lipid interactions in food systems: A review. *J. Food. Sci. Nutr*. 53, 249-260.

Bean, S.R., Lookhart, G.L. (1998). Influence of salts and aggregation of gluten proteins on reduction and extraction of high molecular weight glutenin subunits of wheat. *Cereal Chem*. 75, 75-79.

Belderok, B. (2000a). Survey of gluten proteins and wheat starches. In: *Bread-making quality of wheat: A century of breeding in Europe,* (Donner, D.A, ed.). Kluwer Academic publishers, Dordrecht. pp. 30-39.

Belderok, B. (2000b). Genetic basis of quality in bread wheat. In: *Bread-making quality of wheat: A century of breeding in Europe,* (Donner, D.A, ed.). Kluwer Academic publishers, Dordrecht. pp. 55-66.

Belton, P.S., Colquhoun, I.J., Grant, A., Wellner, N., Field, J.M., Shewry, P.R., Tatham, A.S. (1995). FTIR and NMR-studies on the hydration of high-M(R) subunit of glutenin. *Int. J. Biol. Macromol.* 17, 74-80.

Belton, P.S. (1999). On the elasticity of wheat gluten. *J. Cereal. Sci.* 29, 103-107.

Chen R.Y, Psotka, J.J., Olewnik, M.C., Small, V.W., Seabourn, B.,Okkyung, K.C. (2002). Determination of effect of ingredients and levels thereof on characteristics of dough and batter-based products. *US. Patent Appl. Pub*. US 2002/0137216 A1.

Delwiche, S.R., Weaver, G. (1994). Bread quality of wheat flour by near-infrared spectrophotometry: feasibility of modeling. *J. Food Sci.* 59, 410-415.

Domenek, S., Morel, M.H., Bonicel, J., Guilbert, S. (2002). Polymerization kinetics of wheat gluten upon thermosetting. A mechanistic model. *J. Agr. Food Chem.* 50, 5947-5954.

Don, C., Lichtendonk, W., Plijter, J.J., Hamer, R. J. (2003a) Glutenin macropolymer: a gel formed by glutenin particles. *J. Cereal Sci.* 37, 1-7.

Don, C., Lichtendonk, W., Plijter, J.J., Hamer, R. J. (2003b). Understanding the link between GMP and dough: from glutenin particles in flour towards developed dough. *J. Cereal Sci.* 38, 157-165.

Don, C., Lichtendonk, W., Plijter, J.J., Vliet T.V., Hamer, R. J. (2005). The effect of mixing on glutenin particle properties: aggregation factors that affect gluten function in dough. *J. Cereal Sci.* 41, 69-83.

Edwards, N.M., Peressini, D., Dexter, J.E., Mulvaney, S.J. (2001). Viscoelastic properties of durum wheat and common wheat dough of different strengths. *Rheol. Acta.* 40, 142-153.

Elizalde, B.E., Pilosof, A.M.R. (1999). Kinetics of physico-chemical changes in wheat gluten in the vicinity of the glass transition temperature. *J. Food Eng.* 42, 97-102.

Feeney, K.A., Wellner, N., Gilbert, S.M., Halford, N.G., Tatham, A.S., Shewry, P.R., Belton, P.S. (2003). Molecular structures and interactions of repetitive peptides based on wheat glutenin subunits depend on chain length. *Biopolymers.* 72, 123-131.

Goesaert, H., Brijs, K., Veraverbeke, W.S., Courtin, C.M., Gebruers, K., Delcour, J.A. (2005). Wheat flour constituents: how they impact bread quality, and how to impact their functionality. *Trends Food Sci. Tech*. 16, 12–30.

Grant, A., Belton, P.S., Colquhoun, I.J., Parker, M.L., Shewry, P.R., Tatham, A.S., Wellner, N. (1999). Effects of temperature on sorption of water by wheat gluten determined using deuterium nuclear magnetic resonance. *Cereal. Chem*. 76, 219-226

Griebenow, K., Klibanov, A.M. (1995). Lyophilization-induced reversible changes in the secondary structure of proteins. *Proc. Natl. Acad. Sci. USA*. 92, 10969-10976.

Hamer, R.J., Hoseney, R.C. (1998). *Interactions: The Key to Cereal Quality*, AACC. Saint-Paul, Minnesota.

Hargreaves, J., Popineau, Y., Lemeste, M., Hemminga, M.A (1995). Molecular flexibility in wheat gluten submitted to heating. *FEBS.* 372, 103-107.

Hayta, M., Alpaslan, M. (2001). Effects of processing on biochemical and rheological properties of wheat gluten proteins. *Nahrung*. 45, 304-308.

Irissin-Mangata, J., Bauduin, G., Boutevin, B., Gontard, N. (2001). New plasticizers for wheat gluten films. *Eur. Polymer J.* 37, 1533-1541.

Kinsella, J. E., Hale, M.L. (1984). Hydrophobic associations and gluten consistency: Effects of specific anions. *J. Agr. Food Chem*. 32, 1054-1056.

Law, D.P., Tkachuk, R. (1977). Near infrared diffuse reflectance spectra of wheat and wheat components *Cereal Chem*. 54, 256-265.

Lefebvre, J., Pruska-Kedzior, A., Kedzior, Z., Lavenant, L. (2003). A phenomenological analysis of wheat gluten viscoelastic response in retardation and in dynamic experiments over a large time scale. *J. Cereal Sci.* 38, 257-267.

Lens, J.P., de Graaf, L.A., Stevels, W.M., Dietz, C.H.J., Verhelst, T.K.C.S., Vereijken, J.M., Kolster, P. (2003). Influence of processing and storage conditions on the mechanical and barrier properties of films cast from aqueous wheat gluten dispersions. *Ind. Crop. Prod.* 17, 119-130.

Lewis, R.N.A.H., McElhaney, R.N. (2000). Calorimetric and spectroscopic studies of the thermotropic phase behaviour of lipid bilayer model membranes composed of a homologous series of linear saturated phosphatidylserines. *Biophys. J.* 79, 2043-2055.

Lindsay, M.P., Skerrit, J.H. (1999). The glutenin macropolymer of wheat flour doughs: structure-function perspectives. *Trends Food Sci. Tech.* 10, 247-253.

Liu, K.Z., Mantsch, H.H. (2001). Apoptosis-induced structural changes in leukemia cells identified by IR spectroscopy. *J. Mol. Struct*. 565, 299-304.

Mejri, M., Rogé, B., Bensouissi, A., Michels, F., Mathlouthi, M. (2005). Effects of some additives on wheat gluten solubility: A structural approach. *Food Chem*. 92, 7-15.

Micard,V., Morel, M.H., Bonicel, J., Guilbert, S. (2001). Thermal properties of raw and processed wheat gluten in relation with protein aggregation. *Polymer*. 42, 477-485.

Millar, S., Robert, P., Devaux, M.F., Guy, R.C.E., Maris, P. (1996). Near-infrared spectroscopic measurements of structural changes in starch-containing extruded products. *Appl. Spectrosc*. 50, 1134-1139.

Pezolet, M., Bonenfant, S., Dousseau, F., Popineau, Y. (1992) Functional and solution states as determined by infrared-spectroscopy. *FEBS lett.* 299, 247-250.

Pevnser, A., Diem, M. (2001) Infrared spectroscopic studies of major cellular components. Part I: The effect of hydration on the spectra of proteins. *Appl. Spectrosc*. 55, 788-793.

Popineau, Y., Bonenfants, S., Cornec, M., Pezolet, M. (1994). Study by infrared-spectroscopy of the conformations of gluten proteins differing in their gliadin and glutenin compositions. *J. Cereal Sci*. 20, 15-22.

Rao, V.K., Mulvaney, S.J., Dexter, J.E. (2000). Rheological characterisation of long- and short-mixing flours based on stress-relaxation. *J. Cereal Sci.* 31, 159-171.

Sapirstein, H.D., Fu, B.X. (2002). Evidence for varying interaction of gliadin and glutenin proteins as an explanation for differences in dough strength of different wheats. Special Publications of the Royal Society of Chemistry.

Schofield, J.D. (1996) Wheat proteins: structure and functionality in milling and breadmaking. In: *Wheat. Production properties and quality,* (Bushuk, W., Rasper, V.F., eds). Chapman & Hall, London. pp. 73-99.

Seabourn, B.W. (2002). Determination of protein secondary structure in wheat flour-water systems during mixing using fourier transform horizontal attenuated total reflectance infrared spectroscopy. Ph.d. thesis, Kansas State University, Kansas.

Siesler, H.W., Ozaki, Y., Kawata, S., Heise, H.M. (2002). *Near-infrared spectroscopy. Principles, instruments, applications*. Wiley-VCH, Weinheim. pp. 338.

Tipples, K.R., Kilborn, K.H., Preston, R.H. (1996). Bread-wheat quality defined. In: *Wheat. Production properties and quality,* (Bushuk, W., Rasper, V.F., eds). Chapman & Hall, London. pp. 25-36.

Torres, P.I., Wazquez-Moreno, L., Ledesma-Osuna, A.I., Medina Rodriguez (2000). Contribution of hydrophobic soluble gluten proteins, fractionated by hydrophobic interaction chromatography in highly acetylated agarose, to dough rheological properties. *Cereal. Chem*. 77, 702-707.

Tronsmo, K.M., Faergestad, E.M., Schofield, J.D., Magnus, E.M. (2003). Wheat protein quality in relation to baking performance evaluated by the Chorleywood bread process and a hearth bread baking test. *J. Cereal Sci.* 38, 205-215.

Van de Weert, M. (2001). Fourier transform infrared spectrometric analysis of protein conformations: Effect of sampling method and stress factors. *Anal. Biochem*. 297, 160-169.

Van Velzen, E.J.J., van Duynhoven, J.P.M., Weegels, P.L., van der Maas, J.H. (2003). Factors associated with dough stickiness as sensed by attenuated total reflectance infrared spectroscopy. *Cereal. Chem*. 80, 378-382.

Veraverbeke, W.S., Delcour, J.A. (2002). Wheat protein composition and properties of wheat glutenin in relation to breadmaking functionality. *Crit. Rev. Food. Sci.* 42, 179-208.

Wang, M., Hamer, R.J., van Vliet, T., Oudgenoeg, G. (2002). Interaction of water extractable pentosans with gluten protein: Effect on dough properties and gluten quality. *J. Cereal. Sci.* 36, 25-37.

Wellner, N., Belton, P.S., Tatham, A.S. (1996). Fourier transform IR spectroscopic study of hydration-induced structure changes in the solid state of ω-gliadins. *Biochem. J.* 319, 741-747.

Wellner, N., Bianchini, D., Mills, E.N.C., Belton, P.S. (2003). Effect of selected Hofmeister anions on the secondary structure and dynamics of wheat prolamins in gluten. *Cereal Chem.* 80, 596-600.

Wellner, N., Mills, E.N.C., Browsney, G., Wilson, R.H., Brown, N., Freeman, J., Halford, N.G., Shewry, P. R., Belton, P.S. (2005). Changes in protein secondary structure during gluten deformation studied by dynamic Fourier transform infrared spectroscopy. *Biomacromolecules*. 6, 255-261.

Wesley, I.J., Larsen, N., Osborne, B.G., Skerritt, J.H. (1998). Non-invasive monitoring of dough mixing by near infrared spectroscopy. *J. Cereal Sci.* 27, 61-69.

Wesley, I.J., Blakeney, A.B. (2001). Investigation of starch-protein-water mixtures using dynamic near infrared spectroscopy. *J. Near infrared Spec*. 9, 211-220.

Wrigley, C.W., Bietz, J.A. (1988). Proteins and amino acids. In: *Wheat: Chemistry and Technology,* (Pomeranz, Y., ed.). AACC, St. Paul, Minnesota. pp. 159-275.

Chapter 5: Study of gluten interactions and functionality part 2. Salt effects.

An NIR experiment, involving alterations of the gluten protein conformations and interactions by use of various salts, was carried out with the purpose of investigating the ability of NIR to monitor the protein perturbations in a complex sample of high water content. The NIR spectra were interpreted based on the corresponding ATR-FTIR spectra and *a priori* knowledge of the salt effects. The effects of salts on protein conformations and interactions and the specific effects on gluten functionality are described in section 5.1, whereas the experiment is described in section 5.2

5.1. Salt effects on protein conformations and interactions: Effects on doughs.

NaCl are commonly added to wheat doughs with the purpose of improving the flavour, dough handling properties and the baking result, as addition of NaCl decreases the water absorption, increases dough strength and extensibility and results in larger bread volume [Preston, 1989; He, 1992; Butow, 2002]. Other salts may as well improve the baking result, while some disrupt the gluten functionality and deteriorate the bread quality at high salt concentrations [Preston, 1989; He, 1992]. It has been shown that salts primarily affect the gluten protein hydration, while starch hydration is less affected [Wellner, 2003].

The influence of salts on protein conformations is explained from the Hofmeister series (see chapter 2.4). Ions in the Hofmeister series are arranged according to their salting-in and salting-out properties i.e. their abilities to increase or decrease, respectively, the solubility of a solute (as salts with a low salting-out property have a high salting-in property). The salting property is defined in the Setschenov equation (Eq. 5.1), which outlines a linear relation between the logarithm of the solute solubility c_p and the salt concentration c_s [Grover, 2005; Baldwin, 1996].

Eq. 5.1. $$\ln\left(\frac{c_p}{c_p(0)}\right) = -K_s c_s \qquad (\text{for } c_s > 0.5\text{ M})$$

In Eq. 5.1, $c_p(0)$ is the solubility in water. The proportionality factor is termed the salting coefficient K_s and reflects the interaction between a hydrophobic solute and the specific salt. A negative K_s implies a higher solubility of the solute with increasing salt concentration (salting-in), whereas a positive K_s implies a decreased solubility (salting-out) [Grover, 2005; Kalra, 2001]. Hydrophobic solutes are salted-out by most salts, but for proteins, the ions may salt-in the peptide groups and other polar groups [Baldwin, 1996].

The Hofmeister effects have been related more to ion-surface interactions than to the effect on bulk water structure and is probably the result of an interplay between various factors such as water structure, ionic dispersion forces, ion sizes, co-ion exclusion, hydration forces etc. [Boström, 2004, 2005a,b; Kunz, 2004].

Salt effects at high concentrations

Anion effects: At high salt concentrations (>0.3 M) the salting-in and salting-out of non-polar groups by chaotropic and kosmotropic anions, respectively, is commonly believed to be an indirect effect, caused by the effect on the hydrogen bonding properties of water and thereby on the hydrophobic interactions [Baldwin, 1996]. The salting constant K_s shows correlation to the surface tension increment of the salt (which reflects the kosmotropic/chaotropic property) and is related to the entropic cost of forming a cavity in the water in order to accommodate a hydrophobic solute [Baldwin, 1996]. Likewise, the ability of an ion to accommodate the hydrophobic solute in its hydration shell is suggested a primary factor in the salting-out mechanism. The exclusion of the hydrophobic solute from the volume occupied by the ion and its first coordination shell increases the concentration of the hydrophobic solute in the remaining solvent and promotes the hydrophobic interactions. The exclusion is most effective for the kosmotropic ions, which bind water tightly [Hribar, 2002; Kalra, 2001], whereas the chaotropic anions show low exclusion of the hydrophobic solutes and instead are able to associate to the hydrophobic surfaces [Baldwin, 1996; Di Stasio, 2004]. The solubilising/destabilising effect of the chaotropic anions has been explained from their direct interaction with the exposed peptide bonds on the unfolded form of the protein [Baldwin, 1996] or with some amino acid side chains [Di Stasio, 2004]. The salting-in effect of the most chaotropic ions can also be explained from their high ability to associate to the protein surfaces [Ebel, 1999].

For gluten, the anion effect depends on both protein quantity and quality and is quite complex [He, 1992; Butow, 2002]. In a study by Preston et al (1989), it was found that increasing amounts of the chaotropic ClO_4-, I- and SCN- (as sodium salts) increased the water absorption of the dough, and that the highly chaotropic I- and SCN- in addition decreased the development time and the tolerance against overmixing dramatically. Kinsella et al (1984) also noticed accelerated hydration and decreased dough stability from 1M SCN^-. Wellner et al (2003)showed that the chaotropic NaI and NaBr at 1 M decreased the intra- and intermolecular β-sheet content compared to water and compensated for this by increasing the β-turn content. Thus, they suggested that the chaotropic anions reduce the amount of 'train' regions (which contain both hydrophobic and hydrogen bond interactions) and in turn increase the more hydrated 'loop' regions in gluten. This is similar to an increased solubility. Nevertheless, in the study by Preston et al (1989), the 0.5 M chaotropic salts were found to result in increased aggregation (measured by extensograph maximum height) compared to water and 0.5 M NaCl. This was attributed to the denaturing effect of chaotropic anions and increased hydrophobic interactions between the unfolded chains. Only at higher concentrations could this effect be overcome by the solubilising effect .

In opposition to the chatropic anions, increasing amounts of the neutral Cl^- leads to increased dough development time [Preston, 1989]. Also, the kosmotropic anion SO_4^{2-} causes an increased dough elasticity at increasing salt concentrations, and above 0.1 M, the dough has even been found too elastic for breadmaking [He, 1992]. This reflects the ability of the kosmotropic anions to stabilise the associated forms, due to the promotion of hydrophobic- and thereby also hydrogen bonding interactions. An increased aggregation and resistance to hydration was also noticed by Kinsella et al (1984) for the kosmotropic F-, which also caused a decreased dough consistency. This was related to an impeded protein unfolding during hydration.

Cation effects: There is no clear correlation between the surface tension increment and protein stabilisation for cations [Boström, 2005a; Ebel, 1999]. Instead the direct interaction with the protein is more important, and for example, Arakawa et al showed that the binding of divalent cations to proteins could counteract the surface tension effect. A destabilising effect of kosmotropic cations on proteins is thus explained by their ability to interact with the peptide groups, as seen for chaotropic anions [Arakawa, 1984; Ebel, 1999]. Eggers et al (2001) suggested that some of the effects of cations are actually executed by their influence on the anions, i.e. kosmotropic cations may neutralise the protein stabilising effect of the kosmotropic anions by binding to these. The influence of the kosmotropic cations Ca^{2+} and Li^+ on doughs has been studied by He et al (1992), who found that increasing concentrations of $CaCl_2$ and LiCl decreased the loaf volume, and that LiCl also decreased the dough extensibility in agreement with the described destabilising effect of kosmotropic cations. On the other hand, KCl increased the bread volume to the same extent as NaCl.

Salt effects at low concentrations

At low salt concentrations (< 0.15 M), the neutralisation of charges by electrostatic ion-protein interactions affects the protein solubility. Increased solubility is seen if the ions replace electrostatic intra- or interprotein interactions [Di Stasio, 2004]. In the case the net charge of the protein is abolished by screening, the result may oppositely be a decreased solubility, as electrostatic repulsion is decreased and more protein-protein interaction is allowed [Di Stasio, 2004].

The effect on dough properties of anions at low concentrations has been demonstrated in a study involving sodium salts of Cl^-, Br^-, ClO_4^-, I^- and SCN^-. Both the neutral NaCl and the chaotropic salts at 0.1 M tended to increase the dough development time, but the most chaotropic anions were most effective [Preston, 1989]. Some of the effects could be related to the screening of positive charges on the surface of the gluten proteins and the thereby increased aggregation. However in an ATR-FTIR study, increased intermolecular β-sheet content was shown only for Cl^- and Br^- at concentrations up to 0.2 M, whereas the more chaotropic I^- decreased the intermolecular β-sheet content even at the low concentrations [Wellner, 2003]. Thus, the increased aggregation at low concentrations of I^- and SCN^- is explained from an increased hydrophobic interaction between unfolded proteins and not increased intermolecular β-sheets. Although the dough development time was increased for these salts, the tolerance against overmixing was decreased, and the doughs had decreased breadmaking properties [Preston, 1989]. The weakly kosmotropic Li^+, the neutral Na^+ and the chaotropic K^+ were found to increase the dough strength, but again the effect followed the Hofmeister series, so the most chaotropic cations were the most effective (K^+>Na^+>Li^+) [Butow, 2002]. The stabilising effect was related to their coordination to the proteins. They speculated that the less hydrated K+

could have a higher ability to integrate into the gluten matrix and complex to functional groups on the folded protein [Butow, 2002].

The effects of low and high salt concentrations on dough property and gluten structure are summarised in Table 5.2.

	Anions		Cations	
conc.	Chaotropes	Neutral/Kosmotropes	Chaotropes	Kosmotropes
High	Trains ↓ Loops↑ Denaturation, aggregation Development time ↓ stability ↓	Trains ↑ Loops↓ Development time ↑ Elasticity ↑	Bread volume ↑	Extensibility ↓ Bread volume ↓
Low	Trains ↑ for weak chaotropes Trains ↓ for strong chaotropes Development time↑ stability ↓	Trains ↑ (for Cl^-) Elasticity↑ Extensibility ↑	Dough strength ↑	Dough strength ↑

Table 5.2. Effects of anions and cations at low and high concentration on dough properties and gluten structure. The table summarises the text above.

5.2. Experiment V: NIR analysis of protein structure and interactions in gluten

In this experiment, gluten samples are hydrated in various salt solutions of different chaotropic and kosmotropic properties, so as to vary protein conformations and protein-protein interactions. The effects on gluten protein structure and interactions are analysed by use of the amide I band obtained in ATR-FTIR measurements. The purpose is to investigate the ability of NIR to detect small structural and interaction changes of proteins in a complex sample with a high water content, similar to the changes in food samples.

Methods

Different salts solutions (Na_2SO_4, $MgSO_4$, $MgCl_2$, NaCl, $NaClO_4$, KBr, $MgBr_2$) were prepared at 0.1, 0.2, 0.5 and 1.0 M concentrations in mili-Q water. For each salt solution, 25 ml was added to an aliquot of 10 mg gluten powder (obtained from Sigma), and the samples were mixed until homogeneity. The effect of mixing time was investigated by mixing of two water- and two 1 M NaCl samples for 2-3 min. After storage for ~4 hours at room temperature, the samples were centrifuged 15 min at ~340 g, and excess solution was discarded, where after the water uptake was calculated from the sample weights. The samples were stored at 4°C until few hours before the measurements, which were performed one or two days after sample preparation. Replicates were prepared for the water- and most of the 1M salt samples, and these preparation replicates were measured on different days. Immediately before the NIR or MIR measurements, two to four slices were cut from each lump of gluten and the slices were measured on both instruments. The slices were kept in a closed container in the time between the NIR and MIR measurements. Two to four measurement replicates were obtained in ATR-FTIR, and at least five measurement replicates were obtained in NIR. Measurement replicates include different gluten slices from the same lump of gluten and different placements of the samples. The spectra included in the following analyses are shown in Fig. 5.1.

MIR	water	NaCl	$MgCl_2$	$MgSO_4$	Na_2SO_4	KBr
	•••••••••••••					
0.1		••	•••	••••	••••	••
0.2		••	•••	••••	••••	••
0.5		••	•••	••••	••••	••
1.0		••••••••••••••	•••••••	••••	••••••••	••••
NIR	water	NaCl	$MgCl_2$	$MgSO_4$	Na_2SO_4	KBr
	••••••••					
0.1		•••••••••	•••••	•••••	•••••	•••••
0.2		•••••••••••	••••••	•••••	•••••	•••••
0.5		•••••••••••	••••••	•••••	•••••	•••••
1.0		•••••••••••••••••	•••••	••••••	••••••••••	•••••

Fig. 5.1. Overview of obtained spectra in the experiment. A) ATR-FTIR spectra. B) NIR spectra.

The spectroscopic analyses did not include gluten hydrated in $MgBr_2$, since these samples were not easily measurable. The FTIR spectra from 4000 cm^{-1} to 748 cm^{-1} were recorded on a Bomen spectrometer equipped with a horizontal ZnSe ATR-crystal at a resolution of 4 cm^{-1} and with coaddition of 128 scans. The data interval was 1.93 cm^{-1}. The NIR spectra from 790 nm to 2500 nm were obtained on a Perkin Elmer Spectrum One FT-NIR spectrometer equipped with a reflectance accessory and an InGaAs detector, at a resolution of 16 cm^{-1} and with co-addition of 100 scans. The data interval was 1.67 nm. Each slice was placed directly on the instrument without the use of a sample holder. For the pure salt solutions, a transflector plate was used, with the purpose of directing the light back to the detector. No temperature control was applied in either the ATR-FTIR or NIR measurement series.

Spectral pretreatments (as described in the result section) as well as PCA and PLSR analyses were carried out in Unscrambler 9.2 (Camo).

Results: Effect of salts on water structure

The kosmotropic and chaotropic properties of the salts are reflected in their effect on the water spectrum, as described in chapter 2. The influence on the HOH bending band ν_2 at 1635 cm^{-1} was studied in a PCA analysis, carried out from 1700-1600 cm^{-1} after EMSC correction. The results are shown in Fig. 5.2. The PC1 loading vector indicates a shift of the water band (Fig. 5.2C). Comparing to the score plot (Fig, 5.2A), a shift to lower frequencies occurs with increasing amounts of the bromide salts, and the opposite shift takes place for the sulphate salts in agreement with the chaotropic and kosmotropic property of Br^- and SO_4^{2-}, respectively. However, also the cations influence the PC1 variation, especially Mg^{2+} as expected from its high charge density.

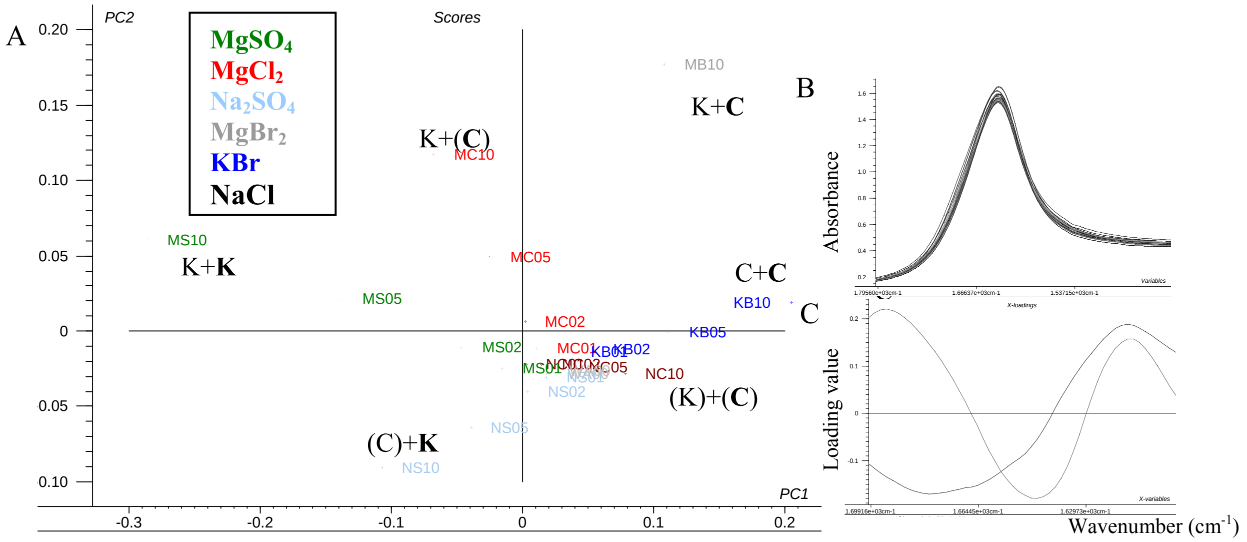


Fig. 5.2. A) Score plot from a PCA, including salt solution spectra (0.1-1.0 M) in the water bending region 1700-1600 cm^{-1}. K=kosmotropic, (K)=weakly kosmotropic, C=chaotropic, (C)=weakly chaotropic. PC1 and PC2 explains 72 % and 24 % respectively. B) Water bending band. C) PC1 and PC2 loading vectors.

PC2 also shows an effect of the magnesium salts. The small effects of NaCl and Na_2SO_4 seen in the score plot are in agreement with NaCl being regarded as a neutral salt in the Hofmeister series and with sulphate being described as fitting into the water structure without disturbing it much.

Results : Water uptake and gluten consistency

The water-uptake depended on the salt type and concentration as shown in Fig. 5.3.

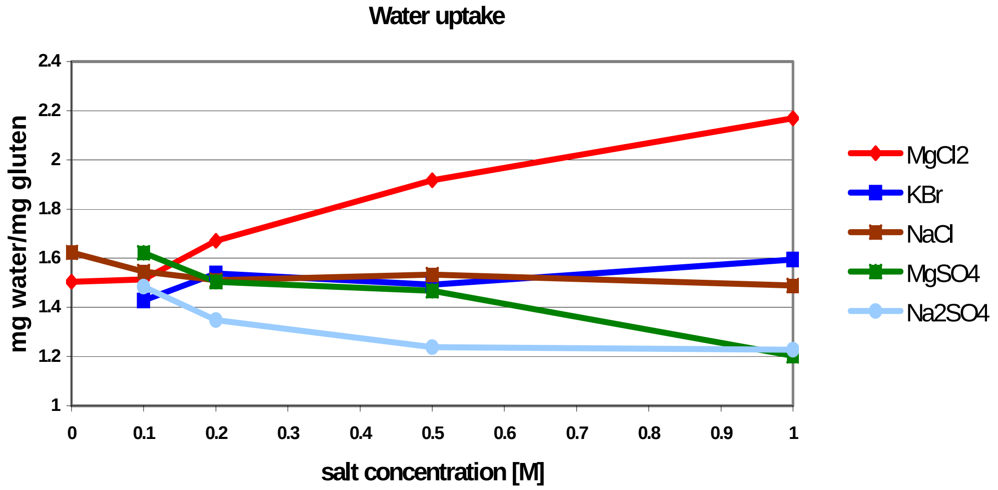


Fig. 5. 3. Water uptake for gluten prepared in different salt solutions. The apparent water uptake includes water trapped in the network and water bound to the proteins. The data series are from single measurements and the figure does not show the replicate variations.

Gluten hydrated in 1.0 M $MgBr_2$ (not shown in Fig 5.3) absorbed ~4 mg water pr mg gluten, while other water absorptions ranged from 1.2 to 2.2 mg water pr mg gluten. Opposite effects of $MgCl_2$/$MgBr_2$ and

Na_2SO_4/$MgSO_4$ on the water uptakes are seen at salt concentrations of 0.2 M and higher. The effects at 0.1 M salt concentrations are small and do not follow the same trends as seen at higher concentrations.

The visible effects on gluten of the different salts (at 1.0 M) are summarised in Table 5.2. Gluten hydrated in 1.0 M Na_2SO_4 and 1.0 M $MgSO_4$ solutions appeared firm, elastic and with reduced extensibility (Na_2SO_4 appeared to have a higher effect than $MgSO_4$). The usual cohesive and homogenous network was disrupted and the samples appeared as compositions of small particles. The effects of $MgCl_2$ and $MgBr_2$ were obviously different from those of $MgSO_4$ and Na_2SO_4. At increasing $MgCl_2$ concentration, the gluten lump became more and more jelly-like, and at 1.0 M salt the sample was soft and sticky. For $MgBr_2$, these effects were enhanced. The effects of NaCl and KBr on the gluten appearance were not significant.

salt	Cation	Anion	Effect on gluten	Water uptake	Appearance
Na_2SO_4	(k)	K	High	Decreased	Firm
$MgSO_4$	K	K	Medium	Decreased	Firm
NaCl	(k)	(c)	Low	No effect	Normal
KBr	C	C	Low	No effect	Normal
$MgCl_2$	K	(c)	High	Increased	Jelly-like, sticky
$MgBr_2$	K	C	High	Increased	Jelly-like, slimy, sticky

Table 5.2. Salt effect on gluten at 1.0 M salt concentration. K=kosmotrope, (k)= weak kosmotrope, C=chaotrope, (c)= weak chaotrope.

The results in Table 5.2 may be interpreted in terms of a stabilising or destabilising effect if water uptake and firmness of the gluten lump are used as indicators, i.e. less binding of water indicates a more compact folding or higher degree of protein-protein interaction. At high salt concentration, a stabilising effect is obtained for the kosmotropic SO_4^{2-} irrespective of the nature of the counter ion (Na^+ or Mg^{2+}), whereas for the weakly chaotropic anion Cl^-, no stabilisation is observed: NaCl did not affect water uptake much above 0.2 M, and $MgCl_2$ even caused destabilisation, perhaps caused by binding of Mg^{2+} to peptide groups. The highly chaotropic anion Br^- caused profound destabilisation as expected, but only when Mg^{2+} was counter ion (and not K^+). The very different outcomes of using $MgBr_2$ and KBr make clear that also cations has high influence on the gluten structure. The destabilising effect of $MgBr_2$ is in agreement with the theory that kosmotropic cations (Mg^{2+}) may cause salting-in, but this effect may apparently be overruled by a stabilising effect of a kosmotropic anion (such as SO_4^{2-}).

The underlying secondary structure changes in the gluten proteins are seen from ATR-FTIR analyses.

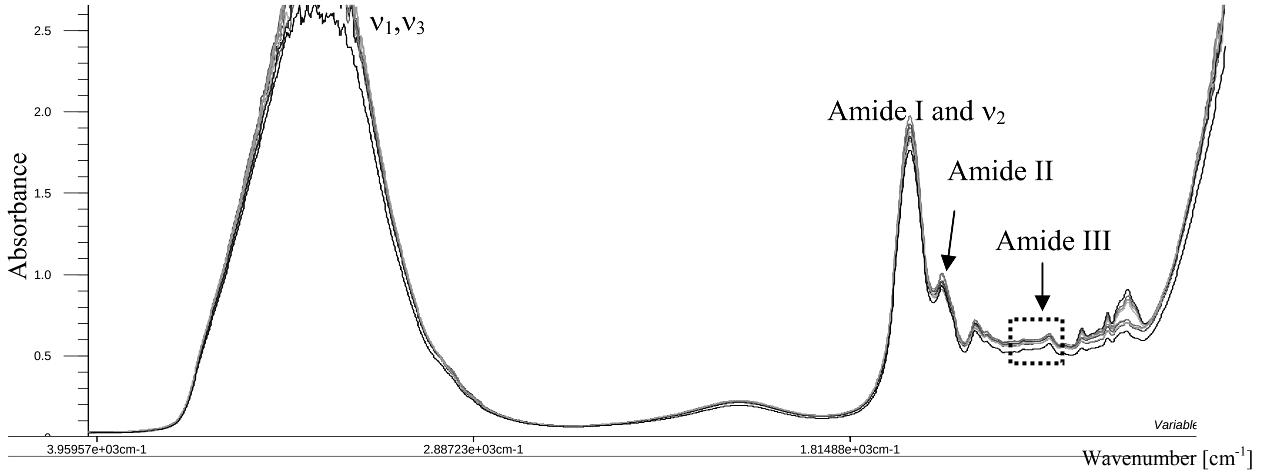


Fig. 5.4. ATR-FTIR spectra of gluten hydrated in water.

ATR-FTIR spectra of gluten hydrated in water are shown in Fig. 5.4. Some replicate variability is seen. This could stem from variations in the contact to the ATR-crystal and from different amounts of surface water (communication with Nicolaus Wellner). Also, adsorption of gluten constituents to the crystal may be another cause of spectral variations that reduce the repeatability of the ATR-FTIR measurements. In addition, the amide I analysis was hampered by a high absorbance in the amide I region, resulting in a low SNR, and by the overlapping water band ν_2. The subtraction of the water band in ATR-FTIR-spectra can be very complicated, due to the dependence of the penetration depth on frequency and refractive index [Wang, 1996]. Therefore, the amide III region, which is not influenced so much by the water absorptions [Fu, 1994], was also analysed for gaining structural information. However, also the EMSC correction is attempted for removal of the water band variations.

Amide III band analyses

Even in the amide III region, the salts caused some variations in the water spectra that related to salt type and concentration. Standard EMSC correction removed most of these baseline effects (not shown). However, an intensive sulphate S-O stretching absorption around 1100 cm^{-1} gave rise to a sloping baseline for the sulphate salt solutions (Fig. 5.5C), and apparently some other variations discriminated between the different salt solutions (seen in a PCA) . The usual subtraction of the buffer spectra could not be carried out in a proper manner and in an attempt to remove this background variation, EMSC with Bad-spectrum subtraction (EMSC-BS) was carried out. The PCs obtained in the PCA on the different salt solution spectra were applied as ‘Bad Spectra’ and subtracted in an EMSC correction of the gluten spectra in the 1350-1200 cm^{-1} range. Some small effects of this correction were seen, e.g. in the Na_2SO_4 –gluten spectra (Fig. 5.5C).

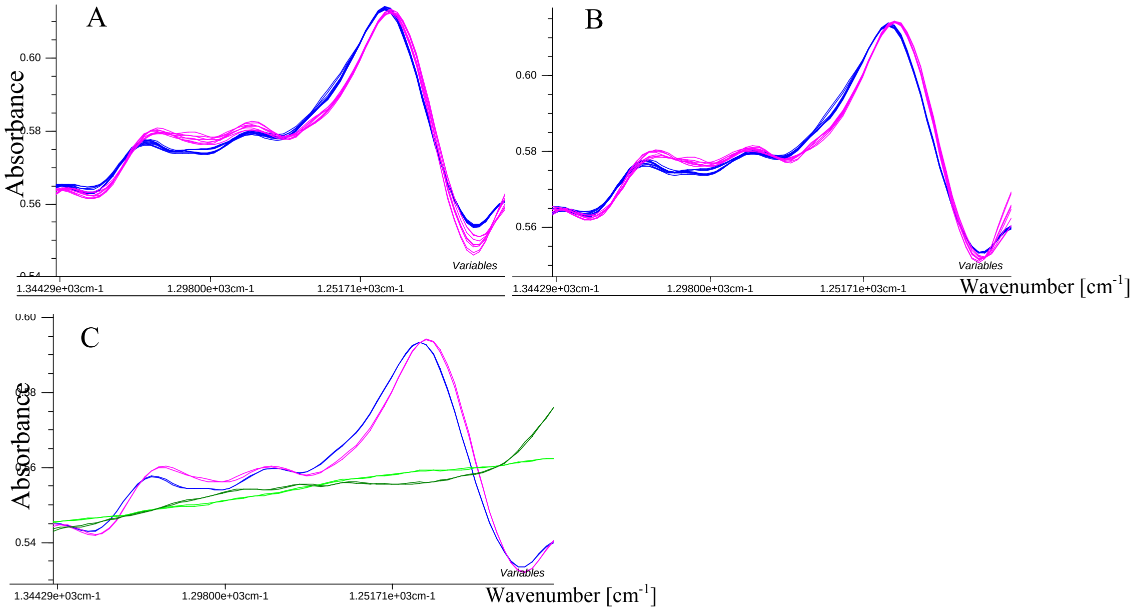


Fig. 5.5. Amide III region of gluten hydrated in water (blue) or in Na2SO4 1.0 M (pink). A) Spectra after standard EMSC correction. B) Spectra after EMSC-BS (described in the text). C) Spectra after EMSC-BS -both gluten and salt solution spectra are shown. Light green=water. Dark green= 1.0 M Na_2SO_4.

Not all background variation could be removed in this EMSC processing (the remaining background variation for Na_2SO_4 solutions is compared to the Na_2SO_4 gluten spectrum variations in Fig. 5.5C).

From the EMSC-BS data, some effects of the different salts at 1.0 M concentration could be seen (all amide III spectra are shown in Appendix V-1, Fig. V-A). The two sulphate salts caused the largest effects compared to water, and both salts caused an increase at 1319-1274 cm^{-1} and narrowing of the 1242 cm^{-1} band together with a shift to lower frequencies of this band (see Fig. 5.5A,B). These differences are similar to those seen between moist and dry gluten in Experiment IV and indicate that the sulphate salts cause some drying out of gluten proteins (in accordance with the low water uptake). The increase at 1319-1274 cm^{-1} could suggest increased amount of α-helix (see assignments to secondary structures in Fig. 4.10).

A PCA analysis that included the remaining salts ($MgCl_2$, KBr and NaCl) at high concentrations was carried out (Fig. 5.6).

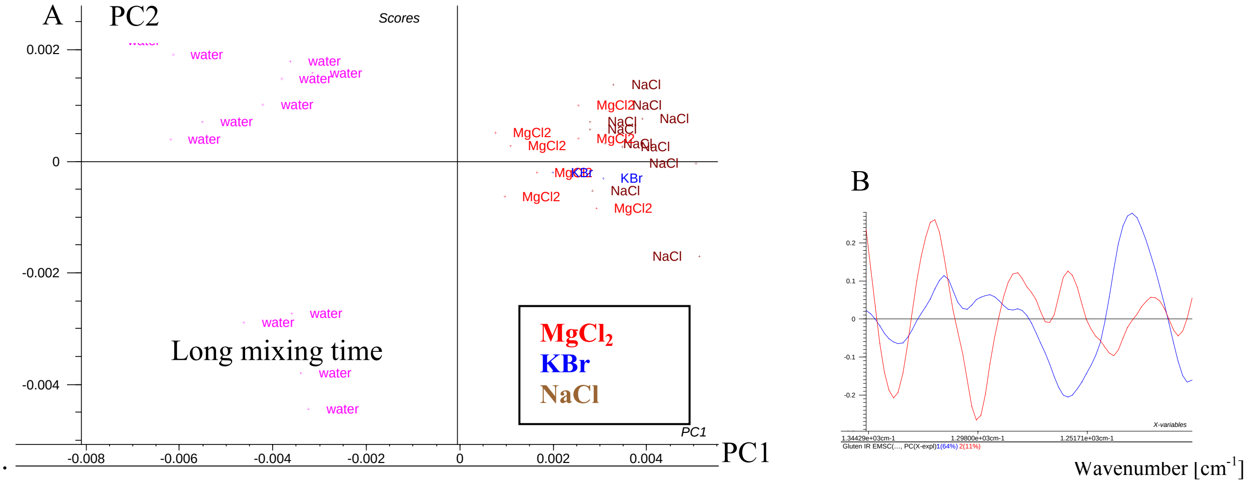


Fig. 5.6. Results from a PCA analysis of the EMSC-BS corrected amide III region. Samples included gluten hydrated in water or 1M salt solutions. A) PC1 vs. PC2 score plot. B) Loading plots. Blue=PC1. Red=PC2.

The score plot in Fig. 5.6 shows an effect of all three salts (compared to water) in PC1, which explains 64 % of the spectral variation. The PC1 loading plot indicates shifts from 1260 to 1230 cm^{-1} and from 1328 to 1311 cm^{-1} as a result of the salts. The shifts to lower frequencies parallel the effects of the sulphate salts, and the indication is that all salts cause some similar changes (compared to water). The different mixing times influence the PC2 variation, which explains 11 % of the variation.

In conclusion from the amide III band analyses, all salts affected the gluten structure in a different way and more extensively than the prolonged mixing. The largest effect was seen from the sulphate salts. The effects of $MgCl_2$, KBr and NaCl on the amide III band were very small but similar to the effects of sulphate salts, even though they seemed to affect gluten less.

Amide I band analysis

The effects of the salts at high concentrations were analysed further from amide I analysis. The gluten amide I spectra were pretreated with EMSC, 2nd derivative, baseline correction and finally mean normalisation.

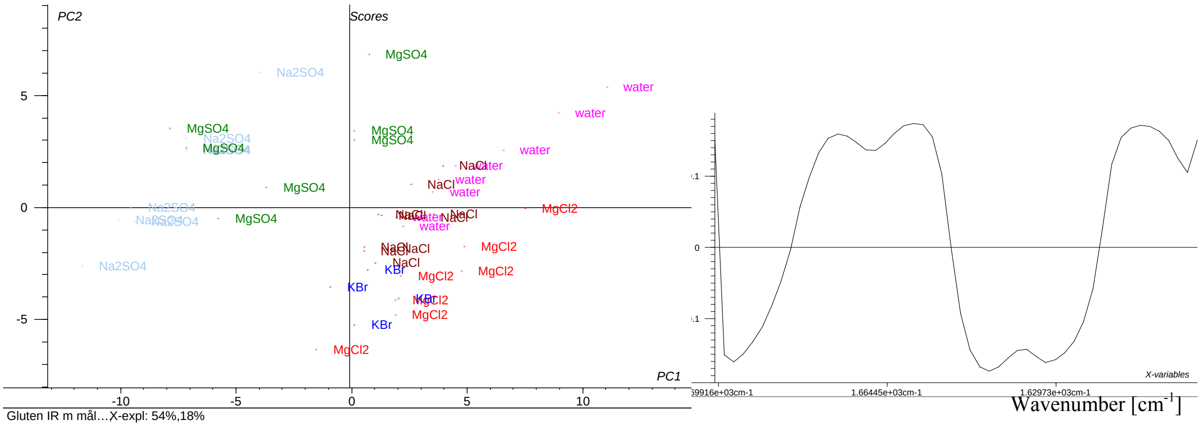


Fig. 5.7. Results from a PCA, including the amide I bands (1700-1600 cm^{-1}) of gluten hydrated in different 1.0 M salt solutions or water. A) Score plot of PC1 vs. PC2. B) PC1 loading. PC1 and PC2 explains 54 and 18 % of the variation, respectively.

A PCA analysis of the pretreated spectra was carried out (results are seen in Fig. 5.7). Comparison of the score plot in Fig. 5.7A to the score plot in Fig. 5.2A of salt solutions reveals that the water spectrum variations is not the dominating effect in the gluten amide I spectra as the two plots are quite different. Opposite effects on the amide I band of the sulphate salts and $MgCl_2$ (in PC1) are seen but only small effects of NaCl and KBr.

In Fig. 5.8. the amide I spectra of gluten in different 1.0 M salt solutions are compared to those of gluten in water. Likewise, the solution spectra are compared to the water spectrum. It is clear that some variations in the amide I band for the different gluten samples may be caused by variations in the water contents and variations in the water spectra. In Paper III, the separation of the variations in the gluten spectrum from those in the water spectrum was attempted by the described EMSC#2 method. By this method, the water spectrum variations caused by temperature and salts were subtracted as well as the gluten spectrum and the light scattering, leaving only small residuals, which should contain the variations due to structure changes. However, these residuals are not analysed thoroughly in the present study due to a shortness of time, but they are presented in Appendix V, Fig. V-B. The spectral changes revealed for gluten in 1.0 M $MgSO_4$ and 1.0 M NaCl from these data seem somewhat in agreement with the changes in the spectra in Fig. 5.8. Therefore, the changes in Fig. 5.8. are interpreted as secondary structure changes in the following.

Seen from Fig 5.7 and Fig. 5.8, 1.0 M NaCl have a small tendency to increase the absorptions between 1650 and 1620 cm^{-1} and to decrease the intensities above 1650 cm^{-1}. These changes could stem from increased random structure (1645 cm^{-1}), β-sheet (1630 cm^{-1}) and perhaps intermolecular β-sheet (1620-1625 cm^{-1}) and instead less α-helix (1654 cm^{-1}) and β-turn (1666 cm^{-1}). The interpretations of the amide I changes caused by the salts at 1M concentration are summarised in Table 5.3.

1M salt	Increase	Interpretation	Decrease	Interpretation
Na_2SO_4	1649-1620	**random (1645)** **β-sheet (1630)** intermolecular β-sheet (1620)	~1685-1650 1616	**β-turns** (1666), α-helix (1654) **extended hydrated structure (1616)**
$MgSO_4$	1649 1633-1616	**random** β-sheet intermolecular β-sheet	~1680-1650	**β-turns,** α-helix
NaCl	1650-1616	random β-sheet intermolecular β-sheet	~1685-1650	β-turn, α-helix
KBr	1633-1616	β-sheet intermolecular β-sheet	~1685-1647	β-turn, **α-helix, random**
$MgCl_2$	1633-1616	β-sheet intermolecular β-sheet	~1700-1664	β-turn

Table 5.3. Amide I interpretations based on the spectra in Fig. 5.8. The effects compared to gluten hydrated in water are shown. Large effects are shown in bold.

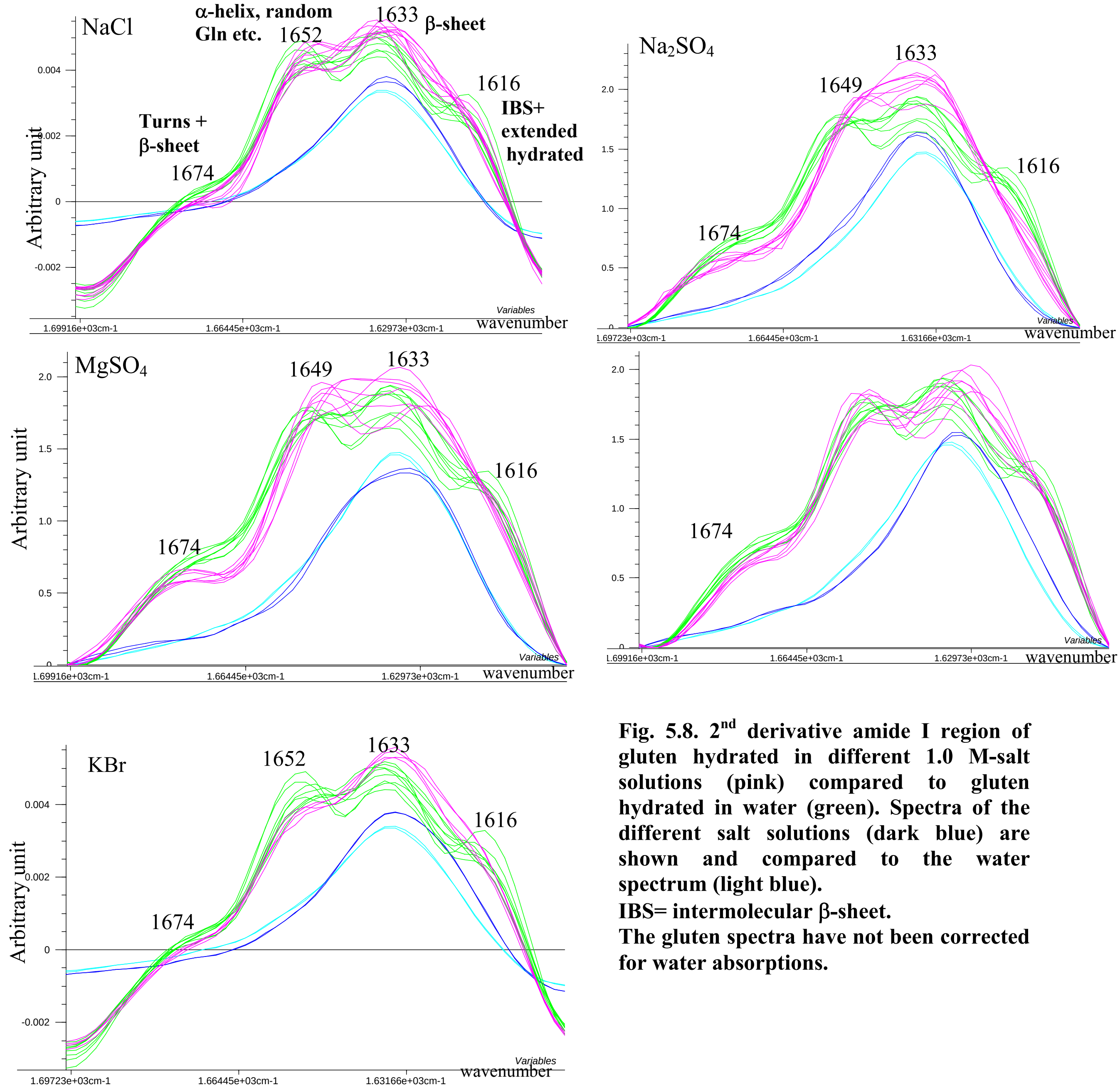


Fig. 5.8. 2nd derivative amide I region of gluten hydrated in different 1.0 M-salt solutions (pink) compared to gluten hydrated in water (green). Spectra of the different salt solutions (dark blue) are shown and compared to the water spectrum (light blue).

IBS= intermolecular β-sheet.

The gluten spectra have not been corrected for water absorptions.

From the described effects of salts in section 5.1, it could be expected that salts in the upper part of Table 5.3 (Na_2SO_4, $MgSO_4$) would increase the train/loop ratio in gluten, whereas the salts in the lower part ($MgCl_2$) would decrease this ratio. In agreement with this hypothesis, the sulphate salts at 1 M are found to decrease the β-turns absorptions at 1666 cm^{-1}, indicating less loop structure. However, also 1M $MgCl_2$ causes a similar and unexpected, though, much smaller decrease at this wavenumber. Also, all salts seemed to increase the absorption in the 1630-~1616 cm^{-1} region, indicative of intermolecular β-sheet and extended structures. This does not agree with the study by Wellner et al (2003), in which the chaotropic anions were found to decrease the intermolecular β-sheet content. However, in the present study the counter ion K^+ may have abolished the effect of the chaotropic Br^-, as K+ has been found to increase the gluten aggregation [Butow, 2002]. A distinctive effect of the kosmotropic sulphate salts is a high absorbance at 1649 cm^{-1}, which could be explained from conversion of α-helix into random structure. The random structure could result from an impairment of the unfolding of gluten proteins during hydration, or it could be associated with the increased intermolecular β-sheet, as Wellner et al (2005) saw that stress relaxation after mixing caused some of the formed intermolecular β-sheet to revert to random structure. They also found that a decrease of α-helix content accompanies the decrease of β-turn content upon mixing or drying of gluten. Thus, the simultaneous decreases of β-turns, α-helix and hydrated extended structures (1616 cm^{-1}) observed for the Na_2SO_4 are both in agreement with the lower water contents of gluten hydrated in this salt solutions.

The above results are not in accordance with the amide III analysis, which indicated increased α-helix for the sulphate salts, and it should be considered that both regions may be affected by perturbations of amino acid side chain absorptions.

Results: NIR analyses

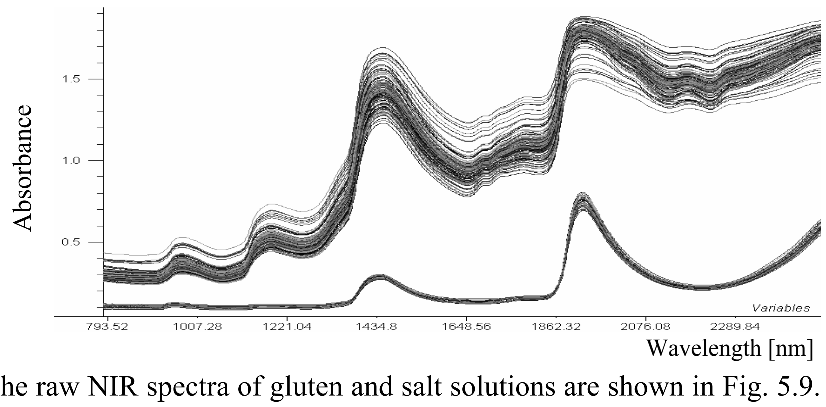


Fig. 5.9. NIR spectra obtained in reflectance mode of gluten hydrated in different salt solutions. Salt solution spectra obtained in transflectance mode are shown below.

The raw NIR spectra of gluten and salt solutions are shown in Fig. 5.9. The pretreatments of the NIR spectra included 2nd derivative transformation and subsequent EMSC in selected regions (1175-1320 nm, 1480-1750 nm and 1960-2360 nm), for removal of physical and quantitative variations. (Removal of the water band variations by use of EMSC with the 2nd derivative water spectra used as "Bad spectra" was not possible for this data set, as the position of the water band at 1930 nm was different in case of gluten and salt solutions, see Appendix V, Fig. V-C). All pretreated spectra are shown in Appendix V, Fig. V-D. An overview of the effects of 1M salts was obtained by submitting the three pretreated regions to a combined PCA.

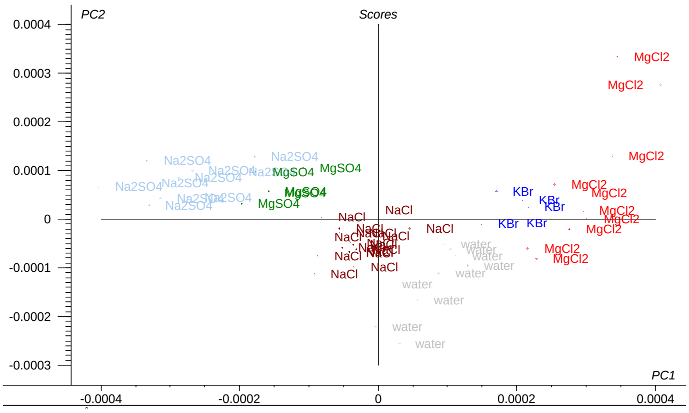


Fig. 5.10. PCA score plot of PC1 vs. PC2. The PCA was based on pretreated NIR spectra of gluten hydrated in 1M-salt solutions or water. Three NIR regions were analysed together.
PC1 and PC2 explains 32 and 8 % of the variance, respectively.
The plot shows mostly the qualitative differences, as the pretreatments have eliminated the physical- and also the gluten concentration variations from the spectra.

The score plot in Fig. 5.10. shows that the NIR spectra gives rise to the same order of samples as obtained from the amide I spectra. However, the NaCl-, water-, KBr- and $MgCl_2$-samples are now more clearly separated (compare to Fig. 5.7). In a similar analysis of the preprocessed salt *solution* spectra, the obtained pattern of samples was different from that in Fig. 5.10, indicating that water structure variations do not affect the results of the gluten analysis (see Appendix V, Fig. V-E). However, the order of gluten samples in Fig. 5.10 is much related to the water contents, as a similar score plot was obtained from a PLSR, in which one PLS component was found to explain the 86 % of the water content based on 28% of the spectral variance. (See Appendix V, Fig. V-F). The concurrent structural and concentration changes causes a difficulty in pointing out the spectral variations that result only from the structure changes.

Effect of salt concentration: The gluten replicate spectra were averaged, and the mean spectra were used in a PCA that included all salt concentrations (0.1, 0.2, 0.5 and 1.0 M). The resulting score plot (Fig. 5.11) shows that the concentration-dependent but specific salt effect is explained in PC1, whereas an effect of the salt concentration in general is explained in PC2. The sample pattern does not show opposite effects of high and low salt concentrations, as could have been expected for some salts according to the theory.

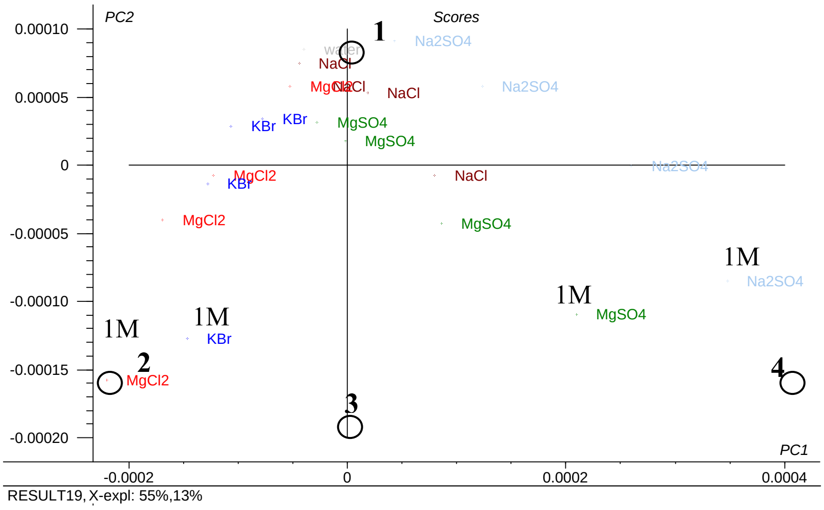


Fig. 5.11. PCA-score plot of PC1 vs. PC2. The PCA was based on pretreated NIR spectra of gluten hydrated in salt solutions (0.1-1.0 M) or water. Replicates were averaged.
PC1 and PC2 explains 55 and 13 %, respectively.
The circles represent artificial spectra that are compared in Fig. 5.12.

The four spectra, indicated with circles in the score plot, were calculated from loading and scores and are shown in Fig. 5.12.

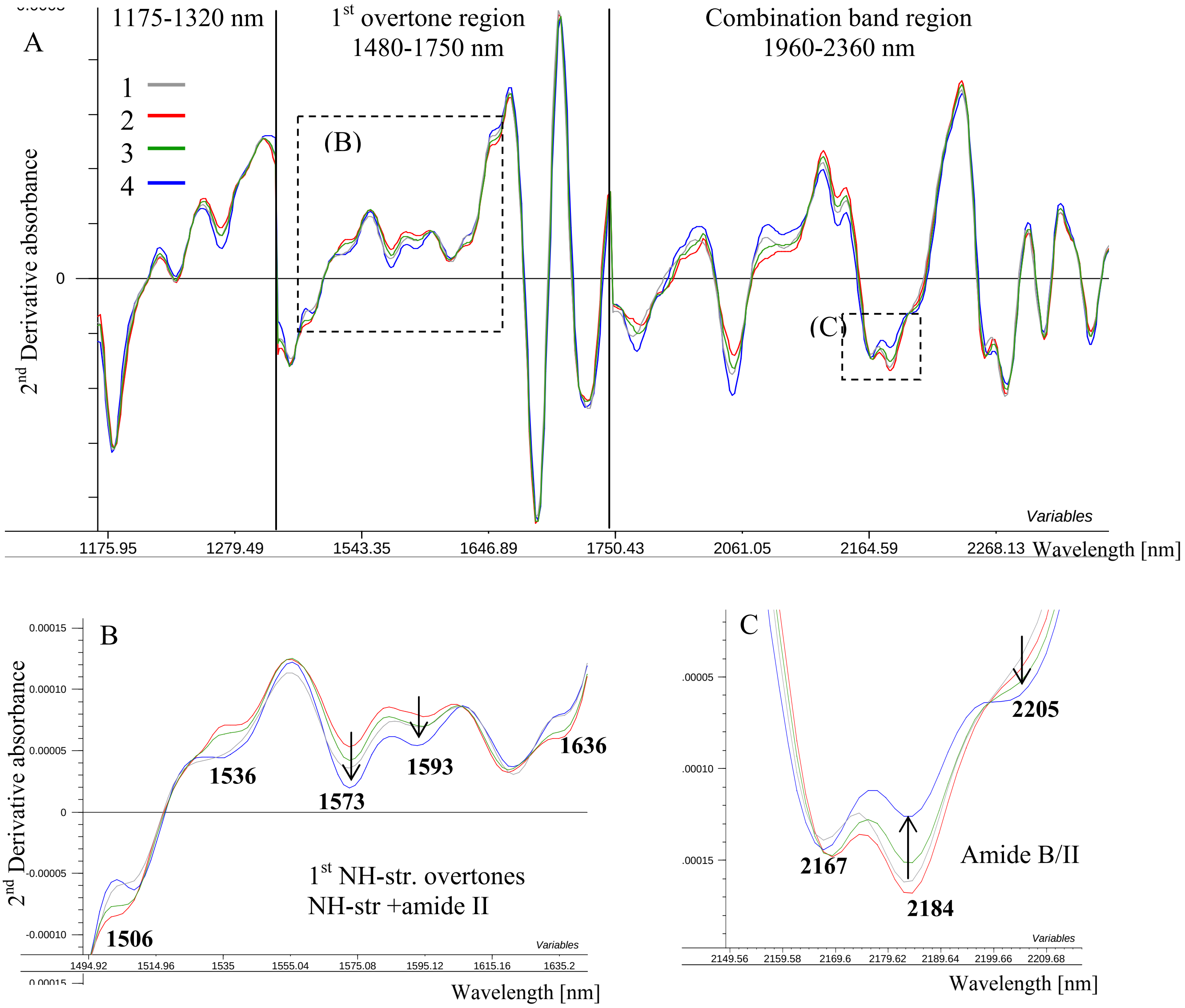


Fig . 5.12. 2nd derivative NIR spectra constructed from PC1- and PC2- scores and loadings (peaks are downward pointing). The PCA score plot is shown in Fig.5.11, and the sample numbers refer to the numbers in this plot. Grey=water, red=$MgCl_2$, green=neutral salt, blue= Na_2SO_4. A) The three analysed regions. B) 1st N-H str. overtones +NH combinations. C) Amide B/II combination band region.

Small effects of Na_2SO_4 at high concentration are seen in Fig. 5.12. In the 1st NH-str. overtone region (1480-1750 nm), Na_2SO_4 causes an increase at 1573 and 1593 nm compared to water, whereas $MgCl_2$ has the opposite effect and also causes at decrease 1536 nm and increases at 1506 and 1636 nm (Fig. 5.12B). These changes could be related to different effects on the hydrogen bonding state of the NH-groups in the peptide backbone and side chains.

In the amide combination band region (1960-2360 nm), Na_2SO_4 causes a decrease in the peak at 2184 nm assigned to amide B/II from primary amide groups (Fig. 5.12C). Comparing to the hydration effect seen in experiment IV (Fig. 4.12), this change is analogous to that caused by drying out of gluten and could perhaps be related to a decrease in the amount of Gln side chains participating in water interactions ($MgCl_2$ has a small but opposite effect on the 2184 nm peak). As there is an increase at 2205 nm, more β-sheet is indicated and is in agreement with the increase at 1633 cm^{-1} in the amide I band. Also at 2260-2280 nm, the effect of Na_2SO_4 is similar to a drying out, and there is an increase at 2268 nm and a decrease at 2278 nm, which could be interpreted as more random and less α-helix structure (in agreement with the amide I analysis). On the other hand, the increase in the amide A/II band at 2056 nm by Na_2SO_4 is in opposition to the drying out effect. Some other effects of the salts are seen, including shifts of some bands.

Results: Combined NIR and amide I analysis

The preprocessed and averaged NIR spectra were combined with the averaged amide I spectra in an analysis, with the purpose of pointing out interesting NIR regions and assist the interpretation of both spectral regions. A PLSR2 model was carried out with X= 2nd derivative NIR spectra and Y= 2nd derivative amide I spectra. Samples of all salt concentrations were included. All variables were standardised and full CV was used for validation. Jack-knifing was employed for removal of non-significant NIR-wavelength. PC1 and PC2 explains together 81.7 % of the NIR variance and 38.2 % of the amide I variance. The correlation loading plot in Fig. 5.13A. was used for identification of the correlations between the NIR and amide variables, and the results are shown in Table 5.4.

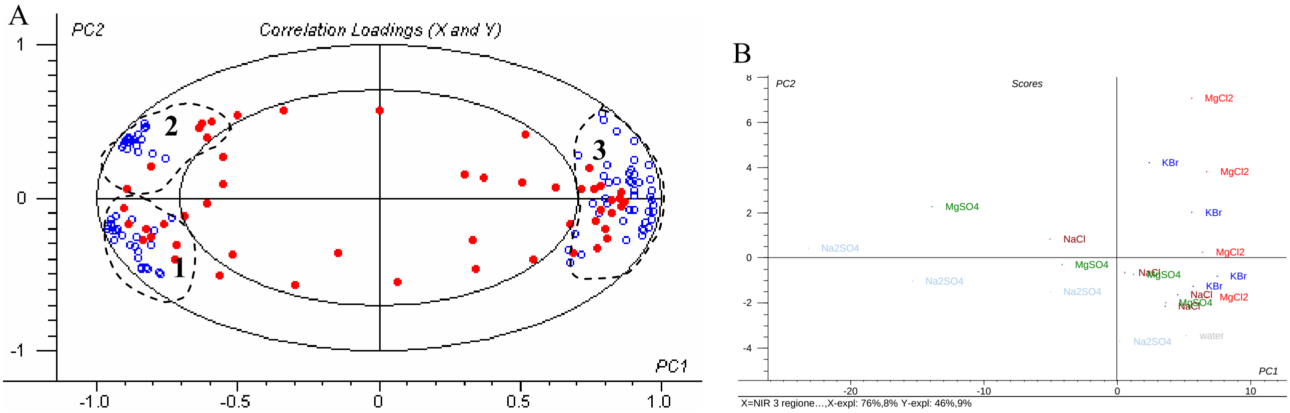


Fig. 5.13. PLSR results from an analysis of gluten hydrated in different salt solutions: X=inverted 2nd derivative NIR spectra (1175-1320, 1480-1750, and 1960-2360 nm), Y=inverted 2nd derivative amide I spectra. A) Correlation loading plot. Red=amide I variables. Blue= NIR variables. B) Score plot of PC1 vs. PC2. PC1 explains 72.7 % of the NIR-variance (X) and 38.5% of the amide I-variance (Y). PC2 explains 9 % of the NIR, but has no explanation of the amide I variance. The three encircled regions (1-3) show the correlations of NIR and amide I variables, on which the results in Table 5.4 are based.

The previous assignments of the NIR regions to secondary structures may assist in the interpretation of the results in Table 5.4. However, only the previous correlation of the 2208-2224 nm region to β-sheet is in agreement with the present correlation of β-sheet frequencies to this NIR-region. In Table 5.4. the increase at 2179-2189 nm is correlated to the increase of β-turns, α-helix and hydrated extended structures and could result from increased water interaction with Gln side chains, as explained. The correlation of the 2251-2256 nm region to the same structures is not in accordance with the attribution of this wavelength region to intermolecular β-sheet (as was suggested in chapter 3).

Amide I wavenumbers (cm^{-1})	Salts with positive influence	Amide I assignment	NIR wavelengths (nm)
1) 1695-1687, 1647-1639 2) 1631-1625	Na_2SO_4, $MgSO_4$	Intra β-sheet random Inter β-sheet	1277-1287, 1571-1580, 1596-1600, 1740-1747, 2057-2061. 1695-1707, **2208-2224**
3) 1676-1654, 1616-1610	Water, $MgCl_2$, KBr	β-turns+ α-helix+ hydrated extended structures	1668-1688, 1712-1723, 2030-2037, 2086-2101, **2179-2189**, 2251-2256.

Table 5.4. Correlations of NIR wavelengths to amide I wavenumbers, based on the PLSR model in Fig. 5.13.

The interesting regions pointed out in Table. 5.4 also include wavelengths from the 1st overtone region and an even lower wavelength region (1277-1287nm). In the lower wavelength regions, the low absorptivities offer a high penetration depth, wherefore these spectral regions contain information of the deeper layers of the dough/gluten samples compared to higher wavelength regions. These regions may therefore be more useful for dough monitoring compared to the combination band region (1960-2360 nm). However, the previous assignments of secondary structures to regions in the amide combination band region offered in this studies some advantages of using this high-wavelength region.

Discussion and conclusion

From their correlation to the amide I band variations, some variations in the NIR spectra were found correlated with the salt-induced secondary structure changes. However, the connection between protein secondary structure changes and the NIR variations was hard to ascertain from this experiment, as other phenomena (e.g. hydration) likely contributed to the NIR variations. Thus, some spectral changes could be explained from the secondary structure changes, as these interpretations agreed with the previous assignments in the NIR region, whereas other spectral changes could be interpreted as different interaction states of amino acid side chains e.g. Gln, participating in either intra/intermolecular β-sheets or in interactions with water. The confirmation of these assignments would require different types of reference analyses. In addition, the reference method for analysis of protein secondary structures could be improved by performing a quantitative amide I analysis by means of deconvolution and curve fitting. However, this analysis would require a lower noise level than in the present measurements, and a more optimal ATR-FTIR setup would be demanded for this purpose.

Other types of reference measurements, detecting various gluten properties, could have shed more light on the correlation between NIR spectral changes and gluten functionality. In an attempt to obtain a quick measure of the gluten functionality, the detection of gluten foam formation and collapse (according to the method of Bombara et al (2004)) was tried. However, the method was originally developed for enzyme-treated gluten and resulted in huge replicate variations for the untreated gluten. Also a few rheometer measurements were done in this study, but the results were not considered further, due to a shortness of time. However, this method could be applied in future analyses. Other approaches to the characterisation of the gluten functionality could involve a baking test or the study of the gluten microstructure e.g. in an environmental scanning electron microscope.

The possibility to use the NIR water bands for monitoring the dough development has been shown by other researchers and results from the sensitivity of these bands to the hydration of starch and proteins during dough mixing [Wesley, 1998, 2001; Alava, 2001]. This type of analysis, focusing on the water signal, could be relevant for studies of other macromolecule-water systems. However, much more information on the molecular level seems contained in the NIR spectra.

In conclusion, the study has shown that protein conformation/interaction changes in the gluten network may be detected by use of NIR with the current instrumentation. In addition, the interpretation of the spectral changes was somewhat possible by help of the complementary FTIR measurements. However, the generalization of the results and the validity in other protein systems needs to be confirmed.

5.3. References

Arakawa, T., Timasheff, S.N. (1984). Mechanism of protein salting in and salting out by divalent-cation salts- balance between hydration and salt binding. *Biochemistry-US*. 23, 5912-5923.

Baldwin, R.L. (1996). How Hofmeister ion interactions affect protein stability. *Biophys. J*. 71, 2056-2063.

Bombara, N., Pilosof, A.M.R., Anon, M.C. (1994). Mathematical model for formation rate and collapse of foams from enzyme-modified wheat flours. *J. Food. Sci*. 59, 626-630.

Boström, V., Kunz, W., Ninham, B.W. (2005a). Hofmeister effects in surface tension of aqueous electrolyte solutions. *Langmuir*. 21, 2619-2623.

Boström, M., Ninham, B.W. (2005b). Energy of an ion crossing a low dielectric membrane: the role of dispersion self-free energy. *Biophys. Chem*. 114, 95-101.

Boström, M., Williams, D.R.M., Ninham, B.W. (2004). Why the properties of proteins in salt solutions follow a Hofmeister series. *Curr. Opin. Colloid In*. 9, 48-52.

Butow, B.J., Gras, P.W., Haraszi, R., Bekes, F. (2002). Effects of different salts on mixing and extension parameters on a diverse group of wheat cultivars using 2-g mixograph and extensigraph methods. *Cereal. Chem*. 79, 826-833.

Di Stasio, E. (2004). Ionic regulation of proteins. *Ital. J. Biochem*. 53, 112-119.

Ebel, C., Faou, P., Kernel, B., Zaccai, G. (1999). Relative role of anions and cations in the stabilization of halophilic malate dehydrogenase. *Biochemistry*. 38, 9039-9047.

Eggers, D.K., Valentine, J.S. (2001). Crowding and hydration effects on protein conformation: A study with sol-gel encapsulated proteins. *J. Mol. Biol*. 314, 911-922.

Fu, F.N., Deoliveira, D.B., Trumble, W.R., Sarkar, H.K., Singh, B.R.(1994). Secondary structure estimation of proteins using the amide III region of fourier-transform infrared-spectroscopy – application to analyze calcium binding-induced structural-changes in calsequestrin. *Appl. Spectrosc*. 48, 1432-1441.

Grover, P.K., Ryall, R.L. (2005). Critical appraisal of salting-out and its implication for chemical and biological sciences. *Chem. Rev*. 105, 1-10.

He, H., Roach, R.R., Hoseney, R.C. (1992). Effect of nonchaotropic salts on flour bread-making properties. *Cereal. Chem.* 69, 366-371.

Hribar, B., Southall, N.T., Vlachy, V., Dill, K.A. (2002). How ions affect the structure of water. *J. Am. Chem. Soc.* 124, 12302-12311.

Kalra, A., Tugcu, N., Cramer, S.M., Garde, S. (2001). Salting-in and salting-out of hydrophobic solutes in aqueous solutions. *J. Phys. Chem. B*. 105, 6380-6386.

Kinsella, J.E., Hale, M.E. (1984). Hydrophobic associations and gluten consistency: Effects of specific anions. *J. Agr. Food Chem*. 32, 1054-1056.

Kunz, W., Nostro, P.L., Ninham, B.W. (2004). The present state of affairs with Hofmeister effects. *Curr. Opin. Colloid In.* 9, 1-18.

Preston, K.R. (1989). Effects of neutral salts of the lyotropic series on the physical dough properties of a canadian red spring wheat-flour. *Cereal Chem*. 66, 144-148.

Wellner, N., Bianchini, D., Mills, E.N.C., Belton, P.S. (2003). Effect of selected Hofmeister anions on the secondary structure and dynamics of wheat prolamins in gluten. *Cereal Chem.* 80, 596-600.

Chapter 6: Discussion, perspectives and future work

The aim of this thesis was to investigate the potential of NIR as a tool for analysis of macromolecule conformations and interactions, mainly in food samples.

In the present studies, proteins have been in focus. The sensitivity of NIR to protein structures and protein-protein interactions was investigated by performing experiments that involved pure protein solutions and more complex protein-based foodstuff (gluten). The sensitivity of NIR to the secondary structure of freeze-dried proteins is previously established, whereas the present work concerns the aqueous, 'wet', and more complex samples, exhibiting weaker protein signals in the infrared. These samples are considered relevant for examining the performance of NIR in food analyses. One goal in the study was to obtain more knowledge of the protein signals in NIR, such as to improve the information that can be obtained from future NIR studies. For this purpose, FTIR was used as a reference analysis.

Improvement of the spectroscopic analyses

A good reference method is pertinent for correct interpretation of the NIR spectra. Thus, improvement of the FTIR analyses by use of various spectral preprocessing tools was attempted (section 2.5 and Appendix II). A preprocessing method for removal of atmospheric absorptions from ATR-FTIR spectra was shown to perform well. The elimination of possible water vapour signals is important in FTIR analyses of protein secondary structures (by use of the amide I band), when an uncontrolled amount of vapour has been present during the measurements. However, the method was not needed for the reference measurements in this study.

Temperature variations during measurements may be another source of irrelevant spectral variations in analyses of aqueous samples. In this work a preprocessing tool, based on EMSC, was found able to eliminate the temperature- and salt-dependent variations in the ATR-FTIR water spectrum, as well as the spectral variations due to physical sample properties. This was demonstrated for ATR-FTIR spectra of both aqueous samples and hydrated gluten samples. Since water is a major constituent in many biological and food samples, this preprocessing method is thought to be an important tool for improving the ATR-FTIR analyses of many samples. The method may also reduce data analysis problems associated with the confounding between e.g. the temperature- or salt-dependent water band variations and the protein band variations. This is important for the amide I analyses, in which the water spectrum variations may obstruct the analysis due to overlapping with the amide I band. Thus, the correction was relevant in Experiment V, as the added salts caused perturbations of both the water bands and the protein bands.

The same preprocessing method is relevant for correction of NIR spectra, and future work will consider the application of the method to the NIR range. For example, in the study of protein solutions in Experiment III, the EMSC preprocessing could possibly have allowed the analysis of the full wavelength range without disturbance from uncontrolled temperature variations.

The methods were made available on the internet for use by others.

NIR analyses of pure protein solutions

NIR and FTIR analyses of different proteins in aqueous solutions at 10 mg/ml (Experiment III) resulted in NIR-assignments that were in agreement with the α-helix and β-sheet assignments established previously for dry proteins. Thus, a sensitivity of NIR to secondary structures of proteins in dilute solutions was shown, although further studies are needed for determining the influence of β-turns, 3_{10}-helices, loops etc. on the NIR spectra as well. (This analysis would have required a larger sample set than in the present study). The analyses indicated inferior performance of NIR compared to FTIR for quantitative analysis of protein secondary structures, and a possible explanation for this could be that amino acid side chain absorptions contribute to many of the protein bands in NIR and therefore cause a high complexity of these spectra.

From protein denaturation studies, the absorptions from intermolecular β-sheet, which is an important type of protein-protein interaction in food, was not shown distinguishable from intramolecular β-sheet absorptions in the NIR region, while these two structures cause distinct absorption patterns in the amide I spectra. The influence of intermolecular interactions on the NIR spectrum was further investigated by analyses of FBP (Appendix III), which undergoes various polymerisation reactions and conformation changes upon buffer-exchange and ligand-binding. The spectral changes confirmed the sensitivity of NIR to protein conformations, while some discrepancies between the NIR and the complementary amide I analysis were attributed to the sensitivity of NIR to changes in the interaction states and microenvironments of the amino acid side chains upon dimerisation/polymerisation. An amino acid side chain combination band at 2260 nm was suggested particularly sensitive to the polymerisation state of FBP. However, this hypothesis needs to be confirmed. Future studies of other proteins that participate in polymerisation reactions could bring evidence to the role of the amino acid side chains in the intermolecular interactions and to the subsequent perturbations of their NIR absorptions. However, such a sensitivity does not necessarily make a band suitable as a marker for protein-protein interactions, as conformational changes may influence these bands as well.

The establishment of NIR as a tool for conformational analysis of proteins in aqueous solutions could lead to several new applications of the spectroscopic method e.g. in the pharmaceutical industry. Some future application of NIR could be: the detection of antibody-antigen binding, characterisation of recombinant proteins compared to the wild types, and detection of denatured and misfolded proteins. With isotopic substitution, it could be possible to obtain information of conformation changes of the individual proteins. The advantages of NIR compared to FTIR are the fast data acquisition, less problems with protein absorption to the cuvette and with water vapour interferences, and no need for an expensive ATR-cell. Furthermore, in the industrial applications, the possibility to measure non-invasively through a glass vial without opening it also offers a great advantage over other methods. The limitation of NIR is caused by the requirement for a relatively high protein concentration, which is not reasonable for some types of pharmaceutical products.

NIR analyses of complex food samples

After having established the sensitivity of NIR to the structure of proteins in pure solutions, NIR was applied to study a protein system with more constituents, namely the gluten complex, for which starch and lipid signals also contributed to the spectra. One purpose in the studies was to assess the performance of NIR in structure-functionality relation studies. Experiment IV and V, involving analyses of gluten at low and high water contents, respectively, revealed qualitative changes in the NIR protein spectra upon hydration, denaturation and influence of various salts. Therefore, the experiments indicated a sensitivity of NIR to protein conformation and interaction changes in the complex gluten system.

In Experiment V, the fully hydrated gluten was used as a model system for protein networks in foods, and different gluten functionalities were obtained by hydration of gluten powder in various salt solutions. Some secondary structure changes were shown by the reference method (ATR-FTIR), which however did not point out intermolecular β-sheet as an obvious marker of the gluten functionality, in contrast to what is reported in the literature. As this type of interaction is of importance to protein-networks in foods, it could, however, be relevant to analyse other food products that contain the intermolecular β-sheet. From the present experiment, it appears to be difficult to conclude upon the intermolecular β-sheet and its spectral fingerprints in the NIR region.

Not all of the NIR spectral changes in Experiment IV and V could be ascribed to protein secondary structure changes, as some changes could result from the concomitant changes in amino acid side chain hydrations and interactions. An enhanced explanation of the NIR spectral changes would require more information regarding the gluten protein hydrations and interactions, of which only the intermolecular β-sheet was identified by the applied reference method. For improved characterisation of the gluten system, e.g. NMR could be used to give more information on the side chain hydrations and interactions.

The possibility to obtain protein hydration information by use of NIR is of interest, as there is a link between gluten protein hydration and the intermolecular β-sheet content, causing the hydration to be a marker of the gluten development. Also, the water spectrum encloses information of macromolecule-water interactions and therefore, it could be of interest to explore the water spectrum more in future analyses (also for analysis of other food systems). For this purpose, the developed EMSC preprocessing method (for ATR-FTIR spectra, described in Paper III) was indicated to be a useful tool, since it extracts information as regards the hydrogen bonding state in water.

The analysis of other well-characterised model systems could provide additional information as regards the capacity of NIR in food protein analysis and in structure-functionality studies. For example, the protein network formed by the milk whey proteins in diary products could be of interest. Model systems of low protein content are also relevant, as many food systems are based on other biomacromolecules and contain only a minor protein fraction. This applies to the dough, in which the chemical processes also are influenced by interactions between different dough components and by the competition between starch and gluten for water. Thus, no conclusion of the performance of NIR in dough analyses can be drawn from the present study of the much more protein-rich gluten fraction. Future analyses should consider starch and lipid conformations and interactions, as these play important roles in many food systems.

The recognition of NIR as a tool for detection of macromolecule conformation, hydration and interactions in complex samples could lead to a wide range of applications for NIR in the food production, as these factors are central to many food properties (e.g. of dough, cheese and other diary products). The replacement of a laborious method for quality control with *on line* NIR measurements could be advantageous in the food industry. A great benefit of NIR is that, besides information on macromolecule conformations and interactions, NIR spectra may simultaneously provide information on the chemical composition as well as the physical sample properties.

The recent developments of fibre-optics and fast diode array instruments have offered the possibility for remote and very fast spectrum collection, which is necessary for application of NIR as an *on line* method. However, applications, marginally possible in the laboratory, may be unsuited for *on line* measurements, which are related with harsher conditions than in the laboratory. Also for *on line* applications, the sample surface needs to be representative for the sample, or the surface property should be related to the sample property. Therefore, this work did not provide information on the performance of NIR for *on line* measurements of macromolecule interactions in foods, and further work is needed to throw light on this.

By providing increased knowledge of the macromolecule interactions in foods, NIR could be a useful tool in the food research, for example in the development of new products and ingredients etc. The design of new macromolecules with the desired functionalities could be one perspective. However, the complexity of most NIR spectra (with overlapping protein, lipid and carbohydrate bands) means that there is a need for a structure sensitive reference method for establishment of a calibration/classification model in each application. On the other hand, NIR is often correlated directly to the functionality measures of e.g. texture, taste etc. by use of chemometrics, and improved interpretation of these models is provided by increasing the knowledge of the biomacromolecule fingerprints in the NIR region, as attempted in this work.

Finally, it is emphasised that additional experiments will be necessary in order to verify the general ability of NIR to detect and characterise macromolecular conformational changes in a food matrix.

Appendix I

Paper I

Multivariate approaches in plant science.

In collaboration with David M. Gottlieb, Jakob Schultz, Susanne Jacobsen and Ib Søndergaard

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Phytochemistry 65 (2004) 1531–1548

PHYTOCHEMISTRY

www.elsevier.com/locate/phytochem

Review

Multivariate approaches in plant science

David M. Gottlieb [a], Jakob Schultz [b], Susanne W. Bruun [b], Susanne Jacobsen [b], Ib Søndergaard [b,*]

[a] Plasma Product Division, Statens Serum Institut, Artillerivej 5, DK-2300 Copenhagen S, Denmark
[b] Biochemistry and Nutrition Group, BioCentrum-DTU, Technical University of Denmark, Søltofts Plads, building 224, DK-2800 Kgs. Lyngby, Denmark

Received 18 December 2003; received in revised form 1 April 2004
Available online 28 May 2004

Abstract

The objective of proteomics is to get an overview of the proteins expressed at a given point in time in a given tissue and to identify the connection to the biochemical status of that tissue. Therefore sample throughput and analysis time are important issues in proteomics. The concept of proteomics is to encircle the identity of proteins of interest. However, the overall relation between proteins must also be explained. Classical proteomics consist of separation and characterization, based on two-dimensional electrophoresis, trypsin digestion, mass spectrometry and database searching. Characterization includes labor intensive work in order to manage, handle and analyze data. The field of classical proteomics should therefore be extended to also include handling of large datasets in an objective way. The separation obtained by two-dimensional electrophoresis and mass spectrometry gives rise to huge amount of data. We present a multivariate approach to the handling of data in proteomics with the advantage that protein patterns can be spotted at an early stage and consequently the proteins selected for sequencing can be selected intelligently. These methods can also be applied to other data generating protein analysis methods like mass spectrometry and near infrared spectroscopy and examples of application to these techniques are also presented. Multivariate data analysis can unravel complicated data structures and may thereby relieve the characterization phase in classical proteomics. Traditionally statistical methods are not suitable for analysis of the huge amounts of data, where the number of variables exceed the number of objects. Multivariate data analysis, on the other hand, may uncover the hidden structures present in these data. This study takes its starting point in the field of classical proteomics and shows how multivariate data analysis can lead to faster ways of finding interesting proteins. Multivariate analysis has shown interesting results as a supplement to classical proteomics and added a new dimension to the field of proteomics.

Keywords: Multivariate data analysis; Chemometrics; Proteomics; Wheat; Gliadins; Gluten; Quality; Mass spectrometry; 2D-gel electrophoresis; NIR

Contents

[*] Corresponding author. Tel.: +45-45252733; fax: +45-45886307.
E-mail address: ibs@biocentrum.dtu.dk (I. Søndergaard).

0031-9422/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2004.04.008

1. Introduction

Two-dimensional gel electrophoresis (2DGE) and mass spectrometry (MS) used in combination constitute a strong analytical tool used in "classical" proteomics (Fig. 1), in which MS is used for identification of proteins. By using the two analytical techniques independently of each other, but coupled with multivariate analysis, we have added a new dimension to the field of proteomics. Multivariate analysis improves the data handling in proteomics, and thereby narrowing down proteins of interest much faster (Fig. 2). Our method should therefore be considered as a strong supplement to the "classical" proteomics.

1.1. Multivariate data analysis

One of the troublesome issues in proteomics is the handling of data with respect to characterization. The field of chemometrics mainly concerns multivariate analysis applied to data from chemistry (Martens and Martens, 2001). Chemometric studies deal with the

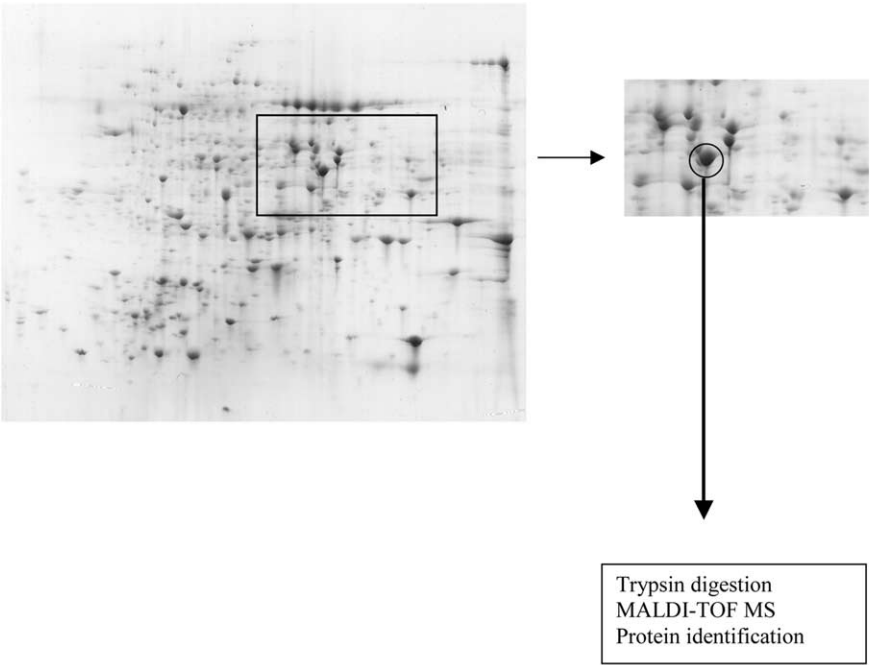


Fig. 1. The water-soluble fraction of a barley variety separated by 2DGE. The proteins are identified after trypsin digestion by mass spectrometry. Labour-intensive work is needed both when many gels are to be compared and when all the protein spots on one gel have to be sequenced and identified.

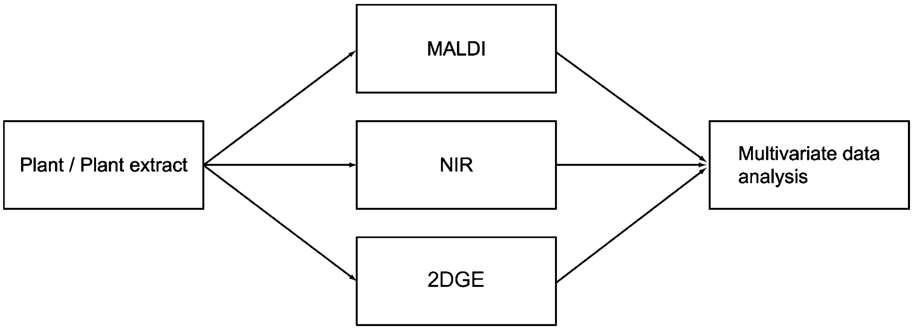


Fig. 2. Multivariate data analysis in plant science.

overall managing, handling and analysis of data collected from, e.g., 2DGE, MS or near infrared (NIR) spectroscopy.

Multivariate analysis builds on the application of statistical and mathematical methods, and includes the analysis of data with many observed variables, as well as the study of systems with many important types of variation (Martens and Martens, 2001).

The performance of multivariate analysis builds on two fundamental principles: (1) formulation of an experiment *before* data analysis (planning) and (2) problem reduction *during* and *after* data analysis (modeling). However, it must be emphasized that prior conventions, theories and expectations may have a restrictive influence on analysis, if used blindly. Multivariate analysis is therefore a balance between prior knowledge and new input gained during analysis.

The explorative data analysis is an important aspect during multivariate analysis. Before any hypotheses are arranged, explorative data analysis can give an insight in the multivariate chaos by means of scores (latent variables) and loading weights. An essential aspect in explorative data analysis is the outlier control. It can always be expected that data include errors as a consequence of typing errors, instrument errors, sampling errors etc. Hypothesis-generating analysis is a natural consequence of the entire concept behind multivariate analysis. In traditional statistical terms, a hypothesis is set up first and then experiments are carried out in order to demonstrate this hypothesis. This is known as deductive analysis. In contrast to traditional statistical methods, multivariate analysis is an inductive analysis, where hypotheses can be set up after having carried out the computational experiments.

Principal Component Analysis (PCA) is an unsupervised multivariate analysis technique used for transforming a set of observed variables into a new set of variables, which then are uncorrelated to one another (Everitt and Dunn, 1991). The basic idea is to find hidden structures in a dataset in order to describe these structures. The strength of PCA is the provision of low-dimensional plots of the data, e.g., to project many dimensions onto a few dimensions. On this basis it is possible to identify outlying observations, clusters of similar observations and other data structures. As the name indicates, the technique is based on principal components, a mathematical technique for an orthogonal orientation to principal axes. A principal component is also referred to as a latent variable. This variable cannot be measured directly but must be expressed as a linear combination of a set of input variables (Martens and Martens, 2001). The PCs form a rearranged multidimensional space based on a bilinear model of the data matrix **X**, meaning that **X** is decomposed into a structural part and an error part. The structural part consists of a scorematrix, **T**, and a transposed loadingmatrix, $\mathbf{P}^{\mathbf{T}}$, while the error part is termed **E** (Esbensen et al., 2000). Equation (1) is the mathematical skeleton of principal component model

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathrm{T}} + \mathbf{E}. \tag{1}$$

PCA is capable of transforming a large number of possible correlated variables to a smaller number of uncorrelated variables, PCs. The original axes are being replaced by PC-axes, where each PC-axis is a linear combination of the original variables.

The relationship of the PCs to the samples (the data rows, t_i) is called scores, and to the variables (data columns, p_i) called loadings. The new uncorrelated variables are represented in decreasing order of importance, which means that the first PC covers as much as possible of the variation in the dataset, and each subsequent component covers as much as possible of the remaining variation. The second PC is calculated orthogonal to the first PC, in that way to ensure complete non-correlation between the first and second PC-axis. The third PC-axis goes through the maximal variation described in the remaining dataset, i.e., not described in the first and second PC. This decomposition continues until all systematic variation is explained. When all variation is explained, the original data matrix has been reduced. It is thus possible to concentrate on only two or three dimensions at a time.

Although most variation is described in the first PC, it does not necessarily make the first PC the most interesting PC.

The **Y**-data structure is used to guide the decomposition of the **X**-matrix in order to reduce the original **X**-data to a small number of latent variables, termed Partial Least Squares (PLS) components (Esbensen et al., 2000). Calibration involves relating the two sets of data by regression modeling:

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{B}, \tag{2}$$

where **B** is a matrix containing b-regression vectors expressing the link between variation in the predictors and variation in the response. **X** and **Y** are defined in equations (3) and (4), respectively:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathbf{T}} + \mathbf{E}, \tag{3}$$

$$\mathbf{Y} = \mathbf{T} \cdot \mathbf{Q}^{\mathbf{T}} + \mathbf{F}, \tag{4}$$

where **T** are the scores, **P** and **Q** are the loadings and **E** and **F** are the residuals. **Y**-variables are predicted on the basis of a well-chosen set of relevant **X**-variables with explanatory or predictive purpose (Esbensen et al., 2000). The **Y**-matrix is therefore usually the property to be calibrated for (the response data), while the **X**-matrix (the descriptor data) is defined as the output of the instrument (Martens and Martens, 2001).

In PCA-calibration one set of loadings from the **X**-matrix (**P**-loadings, cf. equation (3)) is obtained, whereas PLS-calibration also includes the usage of loading weights (termed **W**-loadings). The **P**-loadings express the relationship between the raw data matrix **X** and its scores (**T**), whereas the **W**-loadings represent the effective loadings between **X** and **Y**. The differences between **P**- and **W**-loadings describe the influence of the **Y**-guidance on the decomposition of **X**. The loadings from the **Y**-matrix (**Q**-loadings, cf. equation (4)) are the regression coefficients from the **Y**-variables onto the scores (**U**). Together with the **Q**-loadings, the **W**-loadings are used to interpret the relationships between **X**- and **Y**-variables as well as interpreting the scores related to these loadings (Esbensen et al., 2000).

As an extension to PLS, Nørgaard et al. (2000) have developed iPLS. The purpose of iPLS is to divide the variables into subintervals of equal width in order to carry out local PLS on the subintervals, and thereby detect specific variables of interest. In this way one can get as large score vectors as possible in order to obtain more stable predictions (Höskuldsson, 2001). PLS is then carried out on each of the subintervals (local models) and the full-spectrum (global model). The procedure is a stepwise calculation with the aim of zooming into local models of interest, based on modeling performance between local models and the global model. The selection of intervals by iPLS is furthermore of importance in PCA in order to remove noise from the **X** data matrix.

1.2. Proteomics: how can data handling be improved?

The purpose of proteomics is to find ways of focusing on those proteins that are involved in a particular biological function of interest (Godovac-Zimmermann and Brown, 2001). The field of proteomics mainly consists of the following two stages (Rabilloud and Humphery-Smith, 2000): (1) separation of the proteins to be analyzed and (2) characterization of the separated proteins.

Examinations of proteomic maps have revealed more additional proteins than was expected when compared with the corresponding genomic maps (Corthals et al., 2000). It has therefore been suggested that the additional proteins found via proteome analysis are modified proteins, which could not be accounted for by genome analysis.

For every gene expressed in a cell at a given time, three times as many cellular proteins must be expected as a result of mRNA splicing and posttranslational modifications (Naaby-Hansen et al., 2001). Posttranslational modifications, which include simple proteolytic cleavage as well as covalent modification of specific amino acid residues, like, e.g., glycosylation, phosphorylation and acylation, are not detectable by analysis of RNA (Hille et al., 2001).

Although time-consuming, 2DGE is the favored separation technique in proteomics by virtue of the extremely high resolution obtained (Küster et al., 2001). While separation constitutes the first half of proteomics, characterization constitutes the other half. Proteins of interest, separated by 2DGE and electroblotted, may be submitted to N-terminal sequencing and succeeding database cataloguing in order to determine their identity. The advantage of N-terminal sequencing is the ability to directly sequence the N-terminus of electroblotted proteins without any need of specific preparation procedures (Kinter and Sherman, 2000). Characterization of proteins may also be carried out by application of MS. Matrix assisted laser desorption/ionisation time of flight (MALDI-TOF) MS is the most common type of MS combined with 2DGE (Hille et al., 2001). After 2DGE proteins are concentrated in individual spots as SDS–protein complexes within the polyacrylamide gel matrix. The protein spot must be pre-treated prior to analysis by mass spectrometry (Eckerskorn and Strupat, 2000). If the protein spot is excised from a dry gel, it must first be washed out and then cleaved by site-specific proteolysis with, e.g., trypsin (Naaby-Hansen et al., 2001). The resulting peptide fragments of a given protein spot can thereby be viewed as the third dimension separation, being independent of the two separation modes from 2DGE (Hanash, 2000). The peptide masses obtained from the mass spectrum are subjected to database cataloguing, where theoretically digestions of proteins are available, either in protein databases or in translated genomic databases (Naaby-Hansen et al., 2001).

A general problem in proteomics when starting with 2DGE is the characterization procedure. Approximately one day per gel is needed for analysis. When also including complete structural analysis by MS, about one month is required per gel (Hille et al., 2001). Identification of proteins by combination of MALDI TOF-MS and 2DGE, including the processes of image analysis, spot detection and enzymatic digestion prior to MS, is not possible to carry out automatically (Hille et al., 2001). A compromise between sample throughput and analysis time is therefore an important issue in proteomics of today. Since complete automation is not possible, the field of proteomics must be extended to include methods that can ease the evaluation of results obtained from 2DGE. Focus is turned to optimization of 2DGE and MS, and particularly on how to optimize the combination between these techniques.

Although the concept of proteomics is to encircle the identity of certain proteins of interest, the overall relation between proteins must also be explained. Since large amounts of data are collected, the overview may easily be lost. If the overview is lost, how can the conditions for proteomics then be fully obtained? The field of proteomics should therefore be extended to also include proper handling of large datasets. Image analysis of 2D gels is the basis for characterization of proteins. If the steps in image analysis could be speeded up, and at the same time be improved, it would ease the proteomic procedure essentially. Chemometrics may be the answer to a faster and more reliable analysis in 2DGE.

1.2.1. 2DGE

2DGE separates proteins according to two independent physical and chemical properties. Thousands of different proteins can thus be separated, and information such as the protein pI, the apparent molecular weight, and the amount of each protein is obtained. As the need for high throughput methods in proteomics increases, focus has shifted towards automation (Lopez, 2000; Patterson, 2000), by this new focus the bottleneck has moved from the protocol itself to the gel analysis (Lopez, 2000; Smilansky, 2001). The 2D protocol is still time consuming, however, it is important to notice that the subsequent gel analysis is just as time consuming. To automate gel analysis, several software programs have been developed (Appel et al., 1997; Lopez, 2000; Mahon and Dupree, 2001; Raman et al., 2002; Smilansky, 2001). Gel analysis involves three steps: (1) spot detection, (2) spot/gel alignment and (3) identification of interesting spots (Lopez, 2000). Development of efficient and reliable algorithms to perform the two first steps has been subject to much work (Gustafsson et al., 2002; Kaczmarek et al., 2002; Kriegel et al., 2000; Pleissner et al., 1999; Veeser et al., 2001), which can be seen in the new generation of 2DGE software (Raman et al., 2002). With the continued progress in development of 2DGE analytical software, the full potential of the 2DGE is anticipated in the near future.

The use of multivariate methods in the analysis of 2DGE is an emerging application (Appel et al., 1988; Jessen et al., 2002; Pun et al., 1988; Rabilloud et al., 1985; Tarroux, 1983; Tarroux et al., 1987; Vohradsky, 1997). By mathematical modeling of the data contained in 2D gels, it is possible to make fast extraction of data from gels. Traditionally the use of spot volume data has been applied as this makes direct use of the spot lists generated by most 2D analysis software packages. Alternatively it is possible to use the presence of a spot as indicator, so that the dataset is a binary matrix, where 1 shows that a spot is present in a gel and 0 that it is not (Radzikowski et al., 2002). This makes the classification of gel images based on expression patterns for protein spots possible. Moreover it is possible to deduce biological information from the loading plots, i.e., which spots contribute to the differentiation of the gels. Using an image-analytical approach, it is possible to do much of the same work in an automated and fast process that does not involve the subjective assessments of an operator.

However, although the developments in the 2DGE protocol and instrumentation have greatly improved the reliability and reproducibility of 2DGE, much focus is still on whether the method will turn into what everybody hopes; a fast, reliable method for high throughput proteomic research.

1.2.2. Speed/Automation

The speed and degree of automation are two areas where the 2DGE protocol can be improved. It has been estimated that it can take as much as one month to fully analyze one gel (Hille et al., 2001) with the current degree of automation. Improvements in the degree of automation, however it is estimated, can bring this down to 3–7 days (Hille et al., 2001). In addition, the gel analysis involves subjective assessments by the analyzer, which can make the analysis operator dependent.

Automated units to perform the first and second dimension runs as well as visualization combined with the development of IPG have brought down the number of process steps and operator dependent variables (Görg et al., 2000). Prototypes of fully automated robots for spot identification, excision and analysis with MALDI TOF-MS have likewise been described (Harry et al., 2000; Nordhoff et al., 2001). It is therefore in the gel analysis that we find the bottleneck in large-scale proteomics today.

1.2.3. Problematic proteins

Generally, two groups of proteins have been a problem in 2DGE separations – very basic and/or insoluble membrane proteins. The problem with the basic proteins has been the lack of commercial products to create a pH gradient above 10 and that basic proteins have been difficult to focus in the first dimension because of reverse electro-endosmotic flow. With IPG it is possible to make pH gradients up to 12 and at the same time use a standard protocol (Görg et al., 2000). Narrow pH gradients up to pH 12 require changes to the protocol to minimize the transportation of water from the cathode to the anode. Gradients with pH 10–12 and pH 9–12 have successfully been applied (Görg et al., 2000).

Hydrophobic membrane proteins are very challenging to handle and it is estimated that only about 1% of the membrane proteins are separated in 2DGE with a standard protocol (Fey and Larsen, 2001). The development of sample solubilization has improved the possibility of solubilizing these proteins by use of zwitterionic- and organic-detergents. But it is still an area that needs to be developed (Anderson et al., 2000).

1.2.4. Visualization of proteins

Traditionally proteins are stained with Coomassie brilliant blue (CBB) or silver. CBB staining has a relatively low sensitivity but is compatible with Western blotting and subsequent protein sequencing. Using CBB the spot intensities correlate linearly with protein amount. Silver staining can detect as little as 0.1 ng protein and is thereby much more sensitive than CBB staining but has disadvantages: (1) Silver staining does have a lower reproducibility between replicates; (2) It does only stain quantitatively in a narrow range, which means that silver staining is not useful to study differences in protein expression between different stages; (3) Some proteins are only stained weakly or not at all (Görg et al., 2000).

As an alternative to silver staining, fluorescence staining can be used. Fluorescence staining is less labour intensive and has a detection limit of 1–4 ng protein (Steinberg et al., 2000), which is better than CBB and at the same level as some silver staining procedures. The staining is linearly in a wide interval. The protocol is simple and can therefore more easily be used in an automated system.

Several alternatives to the classic methods of visualizing have been reported. In (Bienvenut et al., 1999; Binz et al., 1999) a molecular scanner is described. In this system all proteins in the gel are digested followed by transfer to a polyvinylidene difluoride (PVDF) membrane with a matrix solution compatible with MALDI TOF-MS. The PVDF membrane is then scanned directly with the MALDI TOF-MS instrument with a resolution of 0.4 mm. The spectrum obtained in each point is automatically submitted to a protein sequence database for identification. In this way a complete map of identified proteins is created. However, the method is not quantitative and the scanning of the gels with the spectrometer is very time consuming. Thus it takes about 36 days to scan one gel measuring 16×16 cm^2, generating approximately 40 GB of data. However, the authors estimate that it will be possible to bring this down to a matter of hours.

The method eliminates some of the classic flaws in 2DGE such as matching of protein spots, sensitivity, identification and to some extent co-migration. As all proteins in principle are identified, the need for matching of protein spots is eliminated. The method is therefore an interesting alternative to the classic 2DGE analysis, in that several steps are combined to one automated process.

In (Walker et al., 2001) a method is described where the second dimension is replaced by MALDI TOF-MS, thus creating a so-called virtual gel. The first dimension is run in the traditional manner, but the IPG strip is prepared for MALDI TOF-MS. By scanning the IPG with the MALDI TOF-MS spectrometer the virtual gel is created where the second dimension is constructed by MS spectra. Thus the method also eliminates some of the classic 2DGE problems.

1.3. Near infrared spectroscopy

Near infrared (NIR) spectroscopy provides a method for rapid, non-destructive and accurate analysis of the composition of a sample. It allows discrimination of various organic compounds and can be used both to acquire qualitative and quantitative information. It not only supplies chemical information, but also information of whether the physical properties of a sample can be obtained.

NIR has been widely used in the field of agriculture, and one of its first applications was the determination of moisture in agricultural products (Pasquini, 2003). Now it is also used in various other fields such as food and medicine, and it is an increasingly accepted tool for academic research and industrial quality control in many areas ranging from chemistry to agriculture and from life science to environmental analysis (Foley et al., 1998; Siesler, 2002). A merit of NIR is the simultaneous determination of multiple constituents in a sample, which also allows for estimation of complex attributes such as the susceptibility of plants to insect attack. NIR is not used for very sensitive analysis since the detection limit in general is only about 0.1% (w/w) for most constituents (Iwamoto and Kawano, 1992).

Infrared (IR) is the part of the electromagnetic spectrum that covers the wavelength region from 0.7 to 200 μm. The region of IR, which is nearest to the region of visible light, is called the near infrared (NIR) region, and it includes the wavelength range from about 780 to 2500 nm. The mid infrared (MIR) spans the higher wavelength range from 2500 to 15,000 nm (Davies, 1993).

The electromagnetic radiation can interact with matter to give rise to an absorption spectrum. In vibrational spectroscopy, which employs the MIR and NIR regions, the absorption bands originate predominantly from radiation energy transferred to mechanical energy associated with the vibration of atoms.

In a molecule, atoms or groups of atoms participating in chemical bonds are displacing one in relation to the other in a frequency that is defined by the type of bond of vibration (Davies, 1993). Absorption of infrared radiation induces the transition between vibrational energy levels, and the frequency and amount of the absorbed radiation gives information about the types and number of bonds between atoms or functional groups in the molecules. Consequently, the absorption spectrum reflects the chemical composition of the material being analyzed, and gives information on the amount of protein, fat, starch or any other organic molecule in a sample. However, NIR is a secondary method requiring calibration against a reference method for the constituent, because of influence also from physical properties (Osborne et al., 1993).

While the MIR-region possesses the energy that is necessary to promote molecules from their lowest excited vibrational states, the NIR region is of higher energy, and the absorptions originate from overtones or combinations of the fundamental absorptions seen in the MIR region.

IR spectroscopy that uses the MIR-region has been a well-established tool for elucidation of structure, because the peaks are relatively distinct and can be attributed to the presence of certain functional groups (Siesler, 2002). In the NIR region, however, direct interpretation of the spectral absorbances is very difficult for complex mixtures because of broad overlapping absorption bands. NIR thus relies on multivariate methods to quantify the properties or constituents of interest.

One of the advantages of NIR over IR is that NIR requires a minimum of sample preparation and provides the possibility for analysis on, e.g., intact fruit and also opaque samples. When a beam of IR radiation containing different frequencies is directed on to a molecule, an absorption spectrum (plot of energy versus wavelength) is produced, because only the radiation of frequencies capable of supplying exactly the energies between allowed transitions is absorbed. Each kind of molecule has a characteristic spectrum depending on the number and types of bonds, since the transition energies are defined by the vibrational frequencies of the different bonds.

Spectra of polyatomic molecules show absorptions from the distinct chemical groups, which vibrate at their characteristic group vibrations. The characteristic vibrations are relatively constant in their frequencies from molecule to molecule, but some adjustment takes place due to influence from different molecular environments and molecular interactions (e.g., degree of hydration) which influence the force constant (Bokobza, 2002). It is therefore possible for example to differentiate C–H stretching stemming from, e.g., alkanes, methanol and ethanoic acid (Osborne et al., 1993).

For polyatomic molecules interbond coupling can occur between stretching and bending vibrations of the same functional group, meaning that their vibrational energies are dependent on each other. This complicates the spectrum, but also causes some distinct vibrations for complex molecules. Proteins, for example, show characteristic absorption bands in the IR and NIR due to the vibrational modes: C═O stretching coupled to N–H bending and C–N stretching (amide I), and N–H bending coupled to C–N stretching (amide II) (Osborne et al., 1993). In the NIR some combination bands involving these modes (and, e.g., N–H stretching) appear, and such a band has been found very useful for estimation of protein concentration. Many bands in the NIR spectrum of protein are sensitive to changes in secondary structure and degree of hydration, and therefore can be used, e.g., for monitoring the denaturation of a protein (Wu et al., 2000). This is the consequence of NH-bands being displaced by hydrogen bonding like any other X–H-band. Hydrogen bonding changes the force constant of the covalent X–H bond thus causing a small shift in the wavelength at which the absorption band appears. This sensitivity of NIR to hydrogen bonding is the reason why NIR also can be used for studying the state of water in foods. The O–H absorption band, however, becomes very broad due to the hydrogen bonding.

2. Results and discussion

2.1. Proteomics 'classic'

The traditional way of doing proteomics is outlined in Fig. 1. Gels are evaluated using image-processing software; interesting spots are pointed out and identified. The visual image of protein spots is invaluable in proteome analysis as far as characterization of single proteins is concerned, which is why 2DGE is the favoured separation technique in proteomics. However, the characterization of proteins from 2D gels often requires many 2D images being compared to each other. However, when adding just a few more gels to the analysis it is almost impossible to maintain an overall view of the data. Image Master® and other software programmes like, e.g., CAROL (Kriegel et al., 2000), Z3 (Smilansky, 2001), PDQuest, Melanie and Progenesis have been developed, making attempts on easing the 2D image analysis. However, the real breakthrough will only appear when a full-automated analysis of 2D images is possible. An unquestionable obstacle towards the full-automated analysis of 2D gels is the problem of gel alignment.

2.2. Explorative data analysis

In order to increase the effectiveness of proper spot selection, the data from 2DGE can be subjected to multivariate analysis in order to point out which combination of spots could be valuable to sequence. This way a lot of time and effort can be saved when only the proper spots are identified. Although the images from 2DGE are obvious subjects to multivariate data analysis by virtue of the many variables they create, there still are some obstacles to pass before it is practically possible. The 2D gel patterns are exposed to geometrical distortions, locally as well as globally, with decisive impact on the grade of reproducibility. In order to analyze 2D images properly by multivariate analysis, they must first of all be aligned.

There are two ways to go. Multivariate data analysis can be used on either the spot list produced by the image processing software after alignment or directly on the aligned images. The first procedure has been used in some studies (Jessen et al., 2002, Radzikowski et al., 2002). In the latter study concerning rye proteins it was shown that the results from the different analysis could then be combined and analyzed by PCA to give an improved characterization of the varieties. The PCA of the 2D spot data was able to group the spots according to the varieties in which they were present and this improved the evaluation of the 2D gels. The resulting data from PCA can also be used to create a dendrogram of the investigated varieties. The PCA of the 2D spot data in combination with the functional properties data showed a similar grouping of the varieties and that there was one spot that was close to the properties, bread volume and bread height. The PCA of the 2D spot data can be useful in any 2D electrophoretic analysis where the aim is to find protein spots that are characteristic for a given sample or find protein spots that are present in a selected group of the investigated samples. The PCA of 2D spot data reduces the time spent on analysis of the results obtained from image analysis of 2D gels, and it also makes it easier to analyze a large number of gels. Another advantage is that it is possible to combine results from many different experiments and analyze them together.

2.3. Spot detection of 2DGE gels

Here, we present an analysis of 2DGE patterns of the storage proteins from ten different wheat varieties by PCA and PLSR. An analysis of the volume spot lists showed that the selected wheat varieties were represented in two groups. To avoid the generation of spot volume lists, i.e., to avoid spot detection, we used a method in which the gels were analyzed as images to test if the gels could be differentiated. The latter approach gave the same classification of the ten varieties as the use of spot volume lists, although without the prior work of spot detection and spot matching which is both time consuming and subjective. For further screening purposes the use of this approach in the initial screening of a large number of gels is therefore a promising alternative to the usual spot detection and matching.

Multivariate analysis is implemented in recent versions of popular 2DGE analysis software packages. The implementation, however, is solely based on the subsequent analysis of spot list data. The present method is based on sampling of real-spot data as basis for the detection.

The algorithm: Based on a data matrix of unfolded spot images we have used a singular value decomposition (SVD) to build a PC Model and used this model to create virtual gels of probability to indicate where the spots are located. This approach is described in Fig. 3. A more detailed description of the algorithm is found in Appendix A.

The algorithm has been used to identify spots on 2DGE gels of wheat storage proteins (Schultz et al., 2004) for 2DGE procedure. In Fig. 4, the spots used to construct the Peak matrix are marked.

Spot identification was done on three different gels of the wheat varieties: Pentium, Hussar and Trintella. All gels were sub-images extracted from whole gels and the background has been subtracted and the intensities adjusted (Fig. 5).

The results from the identifications are shown in Fig. 6. It is clear that almost all spots have been identified. However, there is a tendency towards missing identification of the weakest spots as well as some symmetric noise around each spots. Moreover, it is seen that the algorithm identifies the gravity point in the spot. This is in accordance with what is expected as the spots were sampled from the centre of gravity. Moreover, it should be noted that the intensities of the identified spots reflect the degree to which a specified pixel fits the reference model and not the original spot intensity. On the gel of the variety Hussar an area is seen where the number of spots are difficult to identify on the original gel. In this area three spots have been identified, which are in accordance with the actual gel. The algorithm has also been tested with other spots as reference data, which differ in size and form (data not shown). These tests show that the shape of the spot has less influence on the performance of the algorithm than the size (Fig. 7).

We have here demonstrated an alternative approach to 2DGE spot detection as well as shown how a multivariate approach can be used for other purposes than analysis of spectroscopic spectres. We believe that the algorithm as presented here can contribute to further development of powerful 2DGE analytic software packages, further fuelling the widespread use of the 2DGE technology in modern proteomics research.

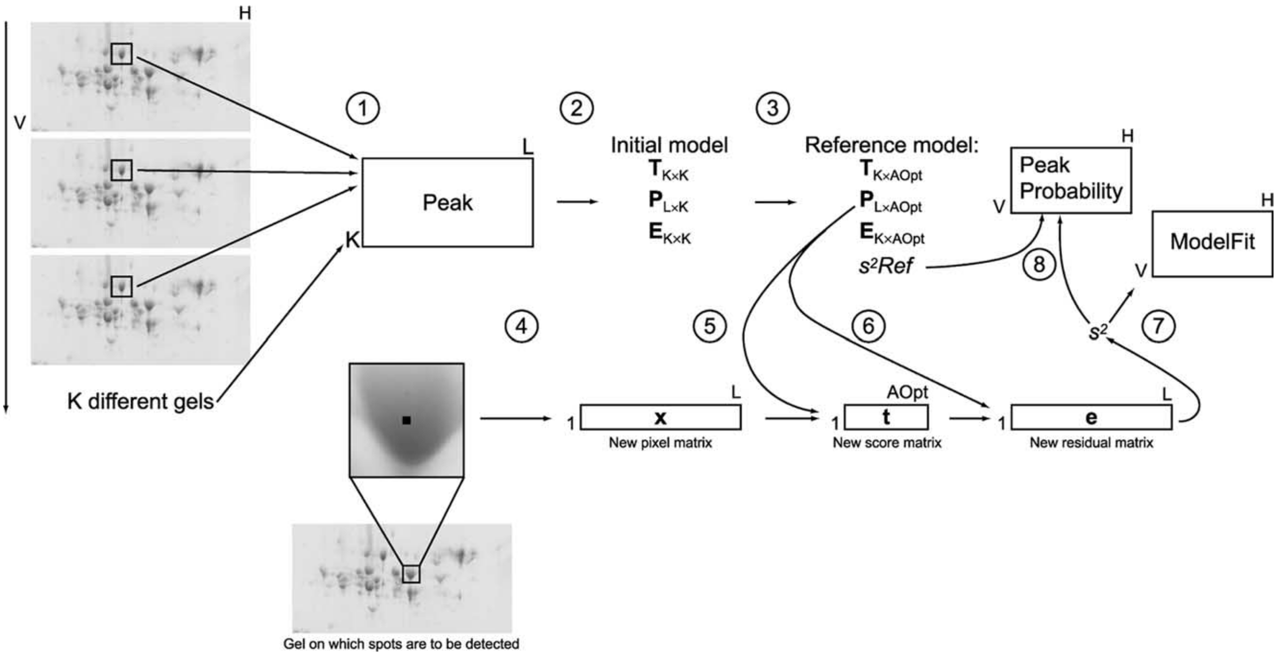


Fig. 3. Diagram of the FindPeak algorithm. (1) The spots are sampled, and the Peak matrix is constructed of the unfolded spot images. Spots are sampled from K different gels. One or more spots from each gel can be sampled. (2) The initial PC model is calculated, consisting of the scores matrix (**T**), the loadings matrix (**P**) and the residual matrix (**E**). (3) The optimal number of principal components is determined (AOpt) and the reference PC model is constructed based on this number of principal components. From the residual matrix the reference variance (s^2Ref) can be calculated. (4) In each pixel in the gel image to be analyzed, a sub-image with centre in the pixel is extracted and is unfolded to **x**. (5) The new score matrix, **t,** for the pixel is calculated from **x** and **P**. (6) The new residual matrix, **e**, for the pixel is calculated from the new score matrix **t** and **P**. (7) From the new residual matrix, the pixel variance, s^2 is calculated. The numeric value of this describes how well the pixel sub-image fits the model and forms the ModelFit matrix. (8) The relationship between s^2 and s^2Ref makes up the PeakProbability matrix.

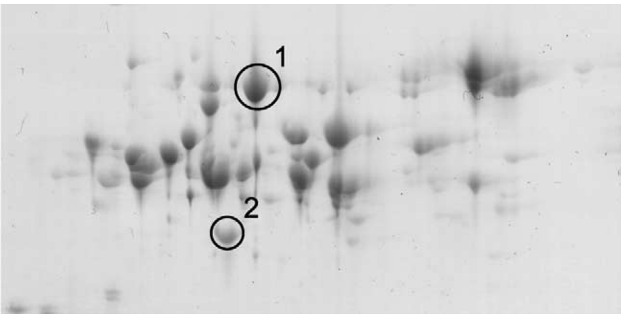


Fig. 4. DGE gel marked with the spots used to construct the Peak matrix. Spots 1 and 2 were sampled from 39 different gels of different wheat varieties.

2.4. Mass spectrometry and multivariate data analysis

MS of whole protein extracts together with chemometrics can be used to classify complex mixtures of proteins. We have used this chemometric approach as a supplement to the proteome analysis of the alcohol-soluble proteins (gliadins) from the wheat gluten complex (Gottlieb et al., 2002). Based on classic proteome analysis with 2DGE, a specific gliadin was found to only be present in wheat varieties unqualified for bread-making. By means of N-terminal sequencing, the identity of the protein was then encircled. Gliadin-data obtained from MALDI-TOF MS ranging 31 kDa were subject to multivariate analysis. By means of multivariate analysis on the MS data narrow molecular weight intervals of interest, with sizes of only few hundreds of Da, were repeatedly detected (Fig. 7). The study revealed that application of multivariate analysis could detect the molecular weight area in which the gliadin of interest was found by the classic labour-intensive proteome analysis. From the study it was concluded that the use of multivariate analysis on data output from separation of gliadins is a strong tool that can contribute substantially to the field of proteomics.

2.5. NIR spectroscopy and multivariate analysis

Quantitative information of a specific analyte is reflected in the intensity at the wavelength at which it absorbs according to Beers law, stating that absorbance presents a linear behavior with the concentration of the analyte for a fixed path length (Heise and Winzen, 2002). The NIR spectra are however often rather featureless, which prevents identification of bands for the analyte of interest. Use of a single wavelength will seldom provide a good model because of the occurrence of overlapping absorption bands and deviations from Beers law. Deviations from Beers law occur, e.g., at high

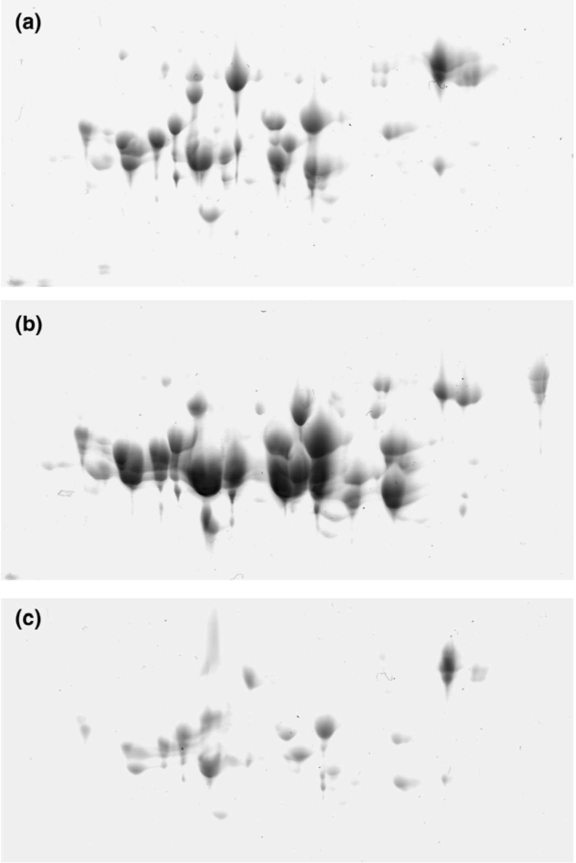


Fig. 5. Gels on which spots have been identified. (a) Pentium; (b) Hussar and (c) Trintella.

analyte concentrations, because of light scattering phenomena in solid samples, and when there are changes in the hydrogen bonding pattern such as temperature or relative concentration changes (Pasquini, 2003). In particulate samples light scattering phenomena further introduce nonlinearities (Martens et al., 2003).

Extraction of quantitative information therefore relies on multivariate models. Multiple linear regression (MLR) that uses only a few wavelengths is a usual method for regression of the reference data on the spectral data. However, PCR and PLSR that can separate out the relevant and reliable covariation patterns from the background noise in the full wavelength range are increasingly being used (Heise and Winzen, 2002). PCA provides a quick overview of the spectral data and reveals clusters and trends which could otherwise be hard to see (Fig. 8).

2.5.1. Barley mutants

NIR spectroscopy has shown promising results for application in plant biotechnology for gaining insight into the phenotypes that result from perturbation of the gene expression by genetic and environmental changes (Jacobsen et al., 2004; Munck et al., 2001). Processing of the NIR spectra by classification techniques yields a metabolic fingerprint of the organisms without differentiation of the individual metabolites. In this way NIR has potential as a fast screening method revealing organisms with altered phenome, but it is also possible to obtain quantitative information of specific metabolites

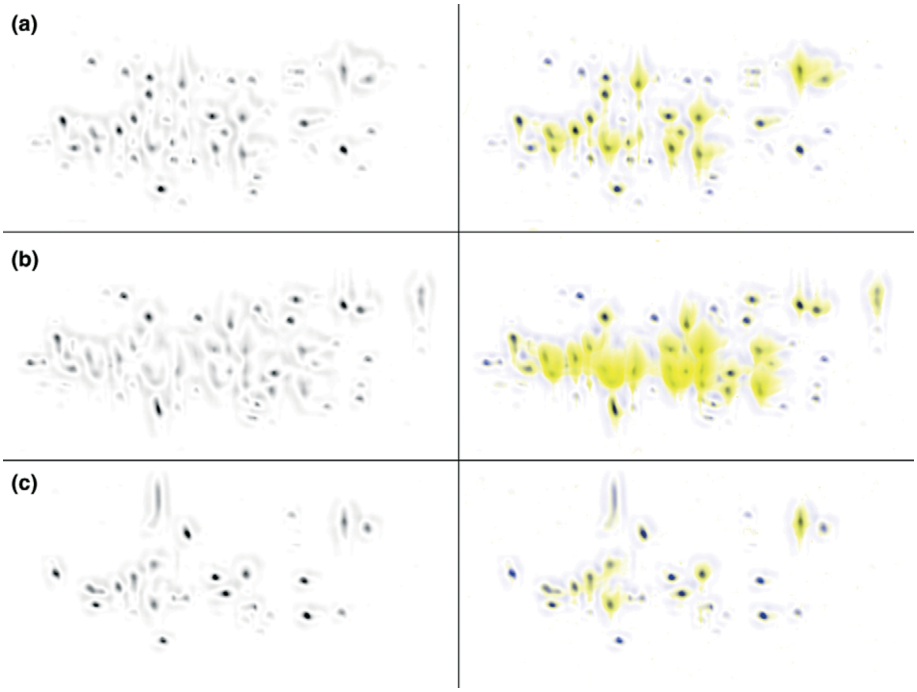


Fig. 6. Left column shows the PeakProbability matrix. Right column shows a composite image of the PeakProbability matrix and the original gel. (a) Pentium; (b) Hussar; (c) Trintella.

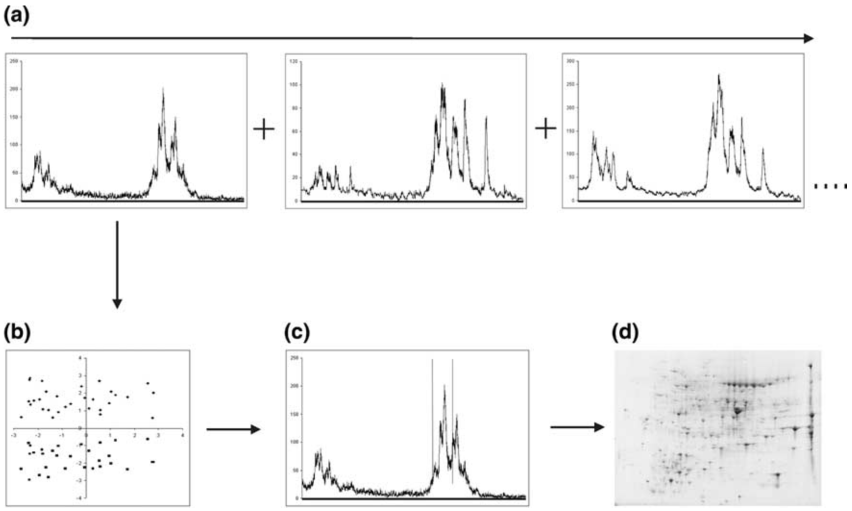


Fig. 7. Multivariate workflow combined with proteomics. (a) Mass spectra are collected, in this case of the alcohol-soluble fraction from wheat varieties. (b) By using PCA to analyse the collected spectra it is possible to compare the objects to each other in a score-plot (each spectrum is represented by a spot). (c) Variables of interest can be detected by iPLS. (d) The high-resolution obtained by 2DGE is then used to isolate the few proteins in the detected interval of interest. Further analysis is done by enzymatic digestion of the selected proteins and final identification by database searching.

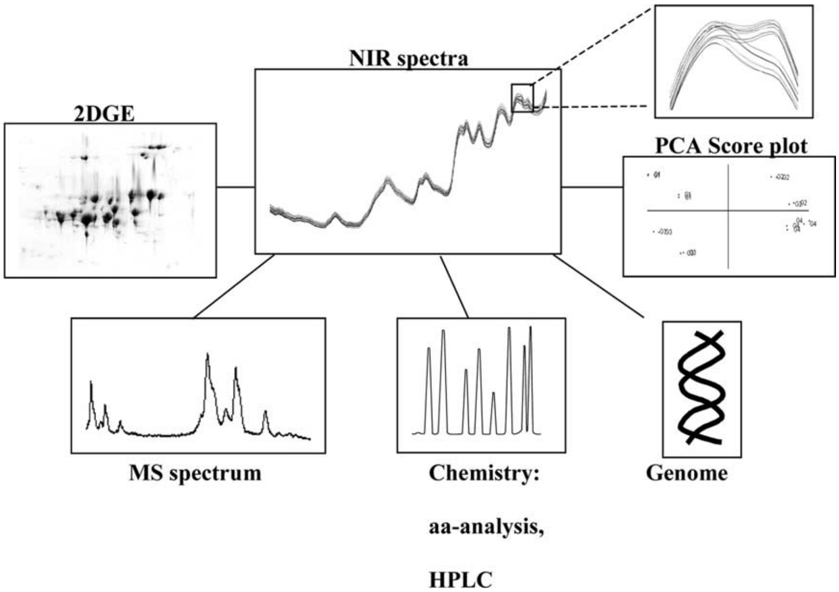


Fig. 8. NIR as an exploratory tool. NIR spectra contain 'hidden' information about the sample compositions, but groupings and trends in samples can easily be surveyed in a PCA based on the spectral data. The basis for the groupings is examined by use of various analytical methods, and gene sequencing reveals the underlying genome.

for use in metabolomics. Accurate and reproducible quantitative methods are necessary to differentiate samples where the result of the changed gene expressions is only quantitative changes in the metabolite concentrations (Sumner et al., 2003).

It has been shown by Munck et al., 2001 that NIR of barley flour provides a spectral fingerprint of the barley endosperm phenome, which can be used for discrimination of normal barley and high-lysine mutants. In the near isogenic background of the advanced barley lines, the effects of the high-lysine genes and also different growth environments were easily detected by NIR. A clear discrimination was seen in a PCA using the wavelength region 400–2500 nm. Also the ability of NIR to discriminate different high-lysine mutant genotypes has been demonstrated. In a work by Jacobsen et al., 2004, even the phenotypic effects of different alleles in the same locus were differentiated in a PCA, where a more extreme mutant was shown to form a distinct cluster.

Comparison of the mean spectra from the PCA clusters lead to identification of the spectroscopic signatures that discriminated the mutant genotypes. A small region (2280–2360 nm) in the spectra was identified as basis for visual discrimination of mutants and also their differentiation from normal barley. Observed absorption bands in this range were in the literature assigned to protein side chains (amino acid-determinant), cellulose and unsaturated fat. The effect of the different high-lysine mutant genotypes on the amino acid composition were thus reflected in the spectral shape, but also the effects of the altered proteome on other constituents such as starch, fat and fibre were evident in the spectra. These pleiotropic effects, which are often forgotten in analysis, are revealed by NIR, and the perspective is that mutants and transformants can be identified from their pleiotropic effects.

The ability of NIR to measure on the intact plant thus provides a holistic fingerprint of the metabolic status in contrast to other chemical methods applying to plant extracts and thus being biased towards specific chemicals. The advantage of using spectral information about the total endosperm composition for classification of unknown barley lines was demonstrated. A barley line that was formerly considered a waxy line due to its low amylose content was recognized as another mutant, since its spectra grouped together with *lys5* mutants in the PCA model. The *lys5* mutants were low in starch but compensated for that by high β-glucan content and thus differed from the classical waxy lines.

The use of the combination of NIR and PCA is a totally exploratory approach. After identification of clusters or outliers in a PCA, the proteome and metabolome can be further investigated by more selective methods like 2DGE, MS, amino acid analysis and other chemical analyses. In this respect, knowledge of the wavelengths at which the different constituents absorb can be of great help for targeting the chemical analysis. Genome analysis reveals the functional relationship between the genomes and the metabolomes. The fingerprinting approach allows for generating new hypotheses about the gene functions and is more objective than the traditional procedure in functional genomics, where only test of the logical response to a perturbation is made (Gidman et al., 2003).

2.5.2. Wheat quality

NIR has long been a recognized method for accurate prediction of the protein content of wheat for assessment of its breadmaking potential (Morris and Rose, 1996). The baking quality of flour, however, relates to both the amount and quality of the gluten proteins and is also determined by the complex interactions of all the biochemical constituents in flour (Veraverbeke and Delcour, 2002). Providing a measure of all the primary constituents simultaneously, NIR should have potential for determination of this quality. Various biochemical and physical properties of dough, relating to the baking quality, have been reasonably estimated by NIR, but a strong correlation between the measured property and the total protein content can lead to wrong conclusions. NIR is, however, sensitive not only to protein content but also to protein quality to some degree (Wesley et al., 2001). The quality of gluten protein is partially determined by the glutenin to gliadin ratio and the weight distribution of glutenins (Wesley et al., 2001). NIR is generally not very sensitive to individual levels of different proteins, but is has been found anyway that the individual contents of gliadin and glutenin can be estimated to some degree from NIR spectra not only because of their correlation to total protein content (Wesley et al., 2001).

2.5.3. Characterization of gluten

Dried gluten is used in the baking industry for improving the bread-making performance of wheat flour. To assure satisfactory performance of the gluten, assesment of both composition and functional end-use properties is required (Czuchajowska and Pomeranz, 1991). An important quality parameter of gluten is the moisture content, since high moisture content (above 10%) promotes deterioration of gluten quality. Other quality parameters include protein, free lipid and ash content as well as particle size and various rheological properties (Czuchajowska and Pomeranz, 1991). NIR has been found useful for determination of all these parameters, though composition was much better predicted than the physical and rheological properties (Czuchajowska and Pomeranz, 1991).

It is more desirable to know how well gluten performs in bread-making than just knowing its individual quality parameters. An accepted test for gluten functionality is measurement of the increase in volume of bread baked from flour fortified by gluten, and NIR has been tested for its ability to predict this end-use property. In the work by Czuchajowska and Pomeranz, 1991, it was found that a calibration model based on three wavelengths had limited power for predicting increase in loaf volume, but it was, however, found that some rheological properties of hydrated gluten (which correlates to its end use properties) could be well predicted by three-wavelengths MLR models.

2.5.4. Experiment: moisture in gluten

In an experiment FT-NIR reflectance spectra were measured on samples of freeze-dried gluten powder, which contained different amounts of moisture. In the beginning of the experiment spectra were measured on the dried gluten powders, and then samples were left to absorb moisture in a moist chamber at room temperature (26 °C). Spectra were again recorded after 2, 4 and 24 h on the same samples (Fig. 9). The final water content after 24 h was around 18% on a wet basis.

Broad water bands appear at around 1450 and 1930 nm for the moist samles and the intensity of the bands increases as the water content becomes higher. At the

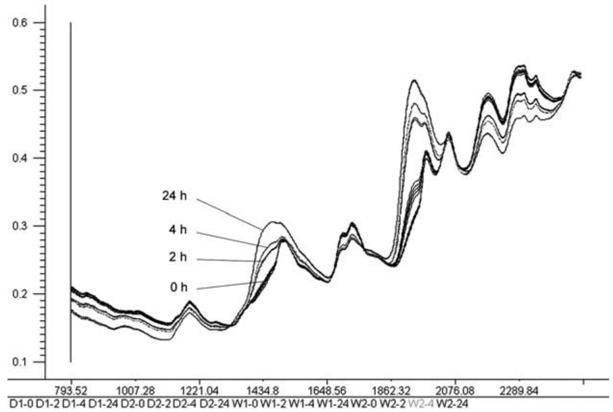


Fig. 9. NIR spectra measured on samples of gluten powder with different water contents.

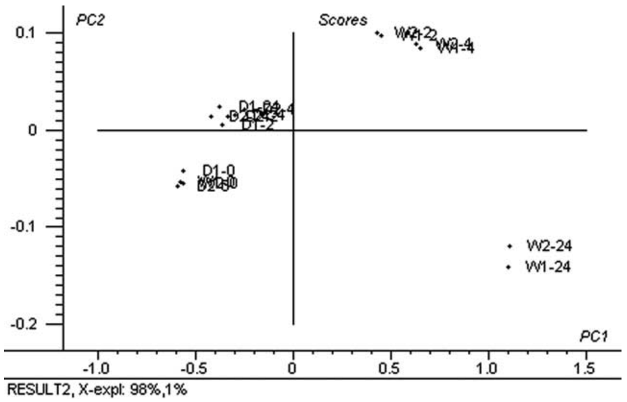


Fig. 10. Scoreplot from a PCA on the NIR spectra seen in Fig. 9. D, dry gluten samples. W, moist gluten samples. Last number refers to hours of hydration.

same time protein bands decrease in intensity. The spectra have been pretreated by multiple scatter correction (MSC) in order to remove undesired spectral variation caused by light scattering. Light scattering results from physical phenomena in the samples like particle shape, size, size distribution and sample packing and it introduces both multiplicative and additive effects in the spectra (Martens et al., 2003). To demonstrate how PCA can provide an overview of the variance and groupings in data a score plot from PCA on the MSC corrected full spectrum is shown in Fig. 10.

PC1 which describes 99% of the spectral variance reflects the increase in water content. PC2 shows another smaller phenomenon taking place.

To get more insight into the spectral changes that take place when water increases and interacts with the gluten proteins, the second derivative was taken. Second derivative spectra are shown in Fig. 11. Taking the second derivative of the spectra facilitates the visual inspection of the spectra since peaks in the original spectrum appear as more clearly separated downward peaks in the second derivative spectrum and at the same time multiplicative and additive effects are removed. The signal to noise ratio is, however, decreased.

The spectral changes upon hydration can originate from the changed concentrations, the changed hydration of the protein, changes in protein secondary structure upon hydration or from changes in protein side chains. Also changes in other minor components cannot be excluded.

A zoom picture of a small wavelength region with interesting spectral changes is seen in Fig. 12. It seems that two peaks that are close to each other change shape

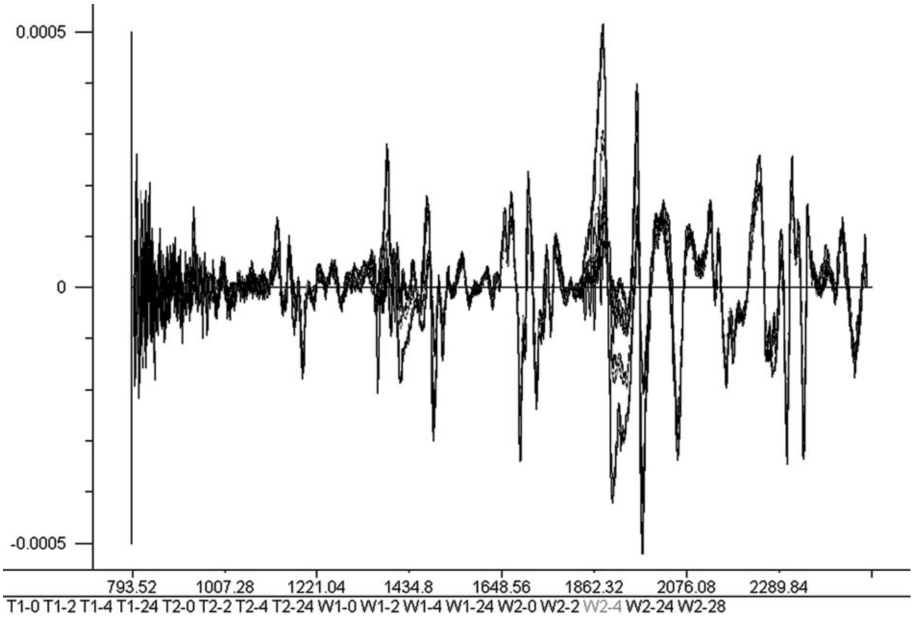


Fig. 11. Second derivative spectra. Original NIR spectra are seen in Fig. 9 and are measured on samples of gluten powder with different water contents.

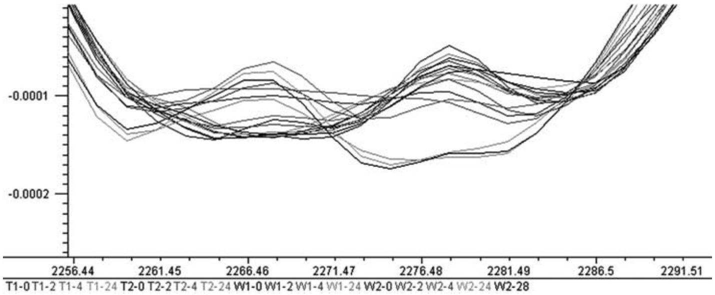


Fig. 12. Zoom picture of second derivative spectra in Fig. 11. Dry and moist gluten samples show distinct spectral features.

and shift wavelength positions to lower wavelengths when the samples become moist.

Changes in protein secondary structure are possibly reflected in the amid combination bands, which are sensitive to the degree of hydration, but also protein side chain vibrations can change upon secondary structure changes (Sefara et al., 1997). The interesting region encloses the wavelength range 2255–2290 nm, and in this area combinations of CH2 stretch and H–C–H bending vibrations from protein side chains and carbohydrates are found (starch content is low). It has been found that increased hydration of some gluten proteins leads to increased β-sheet content and decreased unordered structure, and at a certain point the β-turn content is increased (Belton et al., 1995). The analysis methods that have been used for elucidation of the changes in secondary structure that arise upon hydration, are yet needed for relating these structural events to the changes seen in the NIR spectra.

2.5.5. Applications to fruits and vegetables

The lower absorption intensities of water in the NIR compared to the MIR means that NIR is useful for measurements on high moisture samples such as fruit. A mode for measurements that is especially useful for measurements on intact fruits is called interaction, and it uses a fiber-optic probe, where the probability of the incident beam to interact with the sample is increased (Pasquini, 2003).

NIR is thus used for the assessment of the quality of fruits and vegetables. It has been successfully applied for determination of some of the most important quality aspects of fruits such as soluble solid, sugar and acidity content. Soluble solid content (SSC) or total solid content (TSC) (or dry matter content) has been determined for various fruits such as apples, melons, peaches, tomatoes, kiwis and dates and for vegetables such as onions, potatoes and corn by use of NIR (Kawano, 2000; Lammertyn et al., 2000; Schmilovitch et al., 1999; Slaughter et al., 2003).

NIR spectra have been found useful for determination of sucrose, glucose, fructose, citric acid, malic acid and ascorbic acid content in strawberries (Jin and Cui, 1994). In an experiment using reflectance measurements on potatoe slices, a calibration of sugar content, however, did not perform well, and NIR was also found insensitive to the fructose content (Scanlon et al., 1999). On the other hand, much better calibrations of sugar content have been reported when using transmission measurements on thin potato slices (Mehrubeoglu and Cote, 1997). NIR is in addition able to detect secondary metabolites in plants since, e.g., phenols, alkaloids, tannins and glucosinolate have distinct absorptions in the spectrum (Foley et al., 1998).

Also textural properties of fruits and vegetables can correlate to NIR spectra. Calibration models of NIR spectra of pears could predict fruit hardness, juiciness and mealiness. NIR can also predict the firmness, waxiness and mealiness of boiled potatoes as well as hardness and crispiness of boiled carrots (De Belie et al., 2003).

3. Conclusions

The field of classical proteomics should be extended to also include handling of large datasets by appropriate data analysis. The analysis performed by 2DGE, MS and NIR give rise to many data and multivariate data analysis can unravel the complicated data structures, which can relieve the characterization phase in classical proteomics. Based on analysis of proteins from the wheat gluten complex, we have used this technique to focus on the interesting spots or the interesting part of spectra before the actual identification phase. Multivariate analysis has shown interesting results as a supplement to classical proteomics and added a new dimension to the field of proteomics.

4. Experimental

For preparation of gluten powder, dough was made by mixing commercial wheat flour with distilled water (2:1) by hand. Gluten was washed out manually with distilled water from the dough and freeze-dried. The freeze-dried gluten was ground to powder and sieved through a 500-μm screen.

Moist samples were obtained by placing 2 × 700 mg gluten powder in a sealed container with water in the bottom, and FT-NIR spectra were collected at 2, 4, and 24 h after leaving the samples for moisture absorption. FT-NIR spectra were also obtained from two gluten samples, which were kept dry.

For the FT-NIR measurements, powders were compressed in a sample cup and spectra were recorded using a Spectrum One NTS, Perkin–Elmer spectrometer in reflectance mode. Spectral data were recorded from 793 to 2495 nm at 1.67 nm intervals with co-addition of 50 scans and use of a spectral resolution of 8 cm^{-1}. A Spectralon® diffuse reflectance standard was used as reference. NIR spectra were analyzed by using The Unscrambler Software version 8.0. Spectral preprocessing included MSC.

Acknowledgements

Harald Martens is thanked for very inspiring multivariate data analysis discussions. Ljiljana Nesic and Marianne K. Petersen are greatly appreciated for their crucial experimental assistance. Center for Advanced Food research is gratefully acknowledged. This work is part of the Danish Cereal Network supported by The Ministry of Food, Agriculture and Fisheries.

Appendix A

Here we describe the spot detection algorithm. Notation and syntax are that of Matlab.

The initial model is calculated using Singular Value Decomposition (SVD) as:

$$[\mathbf{U},\mathbf{S},\mathbf{V}] = \text{svd}(\mathbf{X}).$$

From this we get **T** and **P** as:

$$\mathbf{T} = \mathbf{VS},$$

$$\mathbf{P} = \mathbf{U}.$$

The optimal number of principal components to use in the model is obtained by studying the **S** matrix, which is a diagonal matrix of the eigenvalues. The eigenvalues describe how much of the variance is described by each principal component, and is commonly expressed in percents as:

$$\frac{\text{diag}(\mathbf{S})\cdot 100}{\text{sum}(\text{diag}(\mathbf{S}))}.$$

From this the optimal number of principal components (AOpt) is determined and the **T**, **P** and **E** matrixes are constructed and the reference variance is calculated (Fig. 3, step 2):

$$\mathbf{T} = \mathbf{T}(:,1:A\text{Opt}),$$

$$\mathbf{P} = \mathbf{P}(:,1:A\text{Opt}),$$

$$\mathbf{E} = \mathbf{Peak}' - \mathbf{T}\cdot\mathbf{P}',$$

$$s2\text{Ref} = \text{mean}(\mathbf{E}(:).\hat{}2).$$

In every pixel of the gel image to be analyzed is a sub-image the same size as the spot sampling sub-image is extracted and unfolded to row vector **x**, which is used as data in a new model that is calculated from the loadings matrix from the reference model. Hereby it is possible to estimate a new score matrix, by projecting **x** on **P** and hence the residual matrix and variance for the sub-image:

$$\mathbf{x} = \mathbf{t}\cdot\mathbf{P}' + \mathbf{e},$$

$$\mathbf{t} = \mathbf{x}\cdot\mathbf{P}\cdot(\mathbf{PP}')^{-1},$$

$$\mathbf{t} = \mathbf{x}\cdot\mathbf{P},$$

$$\mathbf{e} = \mathbf{x} - \mathbf{t}\cdot\mathbf{P}',$$

$$s^2 = \text{mean}(\mathbf{e}(:).\hat{}2).$$

The variance, $s2$ is used to calculate a peak-probability from the $s2$Ref (Fig. 3, step 7):

$$s2\text{Ratio} = s2/s2\text{Ref},$$

$$\text{Prob} = \text{ScalingFactor} \ * 1/s2\text{Ratio},$$

$$\text{Prob} = \max(\text{Prob},0),$$

$$\text{Prob} = \min(\text{Prob},1),$$

$$\text{PeakProbability} = \text{Prob} + f(\text{leverage}).$$

This is shown below:

```
– Make reference PCA model and calculate reference scores and loadings.
– Study the S matrix and determine the optimal number of principal components ($A$Opt) and estimate the reference
  variance ($s2$Ref) with this number of principal components.
  – for $h = 1 : nh$ (number of horisontal pixels in the gel-image to be analyzed)
    – for $v = 1$:nv (number of vertical pixels in the gel-image to be analyzed)
      – Extraxt a sub-image with centre in the pixel $(v - dv : v + dv, h - dh : h + dh)$
      – Reshape to a row vector, x
      – Estimate the score matrix for the new x data from the reference loadings.
      – Estimate the residual for the new x data: $\mathbf{e} = \mathbf{x} - \mathbf{t} \ * \mathbf{P}'$
      – Calculate the variance: $s2 = \text{mean}(\mathbf{e}(:)).\hat{}2)$
      – Calculate the probability factor for the pixel $(v, h)$
    – end
  – end
```

Before the reference model is made the data is normalised. This is done by Multiplicative Scatter Correction (MSC) (Martens et al., 2003).

References

Anderson, N.L., Matheson, A.D., Steiner, S., 2000. Proteomics: applications in basic and applied biology. Current Opinion in Biotechnology 11, 408–412.

Appel, R., Hochstrasser, D., Rock, C., Funk, M., Muller, A.F., Pellegrini, C., 1988. Automatic classification of two-dimensional gel-electrophoresis pictures by heuristic clustering analysis – a step toward machine learning. Electrophoresis 9, 136–142.

Appel, R.D., Palagi, P.M., Walther, D., Vargas, J.R., Sanchez, J.-C., Ravier, F., Pasquali, C., Hochstrasser, D.F., 1997. Melanie II – a third-generation software package for analysis of two-dimensional electrophoresis images: I. Features and user interface. Electrophoresis 18, 2724–2734.

Belton, P.S., Colquhoun, I.J., Grant, A., Wellner, N., Field, J.M., Shewry, P.R., Tatham, A.S., 1995. FTIR and NMR-studies on the hydration of a high-M(R) subunit of glutenin. International Journal of Biological Macromolecules 17, 74–80.

Bienvenut, W.V., Sanchez, J.-C., Karmine, A., Rouge, V., Rose, K., Binz, P.-A., Hochstrasser, D.F., 1999. Toward a clinical molecular scanner for proteome research: parallel protein chemical processing before and during western blot. Analytical Chemistry 71, 4800–4807.

Binz, P.A., Muller, M., Walther, D., Bienvenut, W.V., Gras, R., Hoogland, C., Bouchet, G., Gasteiger, E., Fabbretti, R., Gay, S., Palagi, P., Wilkins, M.R., Rouge, V., Tonella, L., Paesano, S., Rossellat, G., Karmine, A., Bairoch, A., Sanchez, J.-C., Appel, R.D., Hochstrasser, D.F., 1999. A molecular scanner to automate proteomic research and to display proteome images. Analytical Chemistry 71, 4981–4988.

Bokobza, L., 2002. Origin of near-infrared absorption bands. In: Siesler, H.W., Ozaki, Y., Kawata, S., Heise, H.M. (Eds.), Near-infrared Spectroscopy. Principles, Instruments, Applications. Wiley, Weinheim, pp. 11–41.

Corthals, G.L., Wasinger, V.C., Hochstrasser, D.F., Sanchez, J.-C., 2000. The dynamic range of protein expression: a challenge for proteomic research. Electrophoresis 21, 1104–1115.

Czuchajowska, Z., Pomeranz, Y., 1991. Evaluation of vital dry gluten composition and functionality in breakmaking by near-infrared reflectance spectoscopy. Cereal Foods World 36, 439–464.

Davies, A.M.C., 1993. Introduction to NIR spectroscopy. The Second European Symposium on Near Infrared (NIR) Spectroscopy, 1–16.

De Belie, N., Pedersen, D.K., Martens, M., Bro, R., Munck, L., De Baerdemaeker, J., 2003. The use of visible and near-infrared reflectance measurements to assess sensory changes in carrot texture and sweetness during heat treatment. Biosystems Engineering 85, 213–225.

Eckerskorn, C., Strupat, K., 2000. Mass spectrometry of intact proteins from two-dimensional PAGE. In: Rabilloud, T. (Ed.), Proteome Research: Two-dimensional Gel Electrophoresis and Identification Methods. Springer Verlag, Berlin, pp. 233–244.

Esbensen, K.H., Guyot, D., Westad, F., 2000. Multivariate Data Analysis – In Practice. Camo ASA, Oslo.

Everitt, B.S., Dunn, G., 1991. Applied Multivariate Data Analysis. Edward Arnold, London.

Fey, S.J., Larsen, P.M., 2001. 2D or not 2D. Current Opinion in Chemical Biology 5, 26–33.

Foley, W.J., McIlwee, A., Lawler, I., Aragones, L., Woolnough, A.P., Berding, N., 1998. Ecological applications of near infrared reflectance spectroscopy a tool for rapid, cost-effective prediction of the composition of plant and animal tissues and aspects of animal performance. Oecologia 116, 293–305.

Gidman, E., Goodacre, R., Emmett, B., Smith, A.R., Gwynn-Jones, D., 2003. Investigating plant–plant interference by metabolic fingerprinting. Phytochemistry 63, 705–710.

Godovac-Zimmermann, J., Brown, L.R., 2001. Perspectives for mass spectrometry and functional proteomics. Mass Spectrometry Reviews 20, 1–57.

Gottlieb, D.M., Schultz, J., Petersen, M., Nesic, L., Jacobsen, S., Søndergaard, I., 2002. Determination of wheat quality by mass spectrometry and multivariate data analysis. Rapid Communications In Mass Spectrometry 16, 2034–2039.

Gustafsson, J.S., Blomberg, A., Rudemo, M., 2002. Warping two-dimensional electrophoresis gel images to correct for geometric distortions of the spot pattern. Electrophoresis 23, 1731–1744.

Görg, A., Obermaler, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R., Weiss, W., 2000. The current state of two-dimensional electrophoresis with immobilized pH gradients. Electrophoresis 21, 1037–1053.

Hanash, S.M., 2000. Biomedical applications of two-dimensional electrophoresis using immobilized pH gradients: current status. Electrophoresis 21, 1202–1209.

Harry, J.L., Wilkins, M.R., Herbert, B.R., Packer, N.H., Gooley, A.A., Williams, K.L., 2000. Proteomics: capacity versus utility. Electrophoresis 21, 1071–1081.

Heise, H.M., Winzen, R., 2002. Chemometrics in near-infrared spectroscopy. Principles, instruments, applications. In: Siesler, H.W., Ozaki, Y., Kawata, S., Heise, H.M. (Eds.), Near-infrared Spectroscopy. Wiley, Weinheim, pp. 125–162.

Hille, J.M., Freed, A.L., Wätzig, H., 2001. Possibilities to improve automation, speed and precision of proteome analysis: a comparison of two-dimensional electrophoresis and alternatives. Electrophoresis 22, 4035–4052.

Humphery-Smith, I., 2000. Introduction: the virtue of proteomics. In: Rabilloud, T. (Ed.), Proteome Research: Two-dimensional Gel Electrophoresis and Identification Methods. Springer Verlag, Berlin, pp. 1–8.

Höskuldsson, A., 2001. Variable and subset selection in PLS regression. Chemometrics and Intelligent Laboratory Systems 55, 23–38.

Iwamoto, M., Kawano, S., 1992. Advantages and disadvantages of NIR applications for the food industry. In: Murray, I., Cowe, I.A. (Eds.), Making Light Work: Advances in Near Infrared Spectroscopy. Wiley, Cambridge, UK, pp. 367–375.

Jacobsen, S., Søndergaard, I., Møller, B., Desler, T., Munck, L., 2004. The barley endosperm as a data interface for the expression of genes and gene combinations at different levels of biological organization explored through pattern-recognition data evaluation. J. Cereal Science (submitted).

Jessen, F., Lametsch, R., Bendixen, E., Kjærsgård, I.V.H., Jørgensen, B.M., 2002. Extracting information from two-dimensional electrophoresis gels by partial least squares regression. Proteomics 2, 32–35.

Jin, T.M., Cui, H.C., 1994. A new method for determination of nutrient contents of intact strawberries – near infrared spectrometry. Acta Agriculture Boreal Singapore 9, 120–123.

Kaczmarek, K., Walczak, B., de Jong, S., Vandeginste, B.G.M., 2002. Feature based fuzzy matching of 2D gel electrophoresis images. Journal of Chemical Information and Computer Science 42, 1431–1442.

Kawano, S., 2000. Application to agricultural products and foodstuffs. In: Siesler, H.W., Ozaki, Y., Kawata, S., Heise, H.M. (Eds.), Near-infrared Spectroscopy. Principles, Instruments, Applications. Wiley, Weinheim, pp. 269–287.

Kinter, M., Sherman, N.E., 2000. Protein Sequencing and Identification Using Tandem Mass Spectrometry. Wiley, New York.

Kriegel, K., Seefeldt, I., Hoffmann, F., Schultz, C., Wenk, C., Regitz-Zagrosek, V., Oswald, H., Fleck, E., 2000. An alternative approach to deal with geometric uncertainties in computer analysis of two-dimensional electrophoresis gels. Electrophoresis 21, 2637–2640.

Küster, B., Krogh, T.N., Mørtz, E., Harvey, D.J., 2001. Glycosylation analysis of gel-separated proteins. Proteomics 1, 350–361.

Lammertyn, J., Peirs, A., De Baerdemaeker, J., Nicolai, B., 2000. Light penetration properties of NIR radiation in fruit with respect to non-destructive quality assessment. Postharvest Biology and Technology 18, 121–132.

Lopez, M.F., 2000. Better approaches to finding the needle in a haystack: optimizing proteome analysis through automation. Electrophoresis 21, 1082–1093.

Mahon, P., Dupree, P., 2001. Quantitative and reproducible two-dimensional gel analysis using Phoretix 2D Full. Electrophoresis 22, 2075–2085.

Martens, H., Martens, M., 2001. Multivariate Analysis of Quality – An Introduction. John Wiley & Sons Ltd, Chichester.

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, 394–404.

Mehrubeoglu, M., Cote, G.L., 1997. Determination of total reducing sugars in potato samples using near-infrared spectroscopy. Cereal Foods World 42, 409–413.

Morris, C.F., Rose, S.P., 1996. Wheat. In: Henry, R.J., Kettlewell, P.S. (Eds.), Cereal Grain Quality. Chapmann & Hall, London, pp. 5–54.

Munck, L., Nielsen, J.P., Møller, B., Jacobsen, S., Søndergaard, I., Engelsen, S.B., Nørgaard, L., Bro, R., 2001. Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics. Analytica Chimica Acta 446, 171–186.

Naaby-Hansen, S., Waterfield, M.D., Cramer, R., 2001. Proteomics – post-genomic cartography to understand gene function. Trends in Pharmacological Sciences 22, 376–384.

Nordhoff, E., Egelhofer, V., Giavalisco, P., Eickhoff, H., Horn, M., Przevieslik, T., Theiss, D., Schneider, U., Lehrach, H., Gobom, J., 2001. Large-gel two-dimensional electrophoresis-matrix assisted laser desorption/ionization-time of flight-mass spectrometry: an analytical challenge for studying complex protein mixtures. Electrophoresis 22, 2844–2855.

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, 413–419.

Osborne, B.G., Fearn, T., Hindle, P.H., 1993. Practical NIR Spectroscopy with Applications in Food and Beverage Analysis. Longman Scientific & Technical, Harlow.

Pasquini, C., 2003. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. Journal of the Brazilian Chemical Society 14, 198–219.

Patterson, S.D., 2000. Proteomics: the industrialization of protein chemistry. Current Opinion in Biotechnology 11, 413–418.

Pleissner, K.P., Hoffmann, F., Kriegel, K., Wenk, C., Wegner, S., Sahlstrom, A., Oswald, H., Alt, H., Fleck, E., 1999. Proteome data analysis and management – new algorithmic approaches to protein spot detection and pattern matching in two-dimensional electrophoresis gel databases. Electrophoresis 20, 755–765.

Pun, T., Hochstrasser, D.F., Appel, R.D., Funk, M., Villars-Augsburger, V., Pellegrini, C., 1988. Computerized classification of two-dimensional gel electrophoretograms by correspondence analysis and ascendant hierarchical clustering. Applied Theoretical Electrophoresis 1, 3–9.

Rabilloud, T., Vincens, P., Tarroux, P., 1985. A new tool to study genetic expression using 2-D electrophoresis eata – the functional map concept. FEBS Lett. 189, 171–178.

Radzikowski, L., Nesic, L., Hansen, H.B., Jacobsen, S., Søndergaard, I., 2002. Comparison of ethanol-soluble proteins from different rye (Secale cereale) varieties by two-dimensional electrophoresis. Electrophoresis 23, 4157–4166.

Raman, B., Cheung, A., Marten, M.R., 2002. Quantitative comparison and evaluation of two commercially available, two-dimensional electrophoresis image analysis software packages, Z3 and Melanie. Electrophoresis 23, 2194–2202.

Scanlon, M.G., Pritchard, M.K., Adam, L.R., 1999. Quality evaluation of processing potatoes by near infrared reflectance. Jornal of the Science of Food and Agriculture 79, 763–771.

Schmilovitch, Z., Hoffman, A., Egozi, H., Ben Zvi, R., Bernstein, Z., Alchanatis, V., 1999. Maturity determination of fresh dates by near infrared spectrometry. Journal of the Science of Food and Agriculture 79, 86–90.

Schultz, J., Gottlieb, D.M., Petersen, M., Nesic, L., Jacobsen, S., Søndergaard, I., 2004. Explorative data analysis of 2-D electrophoresis gels. Electrophoresis 25, 502–511.

Sefara, N.L., Magtoto, N.P., Richardson, H.H., 1997. Structural characterization of p-lactoglobulin in solution using two-dimensional FT mid-infrared and FT near-infrared correlation spectroscopy. Applied Spectroscopy 51, 536–540.

Siesler, H.W., 2002. Introduction. In: Siesler, H.W., Ozaki, Y., Kawata, S., Heise, H.M. (Eds.), Near-infrared Spectroscopy. Principles, Instruments, Applications. Wiley, Weinheim, pp. 1–10.

Slaughter, D.C., Thompson, J.F., Tan, E.S., 2003. Nondestructive determination of total and soluble solids in fresh prune using near infrared spectroscopy. Postharvest Biology and Technology 28, 437–444.

Smilansky, Z., 2001. Automatic registration for images of two-dimensional protein gels. Electrophoresis 22, 1616–1626.

Steinberg, T.H., Lauber, W.M., Berggren, K., Kemper, C., Yue, S., Patton, W.F., 2000. Fluorescence detection of proteins in sodium dodecyl sulfate–polyacrylamide gels using environmentally benign, nonfixative, saline solution. Electrophoresis 21, 497–508.

Sumner, L.W., Mendes, P., Dixon, R.A., 2003. Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry 62, 817–836.

Tarroux, P., 1983. Analysis of protein-patterns during differentiation using 2-D electrophoresis and computer multidimensional classification. Electrophoresis 4, 63–70.

Tarroux, P., Vincens, P., Rabilloud, T., 1987. Hermes – A 2nd generation approach to the automatic-analysis of two-dimensional electrophoresis gels. 5. Data-analysis. Electrophoresis 8, 187–199.

Veeser, S., Dunn, M.J., Yang, G.-Z., 2001. Multiresolution image registration for two-dimensional gel electrophoresis. Proteomics 1, 856–870.

Veraverbeke, W.S., Delcour, J.A., 2002. Wheat protein composition and properties of wheat glutenin in relation to breadmaking functionality. Critical Reviews in Food Science and Nutrition 42, 179–208.

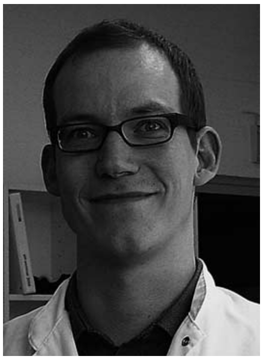
Vohradsky, J., 1997. Adaptive classification of two-dimensional gel electrophoretic spot patterns by neural networks and cluster analysis. Electrophoresis 18, 2749–2754.

Walker, A.K., Rymar, G., Andrews, P.C., 2001. Mass spectrometric imaging of immobilized pH gradient gels and creation of virtual two-dimensional gels. Electrophoresis 22, 933–945.

Wesley, I.J., Larroque, O., Osborne, B.G., Azudin, N., Allen, H., Skerritt, J.H., 2001. Measurement of gliadin and glutenin content of flour by NIR spectroscopy. Jornal of Cereal Science 34, 125–133.

Wu, Y.Q., Czarnik-Matusewicz, B., Murayama, K., Ozaki, Y., 2000. Two-dimensional near infrared spectroscopy study of human serum albumin in aqueous solutions: using overtones and combination modes to monitor temperature-dependent changes in the secondary structure. Journal of Physical Chemistry B 104, 5840–5847.

David M. Gottlieb is Chemist at the Plasma Product Division at Statens Serum Institute in Copenhagen, Denmark. Before entering his current position he worked as a research assistant in the Biochemistry and Nutrition Group at BioCentrum-DTU. The work of Dr. Gottlieb has been centred on the use of multivariate statistics in the analysis of wheat proteins and wheat properties. He graduated from the Technical University of Denmark as chemical engineer under the supervision of Prof. Ib Søndergaard and Prof. Susanne Jacobsen. His studies at the Technical University of Denmark have been supplemented with studies in pharmacology and biochemistry at the University of Liverpool.



Jakob Schultz is working as a research assistant in the Biochemistry and Nutrition Group at BioCentrum-DTU. The work of Dr. Schultz is centred on the development of novel ways to analyse two-dimensional electrophoresis gels, focusing on easy extraction of relevant data. This among other things involves the use of multivariate statistics, multivariate image analysis and proteomics. He graduated from the Technical University of Denmark as chemical engineer under the supervision of Prof. Ib Søndergaard and Prof. Susanne Jacobsen. His studies at the Technical University of Denmark have been supplemented with studies in food technology and brewery technology at the Technical University of Munich.

Susanne Jacobsen is a Ph.D. in biochemistry and associate professor at BioCentrum-DTU, Biochemistry and Nutrition Group, at the Technical University of Denmark in Kgs. Lyngby, Denmark.The work of Prof. Jacobsen has been centred on proteins, proteomics and the integration of biochemical analysis methods – mainly electrophoretic and mass spectrometric methods – with the use of digital image processing, neural networks and multivariate statistical methods. The goal is to gain insight into complex biochemical and biological system using non-destructive and non-invasive methods.

Susanne Wrang Bruun is Ph.D. student in the Biochemistry and Nutrition Group at BioCentrum-DTU. The work of Dr. Bruun is centred on the development of novel ways to analyse cereal protein interactions with proteins and carbohydrates. Near infrared spectrometry is used in combination with multivariate statistics. She graduated from the Biochemistry and Nutrition Group at BioCentrum-DTU, Technical University of Denmark as chemical engineer. As part of the Ph.D. study she is currently studying infrared, near infrared technology and multivariate data analysis at MatForsk, Aas, Norway.

Ib Søndergaard is biochemist and associate professor at BioCentrum-DTU, Biochemistry and Nutrition Group, at the Technical University of Denmark in Kgs. Lyngby, Denmark. Before entering his current position he worked at the Royal Veterinary and Agricultural University, the University Hospital of Copenhagen and the University of Copenhagen. Prof. Søndergaard is currently a member of the Danish Agricultural and Veterinary Research Council and the Danish Nutrition Council. The work of Prof. Søndergaard has been centered on proteins, proteomics and the integration of biochemical analysis methods – mainly electrophoretic and mass spectrometric methods – with the use of digital image processing, neural networks and multivariate statistical methods. The goal is to gain insight into complex biochemical and biological system using non-destructive and non-invasive methods.

Appendix II

Paper II

Correcting ATR-FTIR spectra for water vapour and carbon dioxide.

In collaboration with Achim Kohler, Isabelle Adt, Ganesh D. Sockalingum
Michel Manfait and Harald Martens

To be submitted to Journal of Applied Spectroscopy

SUMMARY

Fourier transform infrared (FTIR) spectroscopy is a valuable technique for characterisation of biological samples, providing a detailed fingerprint of the major chemical constituents. However, water vapour and CO_2 in the beam path often cause interferences in the spectra, which can hamper the data analysis and interpretation of results. In this paper we present a new method for removal of the spectral contributions due to atmospheric water and CO_2 from ATR-FTIR spectra. In the IR spectrum, four separate wavenumber regions were defined, each containing an absorption band from either water vapour or CO_2. From two calibration data sets, gas model spectra were estimated in each of the four spectral regions, and these model spectra were applied for correction of gas absorptions in two independent test sets (spectra of aqueous solutions and a yeast biofilm (*C. albicans*) growing on an ATR, respectively). The amounts of the atmospheric gases as expressed by the model spectra were estimated by regression, using second derivative transformed spectra, and the estimated gas spectra could subsequently be subtracted from the sample spectra. For spectra of the growing yeast biofilm, the gas correction revealed otherwise hidden variations of relevance for modelling the growth dynamics. The presented method has proven to be a valuable tool for filtering atmospheric variation in ATR-FTIR spectra.

Index headings: FTIR, attenuated total reflection (ATR), atmospheric correction, atmospheric absorptions, principal component analysis (PCA).

INTRODUCTION

In biological sciences, Fourier transform infrared (FTIR) spectroscopy has proven to be an important tool for measuring an overall chemical fingerprint of very different samples. For the monitoring of biological processes in fluids like fermentation and enzymatic reactions, the attenuated total reflection (ATR) FTIR technique has turned out to be very useful.[1] This technique prevents saturation of the water peaks for aqueous samples, as a micrometer scale path length is achieved. Thus, FTIR is a versatile tool that allows for non-destructive simultaneous quantification of a rich diversity of chemical constituents (proteins, lipids, carbohydrates, free vs. bound water, etc.) and certain physical properties (e.g. light scattering). The multi-component detection is beneficial, for example in the real-time monitoring of bio-films, grown directly on the ATR-crystal.[2,3] The relevant information is extracted from FTIR spectra by use of chemometric methods.

However, the analysis of FTIR spectra is often hampered by IR absorptions due to uncontrolled, varying amounts of water vapour and CO_2 in the light path. The atmospheric absorption of water vapour demonstrates characteristic absorption bands, each showing a symmetric pattern of fine spectral lines and involving the excitation of a chemical bond vibration. As each vibrational state is associated with many rotational levels (J), spectral lines appear symmetrically in two branches of the band, resulting from transitions involving ΔJ= +1 and ΔJ= -1, respectively. These lines are high-frequent compared to the absorptions from liquids and solids.[4] For CO_2, the absorption lines are less resolved, and a high instrument resolution is needed in order to see the individual lines of the CO_2 bands.

The atmospheric absorptions represent an experimental nuisance, creating unwanted, but systematic patterns which often cannot be completely avoided by purging of the instrument with gaseous N_2 or using background subtraction. The concentrations of water vapour and CO_2 are likely to vary between the background measurement and the sample scanning. This problem is for example encountered when a process is followed over a long time in the ATR-cell (e.g. biofilm development). For ATR-FTIR measurements of aqueous samples, evaporation from the sample itself may increase the humidity in the sample compartment and give rise to the problems, especially if the measurements are carried out at elevated sample temperatures.

The CO_2 bands are not as wide as the water vapour bands and the major band is found in a spectral region devoid of absorptions from the main biochemical components (lipids, proteins and carbohydrates). On the other hand, the water vapour absorptions overlap with some important bands from proteins and lipids and, in particular, the presence of water vapour absorptions in the amide I region (1700-1600 cm^{-1}) presents a problem in analysis of protein secondary structures based on the amide I band. Even small contributions from water vapour may hamper this analysis and lead to incorrect band assignments, as these absorptions are amplified by the resolution enhancement (e.g. by Fourier self deconvolution).[5]

Thus, it appears necessary to correct the spectra for the atmospheric contributions in order to obtain any sample information or to improve accuracy and precision of the FTIR calibration models. However, as the sub bands of amide I show line widths comparable to those of the water vapour lines, information may be lost if simple low-pass filtering is used for reducing the water vapour effects. Commonly, a gas spectrum is collected, whereafter the subtraction from each sample spectrum is accomplished by use of an algorithm in the spectroscopic software. The main problems associated with this approach relate to difficulties in determining the correct subtraction factor and in obtaining exactly the same band shapes for the reference and the sample spectrum. As regards the latter problem, the sample itself may influence the spectral shape due to the influence on the beam geometry. In addition a low instrument resolution causes distortion of the spectral lines and results in non-linearities. The present paper shows a model-based pre-processing tool for ATR-FTIR spectra to minimise water vapour and CO_2 spectral contributions. Successful removal of these interferences is demonstrated for different types of samples.

Notation and terminology

I means measured light intensity, while A means absorbance, defined as $\log_{10}(I)$, or $\log_{10}$ (I/I_0), when mentioned explicitly. $I(\nu)$ means measured intensity at ν cm^{-1}. $A_{i,k}$ means absorbance in sample i at wavenumber channel k.

EXPERIMENTAL

All measurements were performed on a Bruker Equinox 55 FTIR instrument, equipped with a liquid-nitrogen-cooled MCT detector, scanning from 4000 cm^{-1} to 600 cm^{-1}. The nominal instrument resolution was 4 cm^{-1}, but the spectral readings were recorded at 2 cm^{-1} intervals. Aqueous and biofilm samples were measured on a horizontal (ATR) ZnSe crystal with ~5 internal reflections. Each spectrum resulted from coaddition of at least 128 scans (if not otherwise mentioned) obtained in single beam mode. Background spectra were obtained on the empty sample holder.

Four FTIR experiments were carried out, as described in the following. Gas model spectra for water vapour and CO_2 were obtained from Experiment 1 (providing a data set of only atmospheric absorptions) and Experiment 2 (providing a data set with also liquid water absorptions). The model performance is tested on a basic and a realistic/complex data set, obtained from Experiment 3 and 4, respectively.

Experiment 1: Gas calibration measurements and estimation of four primary gas model spectra

A total of 120 FTIR spectra of various concentrations of water vapour and gaseous CO_2 were obtained by acquiring several spectra in an empty sample compartment during the replacement of room air with gaseous N_2 in the beam path. The N_2 purging was started after closing of the sample compartment and spectra were recorded once a minute for 60 minutes. Each spectrum resulted from coaddition of 32 scans. This experiment was performed twice.

Experiment 2: Water calibration extension measurements

Milli-Q water (distilled and ion-exchanged) and various salt solutions of low concentrations (0.2-1.0 M) of NaCl, $MgSO_4$ or $NaClO_4$ (prepared in Milli-Q water) were measured on the ATR crystal, mounted in a closed sample cell, which allowed for temperature control by circulation of heating or cooling water in the space around the ATR-cell. A total of 205 sample spectra were recorded with different levels of water vapour and CO_2 in the sample compartment. Water was measured at different temperatures between 8 and 60°C, while salt solutions were measured at 15, 22 and 29°C. The FTIR instrument was kept at room temperature in all measurements. These ATR-FTIR spectra of aqueous solutions were used to define additional “nuisance” spectra and to develop a between-regions calibration model (see later).

Experiment 3: Water test measurements

Water spectra at different temperatures and with different water vapour and CO_2 levels were measured by placing Milli-Q water in the ATR-cell at one temperature (about 22°C), closing the sample compartment and starting N_2 purging. The spectra were read at consecutive points in time while the sample was changed gradually towards another temperature (about 10°C) by slowly decreasing the temperature of the cooling water. The FTIR instrument itself was kept at room temperature. In total 25 spectra were recorded in this time series. These spectra were used for testing how well the “nuisance” gas modelling worked for samples similar to some of the calibration samples.

Experiment 4: Yeast cell culture test measurements

Candida albicans (strain SC 5314, ATCC collection) was grown in 10 ml Sabouraud Medium (bioMérieux, France) for 24 hours. 300 μl of this culture were added to 3 ml of fresh medium and placed on the ATR crystal at room temperature (21°C) and a biofilm was allowed to develop. Biofilm growth was monitored by ATR-FTIR during 19 hours and a total of 58 spectra were collected at intervals of 20 minutes throughout the growth period. 64 scans were coadded for obtaining each final spectrum. These samples were used for testing how the correction for water vapour and CO_2 performed for sample spectra that were very different from the calibration sample spectra

Determining the primary gas model spectra

The spectra in Experiment 1 were used for defining primary “nuisance” spectra. Temporal absorbance differences $D_{i,k} = A_{i-1,k}- A_{i,k}$, for samples i=1,2,3,.. at wavenumber channels # k=1,2,…1764 (corresponding to the region 4000-600 cm^{-1}) were computed. This was done in order to reduce the impact of possible instrument drift during the hour-long experiments. The H_2O(g) ranges k= 1:400 (4000: 3231 cm^{-1}) and 1001:1450 (2072:1205 cm^{-1}), and the CO_2 ranges k=809:930 (2442:2208 cm^{-1}) and 1601:1764 (914:600 cm^{-1}) were modelled separately; each containing bands from H_2O(g) or CO_2 (see Table 1).

	Wavenumbers [cm^{-1}]	Absorption band
Water vapour region 1	4000-3231	Sym. and asym. stretching $\nu_{1,3}$
Water vapour region 2	2072-1205	Bending ν_2
CO_2 region 1	2442-2208	Asym. stretching
CO_2 Region 2	914-600	Bending

Table 1. The four defined gas regions. Sym: symmetric. Asym: antisymmetric.

An uncentred principal component analysis (PCA), i.e. a singular value decomposition (svd) was performed in each region, and the first two principal components (PC) loading vector of $D_{i,k=k1:k2}$ were in each case extracted and saved as a primary “nuisance” spectrum. (Outside the defined ranges, the values are zero). The first component turned out to be a typical water vapour or CO_2 spectrum, while the second component models the major systematic variation around the typical spectrum in each region (for example caused by non-linearities and by pressure/temperature changes). Thus, loadings for a two-component bi-linear gas model were defined for each of the four gas regions: Two water vapour regions and two CO_2 regions.

Determining the secondary gas model spectra

The eight primary gas model spectra estimated in Experiment 1 (two components in each of the four regions) were applied (see below) to reduce gas contributions in the spectra in Experiment 2. These gas-reduced spectra were subsequently smoothed with a moving-average filter, averaging each channel with its 5 left- and right-hand neighbour channels. The difference between the gas-reduced spectra and their smoothed versions showed one clear remaining high-frequent variation pattern in the two H_2O(g) regions and in the main CO_2 region. Hence, an additional, secondary "nuisance" gas model spectrum was determined by svd of the differences in each of the two H_2O(g) regions and in the first CO_2 region (the second CO_2 region was too noisy). The first loading vectors were used as secondary gas model spectra after orhogonalization to the eight primary gas model spectra. These secondary spectra may represent distortions of the band shapes introduced by the samples or the ATR-crystal, and their inclusion in the model is likely to improve the removal of the atmospheric absorptions for similar measurements.

As a result, the final bi-linear gas models had three model spectra in each of the water vapour regions and in the first CO_2 region and two model spectra in the last CO_2 region, i.e. in total 11 model spectra.

Correction for the gas model spectra

The concentrations of water vapour and CO_2 in the spectra may in principle be estimated (by least squares regression) based on the model of the absorbance ($A_{i,k}$) of sample *i* in wavenumber channel *k* (*k*=1,2,…1764), shown in Eq. 1.

Eq. 1. $$A_{i,k} = \sum_{j=1}^{J} c_{ij} k_{jk} + d_{i,k} + e_{i,k}$$

In Eq. 1, $c_{i,j}$ represents the concentration or score of gas element *j* =1,2,...,*J* (e.g. *J*=11), $k_{j,k}$ represents the model absorbance of gas spectrum *j* at channel *k*, and $d_{i,k}$ and $e_{i,k}$ represent other chemical and physical absorption effects and the measurement error, respectively, at this channel.

However, even when the gas model spectra $k_{j,k}$ are known, it is difficult to estimate their concentrations $c_{i,j}$, because the "interesting" chemical and physical absorption effects $d_{i,k}$ are usually unknown. If these unknown effects are large and ignored in the estimation of the gas scores, then they may create large alias errors in the gas score estimation. Therefore, we here estimate the gas scores only based on the high-frequent part of the spectra where the "non-gas" sample constituents and other phenomena that constitute $d_{i,k}$ have much more smooth features than the gas elements (This assumption may not be correct in all parts of the spectrum, especially for protein absorptions in the second water vapour region, see below). Thus, the gas concentration estimation is here done in the second derivative: For each sample *i* with spectrum $A_{i,k}$, the simplest negative second derivative was computed according to Eq. 2.

Eq. 2 $G_{i,k} = 2A_{i,k} - A_{i,k-1} - A_{i,k+1}$, *k*=2,3,.....

Similarly for each gas model element *j* with spectrum $k_{j,k}$, the simplest negative second derivative was computed according to Eq. 3.

Eq. 3. $h_{j,k} = 2k_{j,k} - k_{j,k-1} - k_{j,k+1}$, *k*=2,3,.....

Hence, we assume that the second derivative of the unknown, but smooth sample contribution spectrum $d_{i,k}$, k=1,2,... can be approximated by a simple unknown offset f_i at all channels in the second derivative. The second derivative model can then be written:

Eq. 4. $$G_{i,k} = \sum_{j=1}^{J} c_{i,j} h_{j,k} + f_i 1_k + e_{i,k}$$

Gas concentrations $c_{i,j}$ in the sample spectra are then estimated by least squares regression of $G_{i,k}$ on $h_{j,k}$, j=1,2,...,J and on vector **1**, by minimizing the residual sum of squares in $e_{i,k}$.

Indirect gas predictions: In case the spectra contain the amide I band from proteins, the above-mentioned assumptions about smoothness of the real sample spectra may not hold. The scores ("concentrations") of the water vapour components in the second $H_2O(g)$ region (containing the amide I region) may instead be estimated indirectly by prediction from three water vapour components in the first region by Eq. 4b, where J=3 represents the three components of the first $H_2O(g)$ region, and M=3 the three components of the second region:

Eq. 4b. $$c_{i,m} = \sum_{j=1}^{J} c_{i,j} b_{j,m} + b_{0,m} \quad m=1,2,..M$$

Likewise, the scores of the M=2 CO_2 components in its second wavenumber region were predicted from J=3 three CO_2 components in the first region by Eq. 4b. The model parameters $b_{j,m}$ and $b_{0,m}$ in Eq. 4b were estimated by full-rank regression for water vapour and for CO_2 separately using the scores $c_{i,m}$ and $c_{i,j}$ obtained in Experiment 2. The indirect score estimation models were then used for the prediction in Experiment 4, since samples in this experiment contain proteins.

With the gas scores $c_{i,j}$ thus estimated, the sample spectra are then gas-corrected by subtracting the gas scores $c_{i,j,}$ multiplied by the corresponding gas model spectrum (Fig. 2a-d). See Eq. 5.

Eq. 5. $$A_{i,k,gas-corrected} = A_{i,k} - \sum_{j=1|}^{J} \hat{c}_{i,j} k_{j,k}$$

Test of the final model

The final extended gas model thus consisted of three model spectra in each of the water vapour regions, three model spectra in the main CO_2 region and two model spectra in the second CO_2 region. This final gas model was applied for correction of the independent test sets from Experiments 3 and 4.

Evaluation of the gas model performance was done by inspection of scores and loadings from a PCA before and after the gas correction. In addition, water vapour indices were used for comparison of water vapour absorptions in the spectra. The index was calculated for each spectrum as A_{max} - A_{min} in the 1847-1837 cm^{-1} region in the second derivative and used as a measure of the water vapour level.[6]

All results were computed and displayed in the authors' software using Matlab (TM) version 7.0.

RESULTS

Experiment 1: Estimation of four primary gas model spectra

Fig. 1a shows the raw data (intensity *I*(ν) spectra) from the first of the time series in Experiment 1, showing water vapour and CO_2 in room air being gradually replaced with N_2 purging gas and hence leaving smaller and smaller intensity reductions. Two characteristic water vapour absorption patterns are evident in the 4000-3500 cm^{-1} and 2000-1200 cm^{-1} ranges.

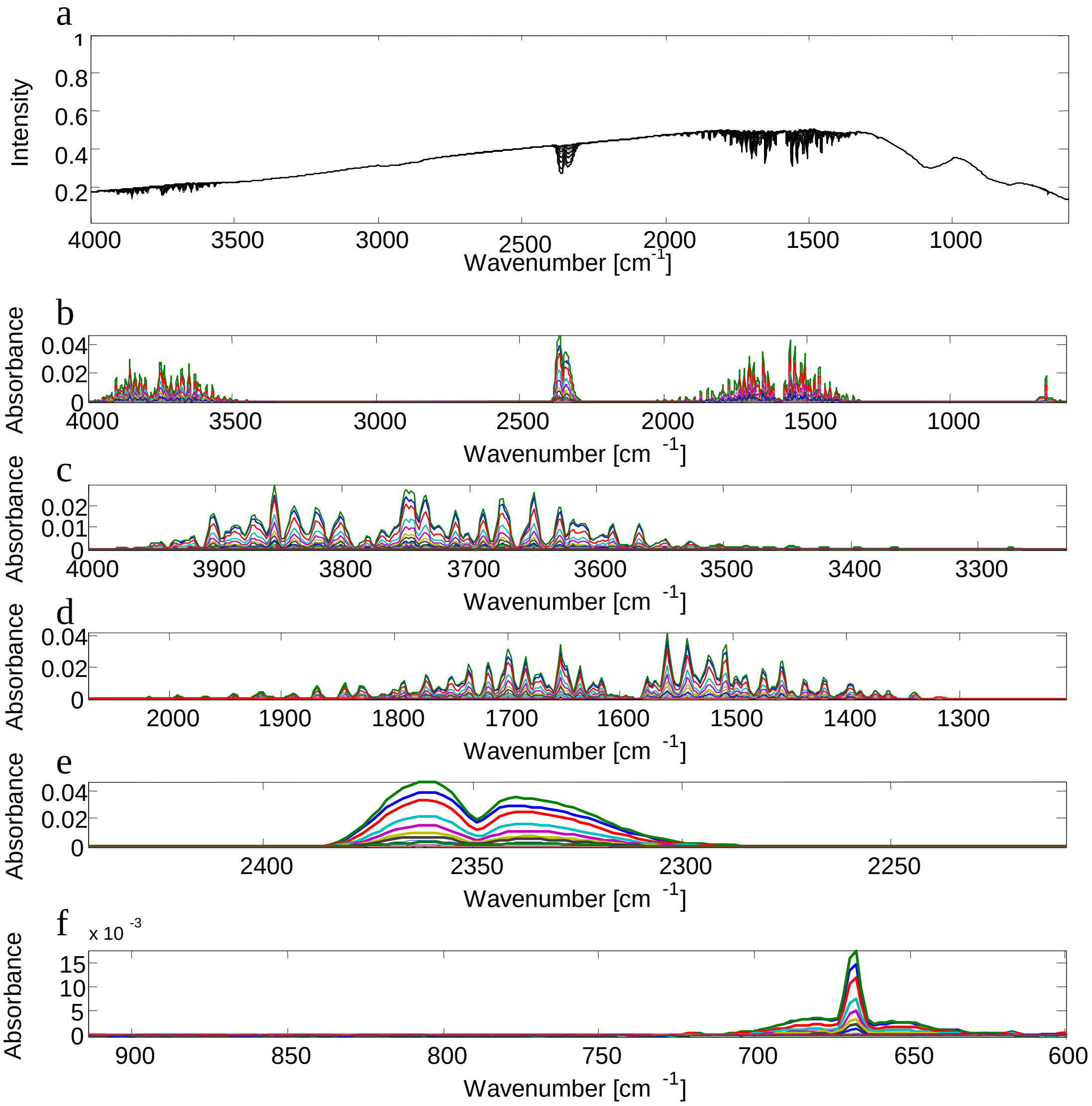


Fig. 1. Calibration spectra from one time-series in Experiment 1. Infrared spectra of room air were measured consecutively in time during purging with $N_2(g)$ to obtain different concentrations of water vapour and CO_2. a) Intensity spectra I(ν). b) Absorbance spectra A(ν)=-$\log_{10}$(I(ν)). c) Water vapour segment 1. d) Water vapour segment 2. e) CO_2 segment 1. f) CO_2 segment 2.

These are the results of rotation-vibration (“rovibrational”) transitions, associated with the stretching and bending vibrations of H_2O, respectively. Moreover, the double CO_2 absorption band around 2350 cm^{-1} is seen, as well as a weaker CO_2 band around 670 cm^{-1}, involving CO_2 antisymmetric stretching and CO_2

bending, respectively.[7] Only the latter CO_2 absorption demonstrates a so-called Q-branch in the middle of the band. Each water vapour and CO_2 band was contained in one of the four defined spectral regions, according to Table 1. In the first of two defined water vapour regions (4000-3231 cm^{-1}), the atmospheric water bands overlap only little with protein absorptions, whereas the interference is more dramatic in the second water vapour region (2072-1205 cm^{-1}), as this region contains amide bands from proteins as well as absorption bands from lipids. The first of the defined CO_2 regions (2442-2208 cm^{-1}) is outside the absorptions regions for the common biochemical constituents, whereas the second region (914-600 cm^{-1}) contains some important backbone vibrations of macromolecules.

Primary "nuisance" spectrum for each spectral region: The absorbance spectra $A(\nu)=-\log_{10}(I(\nu))$, i.e. without background correction, in Fig. 1b show more clearly the four gas absorptions, which are amplified in Fig. 1c-f. The peak sizes decrease with purging time. The svd of each gas element region showed that the first component accounted for more than 99 % of the total variance, while the second component, although small, also showed systematic contributions. This indicated that only one major and one minor spectral pattern of gas variation could be observed in each region. This was seen within each of the two time-series in this experiment. The two replicate PC loading vectors were very similar, for both component 1 and for component 2, in each of the four spectral regions. Hence, their mean (shown in Fig. 2a-d) was subsequently

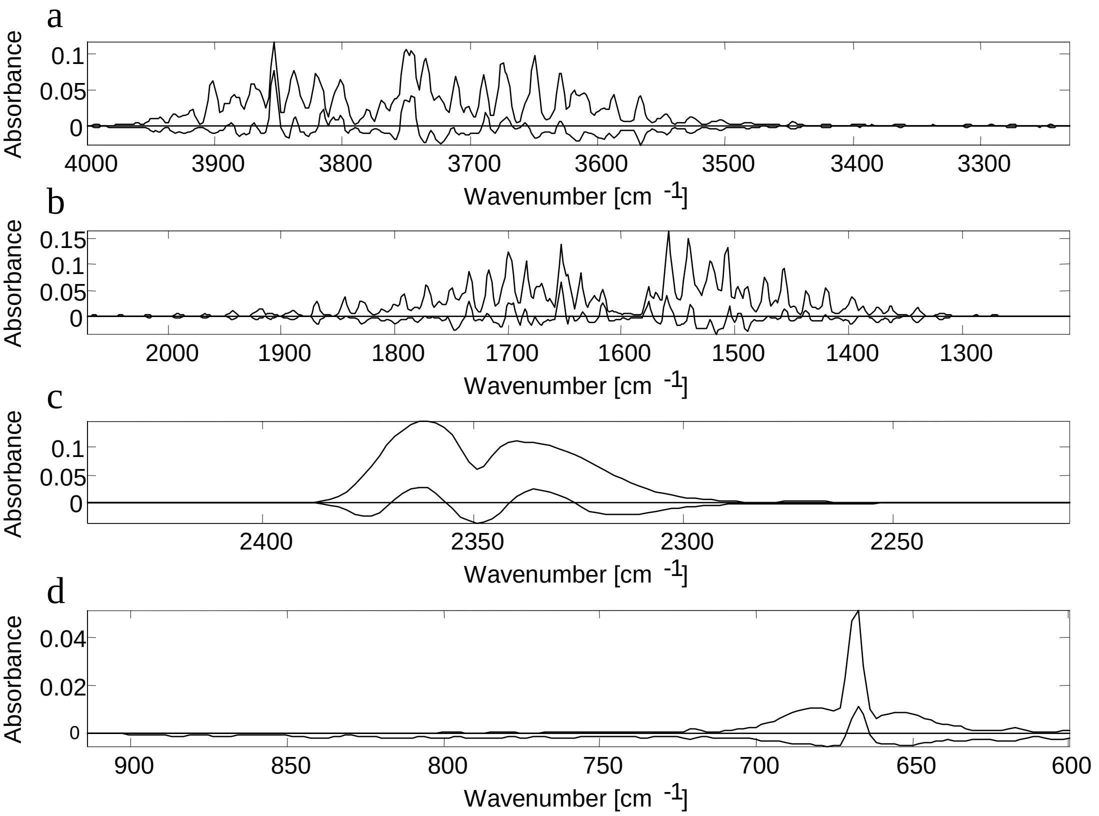
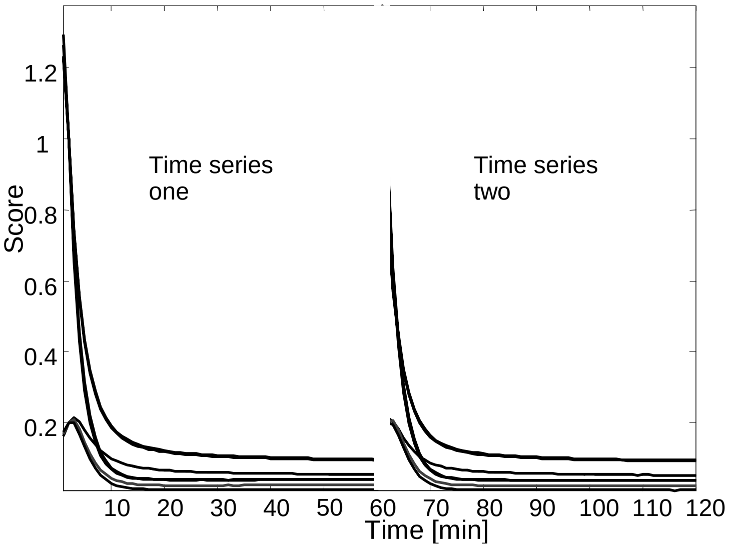


Fig. 2. Primary gas model spectra obtained from Experiment 1 by singular value decomposition (svd) of each gas region. Two component loading spectra are defined in each wavenumber region. In each region, the loading vectors for the first and the second component were normalised to correspond to a score value of 1 and 0.5, respectively, in an arbitrary sample (#2 in time series 1 in Experiment 1).
a) Water vapour segment 1. b) Water vapour segment 2. c) CO_2 segment 1. d) CO_2 segment 2.

used for each region. The first component loading (highest curve in each of the four plots) is seen to be almost non-negative, as expected for chemical absorbances. The second component loading (lowest curve in each of the four plots) has been amplified for visual clarity and it shows both negative and positive absorbances. In each case, the corresponding scores for this second component showed curved but temporally smooth relationship to the score of the first component (not shown here), indicating that it represents a weak but systemati c nonlinear effect (e.g. due to the low instrument resolution).

The levels of the eight primary "nuisance" gas components were estimated in the two time series and are seen to decrease as a function of gas purging time (Fig. 3). Fig. 3 reveals that very similar water vapour levels are estimated from each of the two regions of water vapour absorptions, and the same is found in case of CO_2 estimation from the two CO_2 regions.

Fig. 3. Gas scores for two time series in Experiment 1. Estimated levels of the primary gas components. Upper curves, starting near 1: the first component in the four wavenumber segment models. Lower curves, starting near 0.5: the corresponding second components.



Experiment 2: Estimation of three secondary gas model spectra

Absorbance spectra of the 205 aqueous solutions measured in the ATR-cell are shown in Fig. 4a. These samples come from several measurement series taken over two months. The spectra display the three broad absorbance peaks around 3300, 1640 and 800 cm^{-1} from liquid water plus a major peak around 1100 cm^{-1} (due to the presence of anions SO_4^{2-} and ClO_4^- in some of the samples). However, upon closer inspection, the presence of some high-frequent water vapour contributions is evident, as well as contributions from the first CO_2 double-peak. In contrast, the second CO_2 peak is not readily visible. The spectra were corrected for the eight primary gas "nuisance" spectra by the procedure described by Eq. 1-5 in the Methods section: The second derivative transformed sample spectra were regressed on the second derivative transformed model spectra (plus a local polynomial), and the hereby estimated gas contents were used for determination of the primary spectral gas contributions.

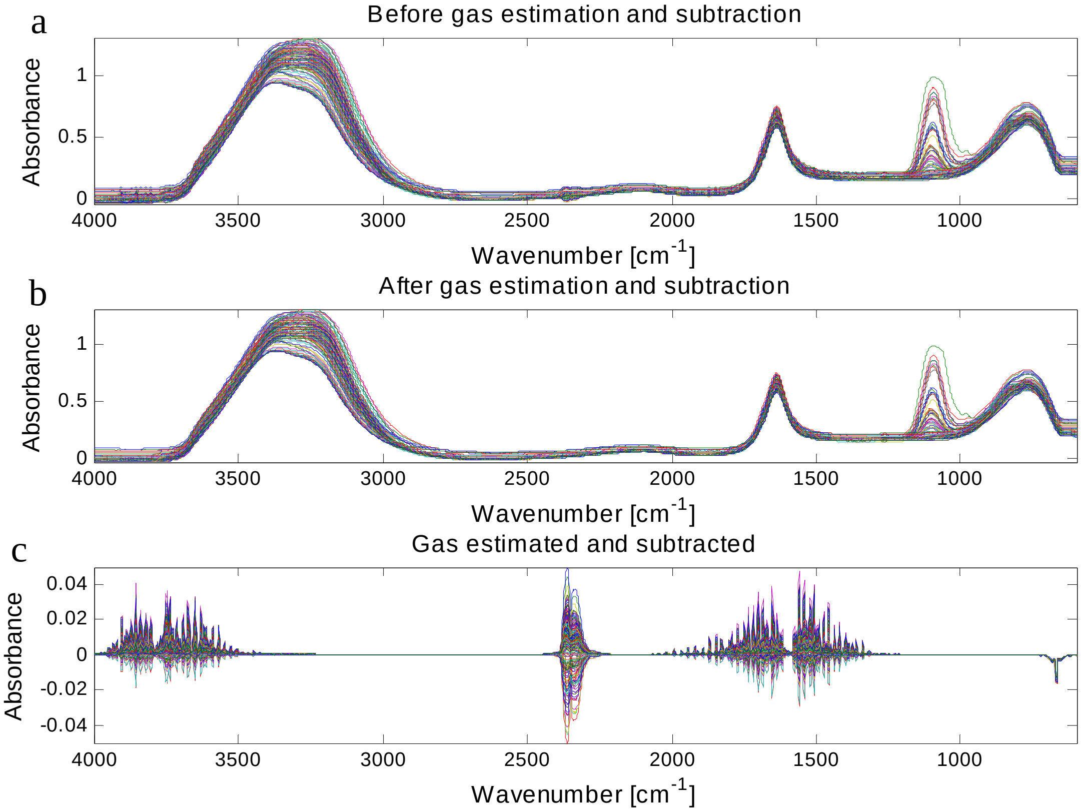


Fig. 4. Calibration spectra of water and salt solutions at different temperatures obtained by ATR-FTIR in Experiment 2. a) Raw absorbance spectra. b) Spectra corrected for the eight primary gas model spectra. c) Estimated gas spectra, which are subtracted from the raw spectra.

These were subsequently removed from the non-transformed spectra. The corrected spectra are seen in Fig. 4b. The estimated and subtracted primary gas contributions (Fig. 4c) either have positive or negative gas peaks in each region, illustrating how the samples may have higher or lower gas content than a reference sample (background spectrum).

Ideally, the high-frequency residuals of the corrected spectra (i.e. after subtracting the low-pass filtered versions of the spectra) should only contain measurement noise. However, upon closer inspection, these high-frequency residuals display some remaining systematic high-frequency patterns in three of the four gas-channel regions (Fig. 5a). Therefore, a separate svd of the residuals in each of these three regions was performed. The first component loading was extracted (Fig. 5b) and these were projected (see e.g. Martens & Naes 1989)[8] on the primary "nuisance" spectra. The orthogonalized residuals (Fig. 5b) were used as secondary loadings i.e. secondary "nuisance" spectra (orthogonal to the primary "nuisance" spectra). They may account for deviations from the atmospheric spectra in Experiment 1, caused by the samples or the ATR crystal.

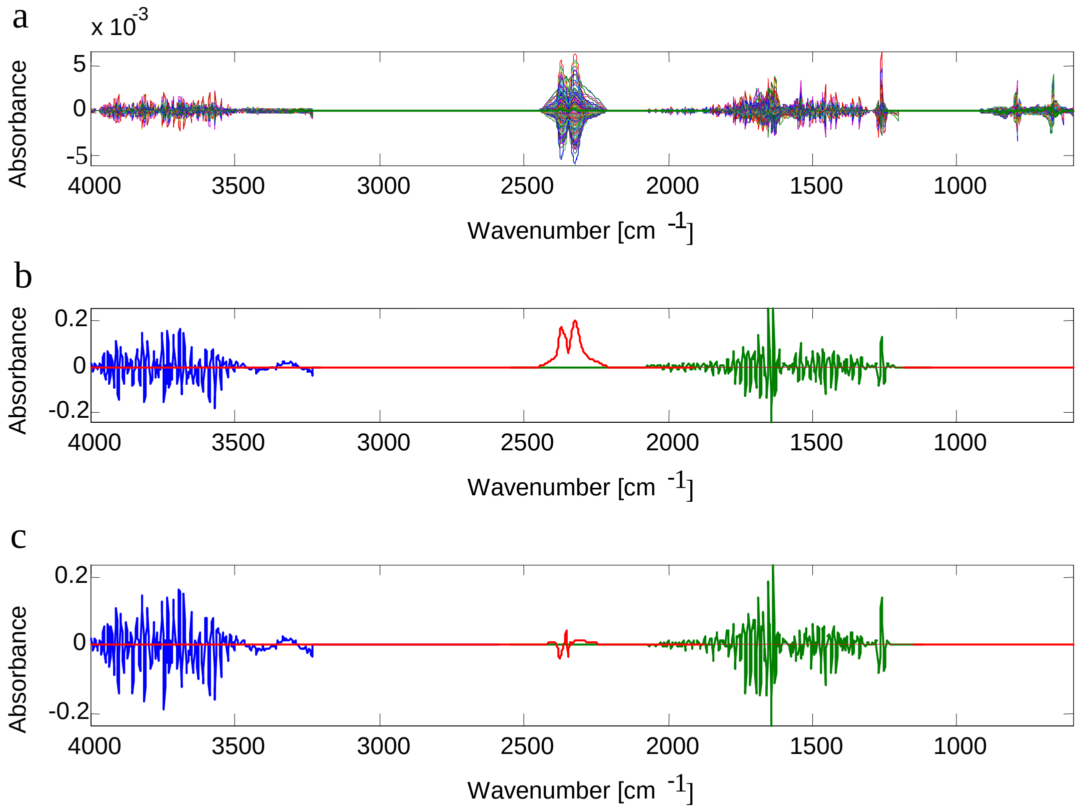


Fig. 5. Secondary gas model spectra from residuals in Experiment 2. a) High-frequency residuals in the spectra from Fig. 3b. b) Loading of first component from svd of the residuals in Fig. 5a for three gas regions (the two water regions and the main CO_2 region). c) The secondary model spectra, obtained by ortogonalizing the spectra in Fig. 5b on the primary model spectra (Fig. 2a-d).

Indirect gas prediction model: To what extent do the two different water vapour regions give similar estimates of water vapour concentration in a given sample's spectrum? Fig. 6 shows various relationships between water vapour scores from regions 1 and 2, for components 1, 2 and 3, respectively. The top row shows the estimate from region 1 vs. the estimate from region 2. Considering that the loading spectra for the first two (primary) components have been obtained from a rather different measurement situation (Experiment 1) than the present, the similarity between the scores from the two regions is good (seen from the closeness to the diagonal where the abscissa equals the ordinate). Only a scaling difference is seen for the first component (Fig. 6a), while an offset is seen for the second and much smaller component (Fig. 6b). As the two regions resulted in rather similar scores in Experiment 1, the deviations seen for Experiment 2 (Fig. 6a,b) may be related to the large differences between the spectra in the two experiments (of the empty sample compartment and of water on the ATR).

Fig. 6c shows that the two regional loading spectra for the third (secondary) components obtained in Experiment 2 reflect much of the same systematic variation phenomenon, presumably related to varying water vapour concentrations. Hence, while the two wavenumber regions gave clearly related score estimates, these are not equal, and the two water vapour regions thus may need to be handled separately.

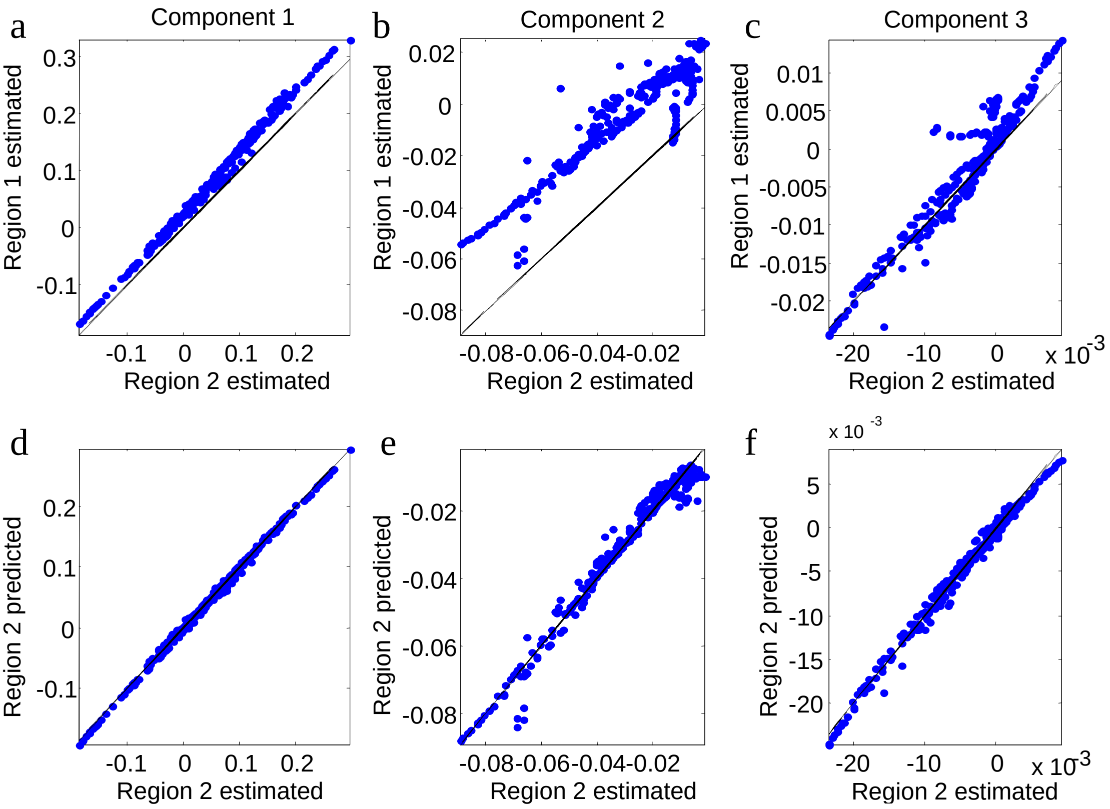


Fig. 6. Score relationships for water vapour in Experiment 2. Left, center, right: components 1, 2 and 3. Top: score estimate in region 1 (ordinate) vs. score estimate in region 2 (abscissa). Bottom: score estimate in region 2, predicted from score estimates in region 1 (ordinate) vs. score estimate in region 2 (abscissa).

To prepare for situations where protein and lipid signals make water vapour concentrations estimation difficult in water vapour region 2, linear full-rank calibration models were established from the spectra in Experiment 2, to predict the water vapour component scores in region 2 from those in region 1 (i.e. the component scores in region 2 were regressed on to the component scores in region 2). Fig. 6d-f show the prediction ability of the models. For each of the three components it seems that the region 2 water vapour scores can be successfully predicted from the region 1 water vapour scores, at least for these types of spectral measurements. The degree of over-fitting in the predicted scores in Fig. 6d-f is considered negligible, since the number of observations (205) is far higher than the number of independent parameters (4) estimated for each component in Eq. 4b. Similar calibration models were developed for predicting CO_2 scores in the second CO_2 range from the first CO_2 range. However, the modelling of (small) CO_2 effects in the second range was obscured by measurement errors in this region, which must be considered unreliable. These calibration models to predict regions 2 from regions 1 for water vapour and CO_2 will be used in Experiment 4 to avoid confounding with proteins and lipids.

Experiment 3: Test of model on ATR-FTIR spectra of aqueous samples

Absorbance ATR-FTIR spectra obtained on pure water, dropping gradually from 21 to 10°C, are shown in Fig. 7a. The absorbance at 4000 cm^{-1} was subtracted from each spectrum to correct for irrelevant baseline variations. The gas peaks and the temperature effect on the water spectrum become apparent after mean centring each channel (Fig. 7b).

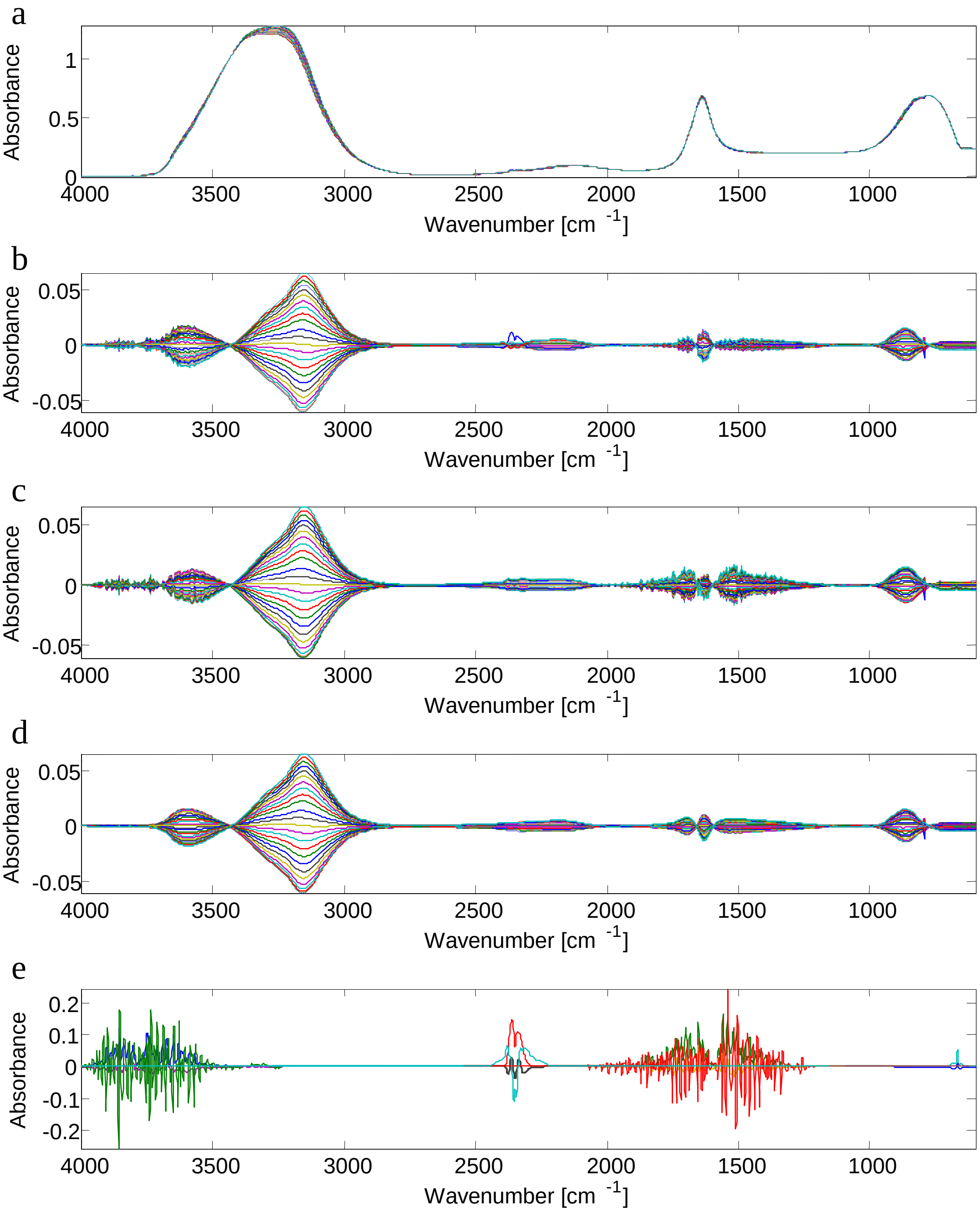


Fig. 7. Test of the gas model, using ATR-FTIR spectra of pure water at different temperatures, obtained in Experiment 3. a-b) Raw absorbance spectra before and after mean centring. c-d) Mean centred spectra after subtraction of gases, estimated in raw absorbance and in second derivative, respectively. e) Gas model spectra.

In Fig. 7c and 7d, the mean centred spectra are seen after gas correction, based on estimation using the raw absorbance data and their second derivatives, respectively. Fig. 7e shows the gas component model.

The high amount of remaining water vapour in the corrected spectra in Fig. 7c illustrates the alias problem of the simpler gas estimation alternative based on the actual absorbance spectra instead of their second derivatives. Here, each spectrum in Fig. 7a have been regressed on the gas model spectra in Fig. 7e without first taking second derivatives and the regression residuals are used as gas-corrected spectra. (The gas model had to be compacted to give meaningful gas score estimates: 3 water vapour component spectra and 3 CO2 component spectra were defined by summing the wavenumber regions 1 and 2 in the original 11 model spectra in Fig. 7e). In contrast, the gas correction based on estimation of gas scores in the second derivative (Fig. 7d) effectively eliminated the gas nuisance effects.

Fig. 8 illustrates the consequences for the subsequent data analysis, when applying no gas correction or when using the two different approaches to the gas correction.

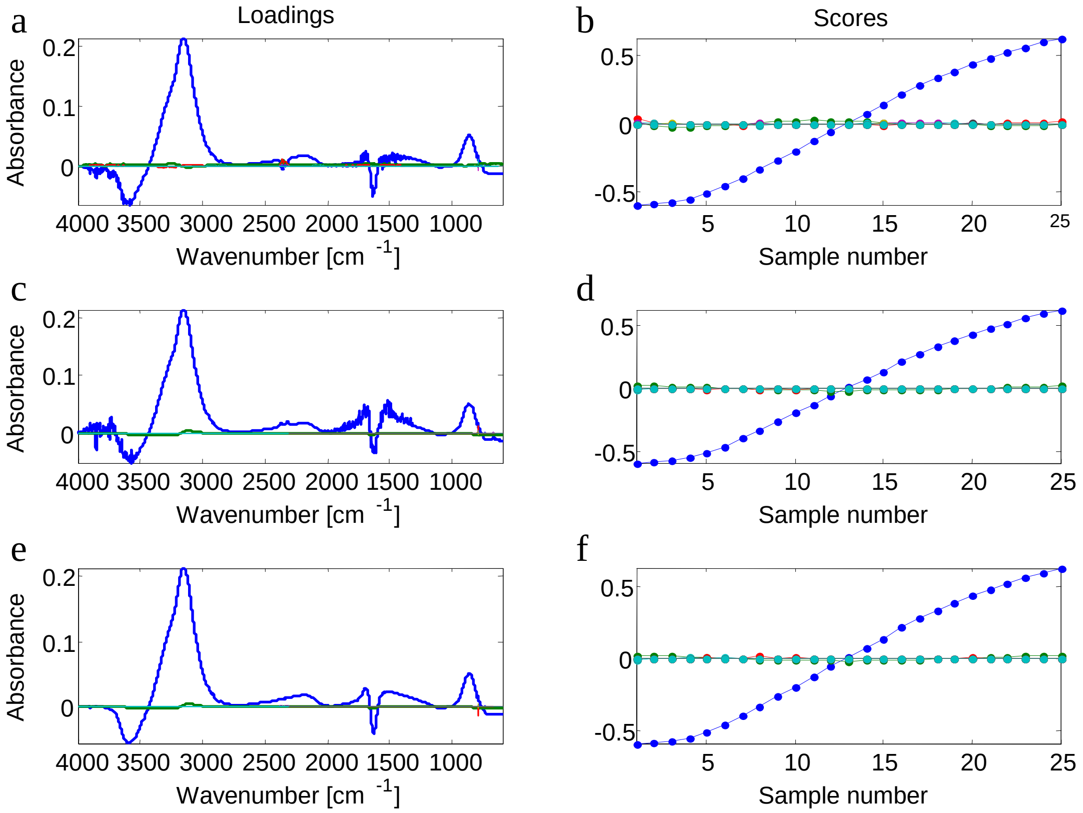


Fig. 8. Results from PCA models of the spectra, obtained in Experiment 3. a,b) Raw spectra. c,d) After gas correction, estimated in raw absorbance. e,f) After gas correction, estimated in second derivative. Left side: the spectral variation patterns found, i.e. PCA loadings vs. wavenumber. Right side: sample models i.e. scores vs. time.

Simple PCA, i.e. svd of the mean-centred spectra in Fig. 7b-d was used in order to extract the major systematic patterns of variation in the FTIR spectra as the liquid water sample cooled from about 21°C to about 10°C. The PC loading spectra and the PC scores, both scaled by the size of the component, are shown on the left and right side, respectively, of Fig. 8. One PC seems to dominate the data in all three cases. This variation is presumed to reflect the temperature-dependent changes between a weakly and a strongly hydrogen bonded state of liquid water. More detailed interpretation of this phenomenon is presented by Martens et al. (2006)[9]. Of more interest presently is the water vapour nuisance effect: The scores of the first

PC is not affected by water vapour and CO_2 nuisances, as these are minor effects compared to the temperature-induced variation in liquid water. However, looking at the loading spectra (left hand-side of Fig. 8), which are intended to represent the effect of temperature on liquid water, these are strongly contaminated by irrelevant high-frequent water vapour signal in case of the untreated data (Fig. 8a) and the over-simplified gas score estimation (Fig. 8c).

It was not possible to use regression for estimation of the water temperature effect, since the precise temperatures were not known for these samples. Anyway, this approach would have been to no avail, as the temporal change in the temperature of the liquid water sample in the ATR is strongly confounded with the temporal change in water vapour concentration inside the instrument.

In contrast, the gas-corrected spectra based on second derivative estimates (Fig. 7d) gave an estimated temperature effect on the liquid water (Fig. 8e) with little or no trace of the water vapour or CO_2.

Experiment 4: Yeast cell culture test measurements

Fig. 9 summarizes the effect of pre-processing on in situ spectra of the live cell culture of *C. albicans*, growing on the ATR surface. The absorbance spectra measured, $A=-\log_{10}(I(\nu))$ (Fig. 9a) are rather similar to that of pure water (Fig. 7a) except for a higher baseline at the lower wavenumbers. After having subtracted the mean of these spectra, Fig. 9b shows that the variation in this data set is dominated by irrelevant water vapour and CO_2 contributions, presumably from inside the FTIR instrument, plus a weaker variation around 1000 cm^{-1} from the samples themselves.

The concentrations of the water vapour and CO_2 components were estimated in second derivative in the upper wavenumber regions (above 2100 cm^{-1}) but not in the lower wavenumber regions (below 2100 cm^{-1}), for fear that informative, high-frequent signals from proteins etc. might be lost. Instead, the scores in the lower wavenumber region were predicted from those in the upper region by the calibration models obtained in Experiment 2 (Fig. 6). The second row in Fig. 9 shows the gas-corrected absorbance spectra before and after mean centring. (The wavenumbers below 856 cm^{-1} were weighted to zero because the measurements were considered unreliable in this region). Apparently, most of the water vapour and CO_2 contributions have been removed, revealing other systematic variation patterns in several regions. However, some minor gas (especially CO_2) contributions are still evident.

Therefore, to further reduce the CO_2 effect, each spectrum was linearly interpolated locally under the CO_2 peak, i.e. from 2442 cm^{-1} to 2208 cm^{-1}. Moreover, the light scattering of the *C. albicans* suspension may be expected to change with growth time. (Light scattering effects in the spectra may result from changes in refractive index of the sample over time. As the refractive index affects the penetration depth of the IR light, this variation may cause multiplicative scaling effects. Also, additive baseline effects appear. For removal of this physical variation, the spectra were subjected to Extended Multiplicative Signal Correction (EMSC)[10] as described in Martens et al. (2003).[11]

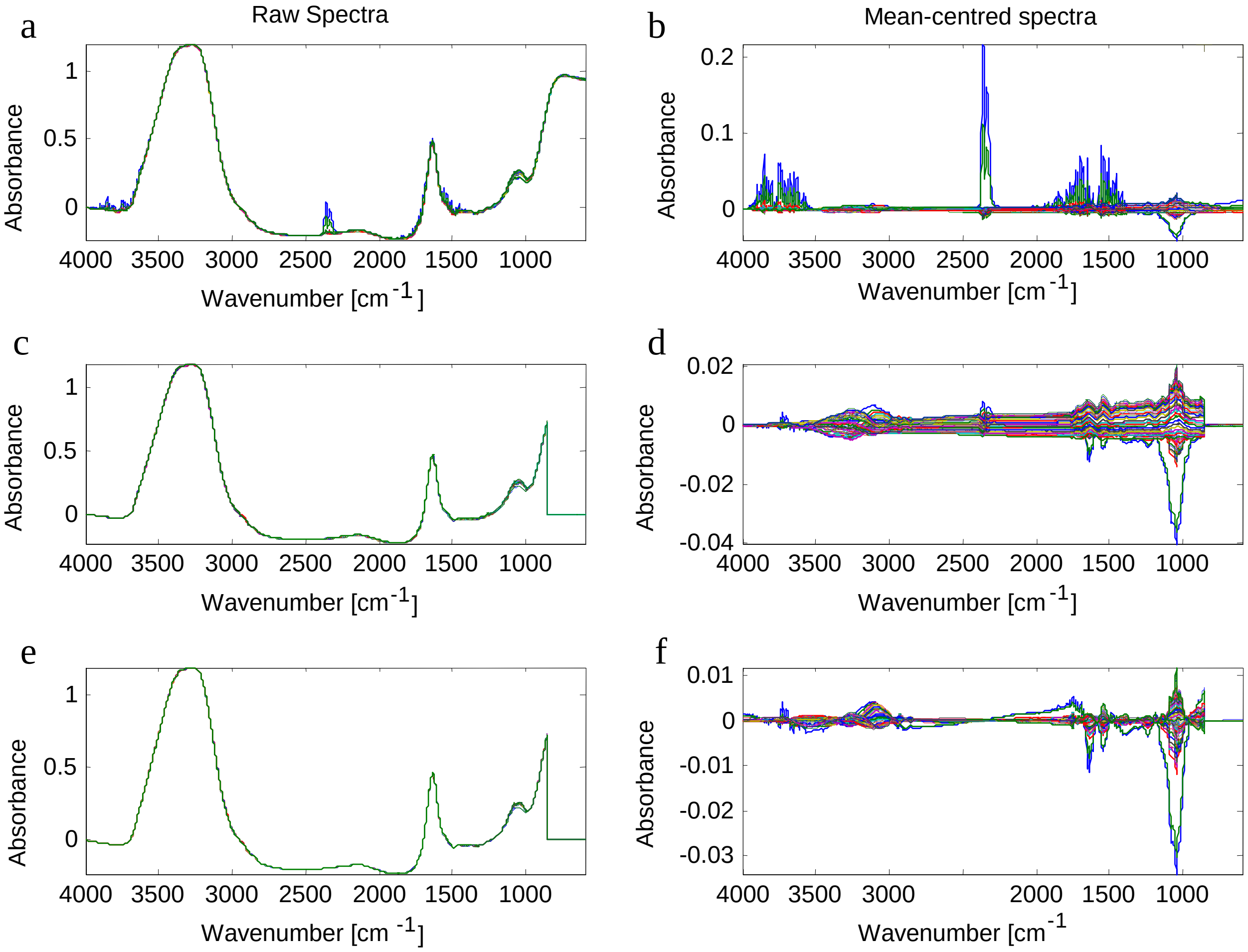


Fig. 9. *In situ* ATR-FTIR spectra of *C. albicans*. Top: input absorbance spectra. Middle: after correction for the primary and secondary water vapour and CO_2 spectra. Bottom: after subsequent interpolation under the main CO_2 peak, and EMSC to correct for light scattering etc. Left side (a,c,e): absorbance data. Right side (b,d,f): mean-centred absorbance data.

The wavenumber regions between 1582 and 1453 cm^{-1} and below 856 cm^{-1} were down-weighted in the EMSC parameter estimation in order to suppress the effects of chemical and instrument variations in the parameter estimation. For catching of the multiplicative light scattering effect and additive baseline variations, each spectrum was approximated by the mean spectrum from Fig. 9c plus a second-degree polynomial in the wavenumber domain. The estimated baseline effects were subtracted and the result divided by the estimated multiplicative light scattering parameter. The corrected spectra, which are shown in the bottom of Fig. 9 (e,f), display many regions with chemical variations. Each of the three data matrices displayed in Fig. 9b, 9d and 9f were submitted to PCA. The loading and score vectors are shown in the left and centre part of Fig. 10, while the two first PC scores are plotted against each other in the right part of the figure. The untreated absorbance data (row 1) give PCA loading vectors dominated by irrelevant gas contributions.

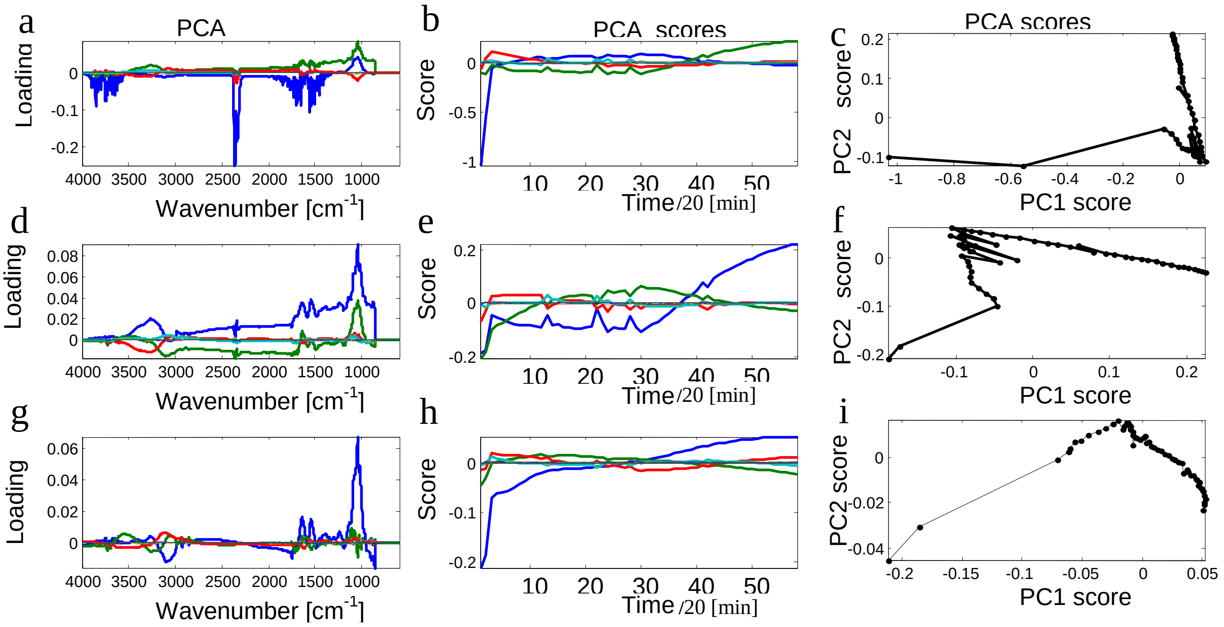


Fig. 10. Results from PCA of *in situ* ATR-FTIR spectra of *C. albicans*. Left side (a,d,g): PCA loadings vs. wavenumber. Middle (b,e,h): PCA scores vs. time. Right side (c,f,i): PCA score for PC1 (abscissa) vs. PC2 (ordinate). “Time=0” represent the start of the measurement series.

Hence, the obtained scores probably reflect the trivial purging of gases. After removal of the gas absorptions (row 2), their contributions to the loadings are greatly reduced, and the loadings now show peaks from the chemical constituents. Thus, the scores reflect chemical changes during the cell growth, but also physical changes, and the jumps observed in the score plots are probably related to some physical measurement variations.

	Input spectra	Gas-corrected (primary model spectra used)	Gas-corrected (primary +secondary model spectra used)
Experiment 1	5.242 (37.754)	0.295 (0.570)	-
Experiment 2	3.181 (9.748)	1.193 (2.783)	0.885 (1.515)
Experiment 3	1.058 (2.392)	-	0.805 (0.826)
Experiment 4	4.977 (21.989)	-	0.143 (0.832)

Table 2. Average water vapour indices (×10^3) for the data sets in Experiment 1-4 before and after subtraction of the estimated primary and secondary gas model contributions. The maximum index for each data set is shown in parenthesis.

The additional step for removal of remaining CO_2 and light scattering effects (row 3) causes the scores to show much smoother development over time. Thus, it seems that the physical variations have been removed in this step, leaving mainly chemical information, and the course of the scores may reflect two growth phases: an initial fast and a final slower phase.

The accomplishment of the gas-removal in all experiments is summarised by the water vapour indices shown in Table 2.

DISCUSSION

The above presented preprocessing tool is suitable for reducing the atmospheric contributions to ATR-FTIR spectra, such as to improve the analysis of the interesting variations in the samples. As seen from the water vapour indices in Table 2, and as shown above, the empirical gas model was able to effectively reduce the contaminating water vapour patterns, even in the spectra of *C. albicans*, which are very different from the calibration spectra. Interestingly, correction of the *C. albicans* spectra caused the largest decrease in the water vapour index (Table 2). Likewise, in Fig. 9 is seen a tremendous effect of removing the gas contributions from the *C. albicans* spectra. After the correction, the scores in the first three principal components are no longer influenced by the gases and it is possible that they could be used for modelling growth dynamics of *C. albicans*. Without removal of the gas absorptions, many more principal components would be needed in this modelling, and the interesting loadings would contain more noise. In addition, if purging rate and growth rate were the same, it could be hard to obtain chemical information about the samples from the loading plots, as these would contain also the gas signatures.

The subtraction of atmospheric absorptions in FTIR spectra is a necessity in protein secondary structure analyses that are based on the amide I band (1700-1600 cm^{-1}). In order to ensure correct amide I band assignments, the overlapping water vapour band needs to be almost completely eliminated prior to the analysis.[5] Thus, mathematical treatments for removal of the water vapour spectrum have been proposed, and in many cases, the algorithms are included in the spectroscopic software. The here proposed method is thought to improve the amide I analysis considerably, as it reduces the high-frequent bands in the amide I region of the second derivative water spectra.

Temperature variations in the measurement chamber could theoretically contribute to some variation in the gas spectra, as the populations of molecules occupying the higher rotational levels become increased on the expense of others with increasing temperature.[4] This causes the maximum of the two branches to move further away from the centre of the rotation-vibration band. In addition, temperature-dependent line broadening effects arise.[4] However, the small variations around room temperature that are commonly found between measurement days are believed to cause only minor temperature effects in the gas spectra. In addition, the calibration samples are expected to cover some temperature variations, due to the different temperatures of the ATR-cell. The performance of the preprocessing tool at very different temperatures was not tested.

Free Matlab (TM) software code for establishing and using model of “nuisance” contributions from water vapour and CO_2 in IR spectra can be downloaded at www.matforsk.no/specmod, where also the actual values of the gas component spectra obtained in the present paper can be found. The methods presented here are considered rather generic. But the gas component spectrum values are considered valid only for the present type of ATR-FTIR measurements in our Bruker instrument; before they can be used in another experimental setup, they may need substantial calibration transfer modification.

ACKNOWLEDGEMENT

Astrid Oust is thanked for providing the empty-instrument gas measurement time series. Isabelle Adt is grateful for financial support by the Conseil Régional de Champagne-Ardenne, Achim Kohler by the Aurora program of the Norwegian Research Council.

REFERENCES

1. K. Kargosha, M. Khanmohammadi, M. Sarokhani, F. Ansari, and M. Ghadiri. J. Pharmaceut. Biomed. **31**, 571 (2003).
2. R.M. Donlan, J.A. Piede, C.D. Heyes, L. Sanii, R. Murga, P. Edmonds, I. El-Sayed, M.A., and El-Sayed. Appl. Environ. Microbiol. **70**, 4980 (2004).
3. P.A. Suci, J.D. Vrany, and M.W. Mittelman. Biomaterials. **19**, 327 (1998).
4. H.M. Heise and H.W. Schrötter. Rotation-vibration spectra of gases. In: *Infrared and Raman spectroscopy: Methods and Applications,* Schrader B., Ed. (Wiley-VCH, Weinheim, 1995). Chap. 4.3, p.256.
5. S.E. Reid, D.J. Moffatt, and J.E. Baenziger. Spectrochim. Acta. A-M. **52**, 1347 (1996).
6. D. Helm, H. Labischinski, and D. Naumann. J. Microbiol. Meth. **14**, 127 (1991).
7. P.J. Hendra. Internet J. Vib. Spec. [www.ijvs.com] **5**, 2 (2001).
8. H. Martens and T. Næs. Multivariate Calibration. (John Wiley and Sons, Chichester, UK, 1989).
9. H. Martens, S.W. Bruun, I. Ad, G. Sockalingum, and Kohler, A. Correction of temperature and salt effects in water in FTIR biospectroscopy. Paper submitted. (2006).
10. H. Martens and E. Stark. J. Pharmaceut. Biomed. **9**, 625 (1991).
11. H. Martens, J.P. Nielsen, and S.B. Engelsen. Anal. Chem. **75**, 394 (2003).

Appendix III

Paper IV

Application of NIR and FTIR in characterisation of ligand-induced conformation changes in folate binding protein purified from bovine milk. Influence of buffer type and pH.

In collaboration with Jan Holm, Steen Ingemann Hansen and Susanne Jacobsen.

Application of NIR and FTIR in characterisation of ligand-induced conformation changes in folate binding protein purified from bovine milk. Influence of buffer type and pH.

Susanne W. Bruun[1], Jan Holm[2], Steen Ingemann Hansen[2], Susanne Jacobsen[1].

(1) Biochemistry and Nutrition Group, BioCentrum-DTU, Technical University of Denmark, Kgs.Lyngby, Denmark (2) Department of Clinical Chemistry, Hillerød Hospital, Hillerød, Denmark.

Submitted to Journal of Applied Spectroscopy, January 28, 2006.

SUMMARY

Fourier transform infrared (FTIR) and near-infrared (NIR) spectroscopy have been applied to detect structural alterations in folate binding protein (FBP) induced by ligation in different buffer types. The amide I region pointed to a β-sheet to α-helix transition upon ligation in acetate and phosphate buffers, and the formation of intermolecular β-sheet was indicated at pH 5.0, in agreement with a dimerisation of FBP taking place at this pH. The ligand-induced changes in the 2100-2300 nm NIR region were significant for FBP in acetate and phosphate buffers of pH 5.0, and the variations were interpreted as secondary structure changes, based on previous assignments of secondary structures to the combination bands in NIR. In case of acetate buffer, variations in the amide combination bands agreed with the amide I analysis, but for the other buffer types some discrepancies were found and explained by side chain contributions to the NIR, which could reflect the tertiary and quaternary structure differences.

NIR spectra of FBP at pH 7.4 and 5.0 revealed contradictory effects on the side chains, reflecting different polymerization events at the two pH-values, whereas the amide I region indicated similar changes at the two pH-values. Therefore, FTIR and NIR may complement each others.

INTRODUCTION

FTIR is a powerful and widely used technique for determining the secondary structure of proteins in aqueous solution. The characteristic amide I band between 1700 and 1600 cm^{-1} (in the mid-IR range) provides information on protein secondary structure owing to a sensitivity of the amide I frequency to the hydrogen bonding pattern and dipolar couplings in the protein backbone.[1-3] The empirical relation between different secondary structure types and absorptions at specific frequency intervals in the amide I band was discovered by Elliot and Ambrose in 1950.[4] Since then, FTIR has been applied in a number of qualitative and quantitative conformation studies. These have dealt with the denaturation/unfolding and aggregation of proteins induced by heat,[5-9] pH,[10] denaturants,[7] buffers[8] etc, and with protein conformational changes upon enzymatic cleavage[11], ligand binding, oligomerisation[9] etc. Also, a change in the secondary structure of human serum albumin (HSA) upon coordination of cis-platin to the protein backbone has been elucidated by FTIR[12].

In addition to the fundamental amide bands in the mid-infrared region (4000-500 cm^{-1}), proteins have their complementary but much weaker fingerprints in the near-infrared region (14000-4000 cm^{-1} equal to 714-2500 nm), in which overtones and combinations of the fundamentals appear. These broad and overlapping bands have rarely been used for protein structure analysis. However, they are sensitive to the interaction state of the different chemical groups, and a relation between protein secondary structure and overtone and combination bands has been established from several experiments.[13-15] For example, detailed information about the denaturation process of HSA and ovalbumin (OVA) has been obtained from the near-infrared (NIR) spectra.[16-17]

The purpose of the present study was to apply the two techniques; FTIR and NIR, to study ligand- and buffer-induced conformational changes in a protein and compare the information obtained by use of the two spectroscopic approaches, as they may offer complementary information.[18]

We study the high-affinity folate binding protein (FBP), which is present in most mammalian tissues and body fluids at nanomolar concentrations and seems to regulate homeostasis of folate in the body via its involvement in conservation and protection of folates against biological degradation as well as distribution, excretion and intracellular trafficking of folates.[19-22]

Direct and indirect evidence suggests that changes in the secondary structure of bovine milk FBP occurs after binding of folate. The CD-spectrum shows a decrease in antiparallel β-strands and increase in α-helix after ligation,[23] the data, however being incomplete due to the inability to measure the hydrophobic unligated (apo) FBP at near-neutral pH. At near-neutral pH, ligand binding enhanced the concentration-dependent aggregation of FBP[24,25] and caused a conversion of the hydrophobic aggregates into hydrophilic aggregates.[26,27] Ligand binding also induced dimerisation at pH 5.0, at which pH, the apo FBP exists as a cationic hydrophilic monomer.[24,27, 28]

Changes in the conformational structure of FBP subsequent to ligand binding could be of great physiological importance in the following respects: An altered structure of the ligated (holo) FBP might “wrap up” the ligand, and thereby protect it from biological degradation.[29] After ligation, the high-affinity folate receptor (FR) on the surface of the cell membrane is internalized via endocytosis and trafficks through endosomal compartments, and recycles back to the plasma membrane after release of ligand into acidic intracellular compartments.[30] The process initiating cycling/recycling of the ligated/unligated receptor is still obscure, but it is a well-known fact that ligand binding to many receptors is accompanied by conformation changes/dimerisation that initiate a sequence of biological processes. By analogy one could propose that ligation/unligation of the FR and the conformational change associated with that process is the initial event or signal that triggers its cycling/recycling.

The FR is a promising therapeutic target for antifolate drugs in tumor cells expressing high levels of the receptor. Detailed knowledge on how these drugs might affect the conformational structure of FR would thus be of great interest.

Dissociation of folate from FR in the intracellular acidic compartment occurs at a pH of 5.0, range 4.7-5.8[31] or markedly higher than that (pH 3.5) normally required for release of folate from FBP in vitro.[19-22] One could hypothesize that the organic acids and metabolites acetate and citrate in some mode (conformational change?) downregulate the affinity of FR for folate in the intracellular compartment, since no binding of folate to physiological concentrations of FBP occurs at pH 5.0 in acetate or citrate buffers.[28]

We analyze the buffer effects as well as the effect of ligation by means of the spectroscopic techniques. Very limited data is available concerning the structure of FBP, wherefore this study contributes with some new information to this field as well as insight into the performance of NIR in protein structural analysis.

Due to the weak protein signal compared to the water signal in NIR, the chemometric techniques have commonly been applied for analyzing NIR spectra of proteins in solution. Data analysis methods have included e.g. two-dimensional correlation spectroscopy[16,17,32-34] and Principal Component Analysis (PCA).[33,35] PCA is a factor based methods that takes advantage of the intercorrelations between spectral variables to reduce the dimensionality of the data and allows the extraction of information from the complex data sets.[36]

Here we use PCA supplemented with 2nd derivative and difference spectroscopy.

MATERIALS AND METHODS

The radiochemical [^{14}C] folate (pteroylglutamate) with a specific activity of 52.4 Ci/mol, and [^{3}H] folate with a specific activity of 26-45 Ci/mol were obtained from Amersham International Ltd., Amersham, U.K. Folate and standard proteins were supplied by Sigma.

FBP purification: A large scale purification of FBP from cow's whey powder was performed as previously described[26] by a combination of cation exchange - and methotrexate affinity chromatography. The purified protein was characterized with regard to primary and secondary structure as well as ligand binding characteristics.[24-27,37,38] All FBP solutions were dialyzed against 0.2 M acetate buffer, pH 3.5 at 4 °C to remove endogenous folate.

Binding study: Equilibrium dialysis was performed as described previously for periods of 20 h at 37 °C and pH 5.0 in different buffer types[25] (Fig. 3) with FBP in the internal (1000 µl) and radioligand in the external solution (200 ml).

Sample preparation for spectroscopic measurements: FBP solutions were prepared by dialysis of the stock solution against acetate-, citrate-, formate- and phosphate buffers, all of pH 5.0 and also phosphate buffer of pH 7.4. The final protein concentrations in the buffers varied from 8 to 19 mg/ml (265-630 µM binding activity).

At pH 5.0, an adequate amount of folate in solid form was added to an aliquot of each FBP solution, and after mixing for 1 min and resting for ~1 hour, all samples with and without folate were centrifuged for 1 min at 14000 g. Precipitated folate was discarded. Solubility of folate at pH 5.0, 25°C: 200 µM. At pH 7.4, removal of unbound and dissolved folate from samples was performed by dialysis against the pure buffer for 2 days. Solubility of folate at pH 7.4, 25°C: > 100 mM.

Standard proteins of known secondary structures bovine serum albumin (BSA), β-lactoglobulin (BLG), OVA, lysozyme (LYS) and casein (CAS) were prepared at 10 mg/ml concentrations in phosphate buffer of pH 7.4. BSA and BLG solutions were also prepared in acetate pH 5.0 and phosphate pH 5.0, and all solutions were treated with folate as described above and measured as calibration and control samples.

Spectroscopic measurements

The clear protein solutions were measured in a 1 mm cuvette in transmission mode on a Perkin-Elmer FT-NIR Spectrum One NTS spectrometer equipped with a deuterated triglycine sulfate (DTGS) detector. The resolution was 16 cm^{-1} and 100 scans were co-added. The data interval was 1.67 nm. Absorbance spectra were calculated by subtraction of the background spectrum (empty cuvette). Each sample was measured in replicates and on different days to include the measurement- and day-to-day- variations.

FTIR transmission spectra were acquired on an Bomen MB-100 series FTIR spectrometer (Bomen, Quebec, Canada) equipped with a DTGS detector and continuously purged with dry air. A BioCell (BioTools, Wauconda, IL, USA) with CaF windows and a predrilled 6 µm depression was used.

The resolution was 2 cm^{-1} and 256 scans were co-added. The empty cell was used as background.

Spectral preprocessing and data analysis

NIR spectra were preprocessed and analyzed by PCA in Unscrambler 9.2 (Camo, Oslo).

The NIR spectra were truncated to the range 2100-2300 nm, since this region is rich in protein information and avoids the large absorption bands from water. Extended Multiplicative Scatter Correction (EMSC)[39] was applied to this range, with the purpose of removing the additive and multiplicative effects etc. that result from a variable path length and light scattering phenomena. In the EMSC model, each spectrum $\mathbf{z}_i$ is expressed as a modification of the ideal chemical spectrum $\mathbf{z}_{i,chem}$ (Eq. 1).

(Eq. 1) $$\mathbf{z}_i = a_i\mathbf{1}+b_i\mathbf{z}_{i,chem} +d_i\boldsymbol{\lambda} + e_i\boldsymbol{\lambda}^2+\varepsilon_i.$$

The a_i and b_i coefficients represent the additive and multiplicative effects, respectively, and linear and quadratic wavelength effects are included by the terms $d_i\boldsymbol{\lambda}$ and $e_i\boldsymbol{\lambda}^2$.

The ideal chemical spectrum is in the standard EMSC approach taken as a variance around the mean spectrum $\mathbf{m}$ as expressed in Eq. 2.

(Eq. 2) $$b_i\mathbf{z}_{i,chem}= \mathbf{m}+\boldsymbol{\delta}.$$

Thus, the EMSC coefficients are estimated from the whole data set and used for correction of each spectrum (Eq. 3).

(Eq. 3) $$\mathbf{z}_{i,corr}=(\mathbf{z}_i-a_i\mathbf{1}-d_i\boldsymbol{\lambda}-e_i\boldsymbol{\lambda}^2)/b_i.$$

After EMSC correction of the NIR spectra, the matching buffer spectrum was subtracted from each sample spectrum, and EMSC was applied once more. By this preprocessing method, the protein concentration differences were eliminated from the data set. Finally 2nd derivative transformation was carried out by use of the Savitzky-Golay algorithm (9 or 13 data point smoothing). The inverse spectra were calculated by multiplication by –1. Difference spectra (ligated-unligated) were calculated from the non-transformed spectra.

The buffer subtractions from transmission spectra were carried out in Grams software, and a flat baseline in the 2000-1800 cm^{-1} range was used as criteria. The atmospheric water vapor spectrum was obtained before the experiment and subtracted from each spectrum. The 2nd derivative spectra were calculated by use of the Savitzky-Golay algorithm (13-17 data points smoothing). The spectra were inverted, and baseline correction and mean normalization was applied to the 1700-1600 cm^{-1} region. Difference spectra (ligated-unligated) were calculated from these spectra.

RESULTS AND DISCUSSION

COMPARISONS OF UNLIGATED FBP IN DIFFERENT BUFFER TYPES

IR-results

We examined the influence of three buffer types (phosphate pH 7.4, phosphate pH 5.0, and acetate pH 5.0) on the amide I spectra of FBP. The preprocessed 2nd derivative spectra are seen in Fig. 1.

The contributions from α-helix and β-sheet to the amide I band appear at 1660-1650 cm^{-1} and 1640-1625 cm^{-1}, respectively, and the contours of the amide I bands in Fig. 1 reflect a protein with no predominant single structure. This is in agreement with CD measurements (however, only performed at pH 5.0) by Kaarsholm et al.,[23] who found 22 % α-helix, 30 % β-sheet, 31 % random, 17 % turns in FBP.

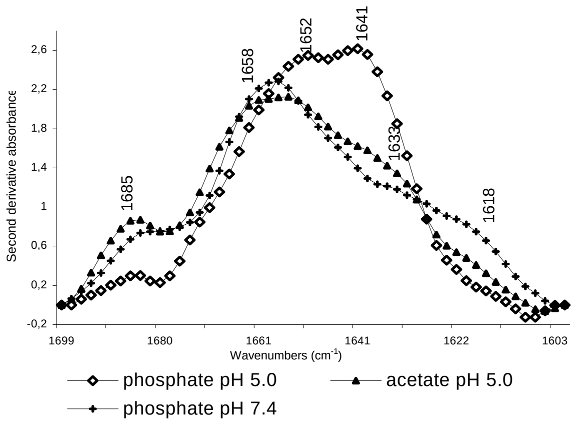


Fig. 1. 2nd derivative amide I spectra of apo FBP in different buffer types.

FBP in phosphate buffer of pH 5.0 exhibits an amide I band shape that is highly different from that in the two other buffer types (acetate pH 5.0 and phosphate pH 7.4). 1) The high absorbance at 1641 cm^{-1} implies either an increased amount of intramolecular β-sheet or random coil (1645 cm^{-1}) or both. 2) A band that can be ascribed to α-helix has a maximum at 1652 cm^{-1} instead of 1658 cm^{-1} as seen in the other buffer types. This could suggest that more solvated α-helices exist in phosphate buffer of pH 5.0, as solvation causes helices to appear at lower frequencies[40] or that α-helices have a different length in this buffer type.[41] 3) The decrease in the high-frequency region could also be related to a lower content of β-turns in phosphate buffer of pH 5.0, as β-turns may cause absorptions in the region 1660-1690 cm^{-1}. [42]

At low frequencies, a weak band that is usually ascribed to intermolecular β-sheet (1618 cm^{-1})[9] is apparent. For the two buffers of pH 5.0, the low absorption at 1618 cm^{-1} indicates less intermolecular β-sheet than in the phosphate buffer of pH 7.4. As intermolecular β-sheet is frequently found to participate in oligomerisations,[43] this observation is in agreement with apo FBP being monomeric at pH 5.0 while forming polymers at pH 7.4.[24,25]

NIR-results

Fig. 2A shows the buffer-subtracted 2nd derivative NIR spectra and compares FBP in five different buffer types (acetate-, formate- and citrate buffers of pH 5.0 and phosphate buffers of pH 5.0 and pH 7.4).

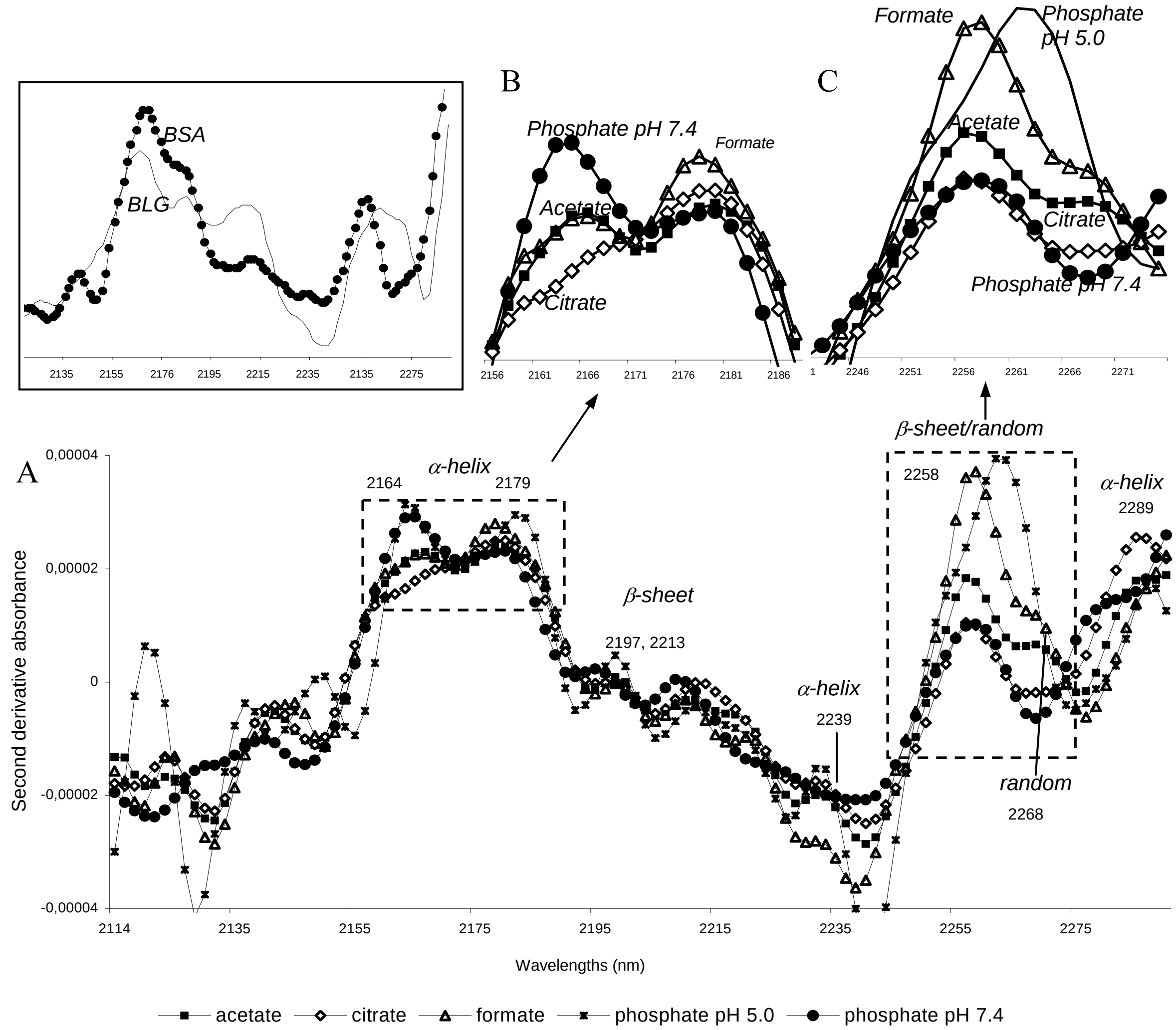


Fig. 2. NIR spectra of apo FBP in different buffer types. A) 2nd derivative spectra in the 2100-2300 nm region. B) Enlargement of region 1, showing the amide combination band at 2170 nm. Phosphate buffer of pH 5.0 is not included. C) Enlargement of region 2, showing the side chain combination band at 2261 nm. The insert in the upper left corner shows the spectra of BSA and BLG in phosphate buffer of pH 7.4.

The absorption bands seen below 2230 nm result from combinations of the amide modes i.e. the coupled vibrations in the protein backbone. The band at 2170 nm is usually ascribed to the amide B/II combination[32,44] (basically N-H stretching combined with N-H bending), but several different amide combinations bands likely appear in the region 2100-2230 nm. The 2nd derivative transformation allows for the observation of several sub bands in the region 2120-2230 nm, as seen in Fig. 2A, whereas a single absorption band is seen in the raw spectra (not shown). In this region, Miyazawa et al.[14] observed for different globular proteins six sub bands at 2141, 2168, 2186, 2200, 2209 and 2213 nm. In Fig. 2A, sub bands at similar positions can be seen. However, the regions 2114-2150 nm and 2190-2220 nm are much influenced by noise, and the band positions can not be precisely established. The absorption bands above 2230 nm result from vibrations in the amino acid side chains and are mainly the combinations of C-H stretching and C-H bending vibrations, but also the combination of O-H stretching and O-H bending and the combinations of C-H stretching and amide III are suggested to appear at 2255 nm and 2290 nm, respectively.[33,34]

In the literature, wavelength regions characteristic of the different secondary structures have been identified, as indicated in Fig. 2A and Table 1. For example, the intensity of the NIR band at 2170-2180 nm has been found to receive the highest contribution from α-helix compared to β-sheet and random structure.[15] This is also illustrated by the spectra of BSA (α-helix protein) and BLG (β-sheet protein) in the insert in Fig. 2. The spectrum of apo FBP was influenced by buffer type as seen in Fig. 2A-C. In the α-helix region around 2170 nm is seen a variation in the A_{2164nm}/A_{2179nm} ratio (Fig. 2B). The high-wavelength band has often been ascribed to the primary amide groups in Gln and Asn. Yuan et al.[35] observed a splitting of the band at 2169 nm into two bands at 2164 nm and 2176 nm upon denaturation of BSA. Likewise, Miyazawa et al.[14] observed that, while a 2186 nm band was only observed for some of the proteins, the 2168 nm band was common to all the measured proteins. For FBP, the low-frequency band at 2164 nm shows most variation in the different buffers, with citrate buffer showing the weakest absorption and the phosphate buffer of pH 7.4 the highest absorption (the phosphate buffer of pH 5.0 could not be compared due to the high replicate variability in this region).

Structure	Characteristic wavelengths [nm]
α-helix	2172-2186, 2239, 2289
β-sheet	2200-2213, 2264
random	2265

Table 1. Assignments in the 2100-2300 nm NIR region to different secondary structures.[13,14]

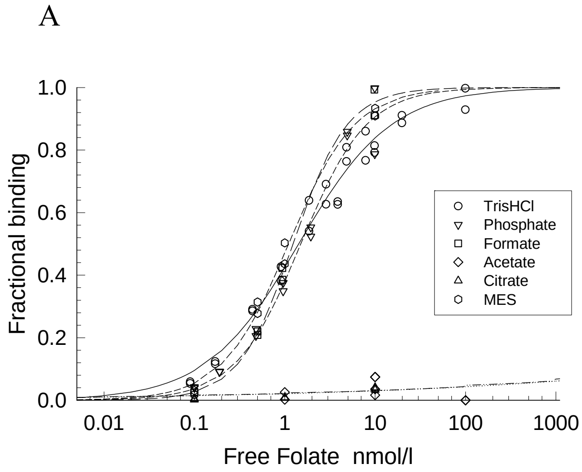


Fig. 3. Saturation curves for high-affinity binding of radiolabeled folate to FBP dissolved in different buffer solutions at pH 5.0. FBP concentrations were 6 nM and buffer concentrations were 0.2 M (phosphate, formate, acetate, citrate), 0.17 M (Tris) or 0.1 M (MES).

Another large and remarkable variation in the

FBP spectra, resulting from the different buffer types, is found in the side chain absorption at ~2260 nm (Fig. 2C). Formate- and phosphate buffer samples of pH 5.0 show a high absorbance in this region compared to the other two buffers of pH 5.0 (acetate and citrate) as well as phosphate buffer of pH 7.4. The large peak is found at 2258 nm for the formate buffer samples and at 2263 nm for the phosphate buffer samples. The increased absorbance could result from a high content of β-sheet (2264 nm) and/or random coil (2265 nm) in the two buffers, according to Table 1. This would be in agreement with the high absorption at 1641 cm^{-1} (in the amide I band) for FBP in phosphate buffer of pH 5.0 (Fig.1). The indication that formate and phosphate buffers of pH 5.0 have a different influence than acetate and citrate buffers on FBP is interesting in view of binding studies at pH 5.0, demonstrating binding of folate in formate, and phosphate buffers (and also in Tris and MES buffers) but not in acetate and citrate buffers at low FBP concentrations (Fig. 3). Furthermore, 50 % compared to 14 % of folate bound to FBP was retained after wash-out at pH 5.0 in Tris HCl and acetate buffers, respectively (unpublished results). At high FBP concentrations, as used in the present NIR/FTIR study, the binding can take place also in acetate buffer.[28] Still, it is possible that the buffer effects are reflected in the spectra even at the high FBP concentrations. The applied anions are kosmotropes (water-structuring ions) of varying strength, and they may stabilize the native protein conformation and the hydrophobic protein-protein interactions to different degrees.[45] That no simple explanation of the observed buffer effects could be given probably reflects the complexity of these anion effects. The high absorbance at 2260 nm in case of formate and phosphate buffer is not accompanied by a high absorbance in other β-sheet regions (2200-2213 nm), wherefore, it is more likely that the increase at 2260 nm reflects random coil and not β-sheet. However, it should be stressed that the variations in the side chain combination bands are not exclusively the results of different secondary structures but also of other conditions that influence the micro-environments of the side chains and their interactions.[17,34,46] Thus, for example the solvation and oligomerisation state of FBP in the various buffers may also account for the differences in the side chain absorptions. A demonstration of changes in the side chain absorptions, occurring despite of no significant secondary structure changes, was obtained from spectra of BSA, for which the 2260 nm band was increased in acetate buffer of pH 5.0 compared to phosphate buffer of pH 5.0. Sefara et al.[46] observed uncoupled changes in the side chain absorptions at 2252, 2257 and 2262 nm upon unfolding of BLG in bromoethanol.

LIGATION IN DIFFERENT BUFFER TYPES

IR results

Ligand-induced conformation changes of FBP were analyzed by transmission FTIR in three buffers (acetate buffer of pH 5.0 and phosphate buffers of pH 5.0 and 7.4). We investigated the difference spectra instead of performing curve-fitting analyses on the amide I bands.

At pH 5.0, the amide I bands from FTIR-measurements of FBP in acetate and phosphate buffer are shown in Fig. 4A-B. The spectral changes are most prominent for phosphate buffer of pH 5.0, but for both buffer types, the variations indicate a decrease of intramolecular β-sheets and an increase of α-helix upon ligation, as shown by the negative peaks at $\sim$1640 cm^{-1} and the positive peaks at 1660 cm^{-1} in the difference spectra. These findings are in agreement with CD studies[23] at pH 5.0, in which a decrease from 25 to 15 % of antiparallel β-sheet and a small increase in α-helix and turn structures was observed. The decrease in the region 1650-1630 cm^{-1} may also stem from a decrease of random coil content, although the CD studies[23] suggested a minor increase of random coil upon ligation.

A small increase of the extreme high- and low-frequency bands (1685 and 1620/1610 cm^{-1}) for both buffer types can be ascribed to an increase of intermolecular β-sheets or extended β-strands, in

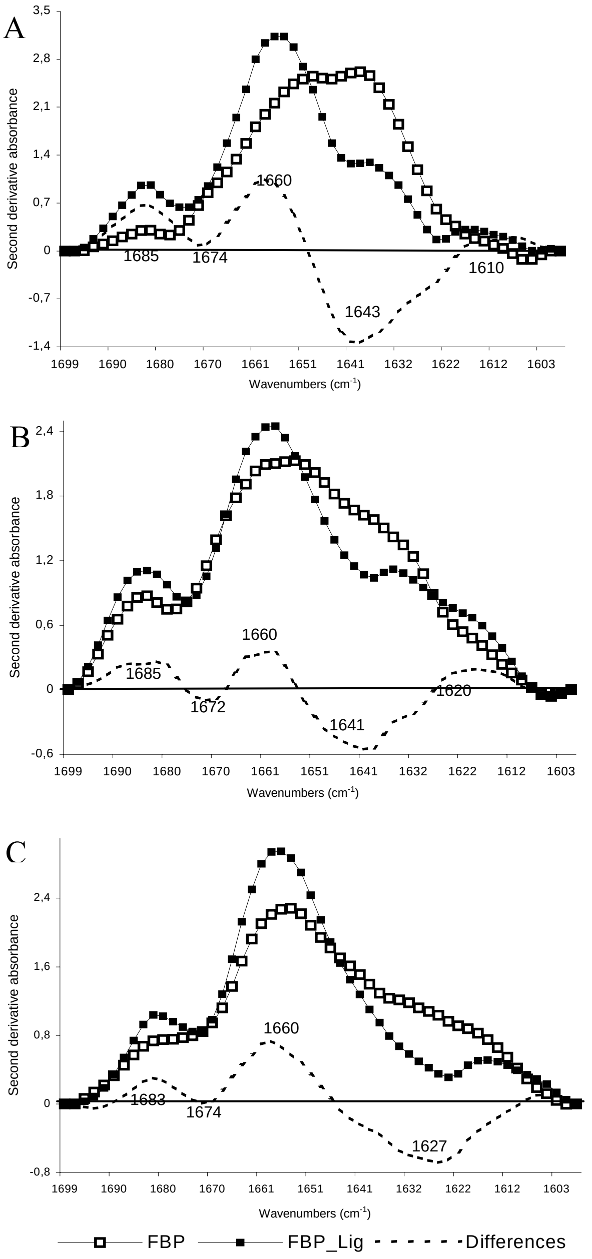


Fig. 4. 2nd derivative amide I spectra of holo and apo FBP in different buffers. The difference spectra (ligated-unligated) are shown as well. A) Acetate buffer of pH 5.0. B) Phosphate buffer og pH 5.0. C) Phosphate buffer of pH 7.4.

agreement with the dimerisation occurring upon ligation at pH 5.0.[24,28] A high-frequency component usually accompanies the low-frequency component from intermolecular β-sheets.[1]

The amide I variations induced by ligation of FBP in phosphate buffer of pH 7.4 were studied, and the 2nd derivative spectra and the difference spectra are shown in Fig. 4C. The changes are quite similar to those seen at pH 5.0 (Fig. 4A-B), although the negative peak has its maximum shifted to a lower frequency (1627 cm^{-1}) than at pH 5.0 (1641 cm^{-1}). The changes are consistent with a transition from intramolecular β-sheet to α-helix. The bands, ascribed to intermolecular β-sheet, are not increased as seen at pH 5.0, and this can be explained by the fact that polymers already exist for the apo FBPs at pH 7.4. Ligation changes the polymers to a more hydrophilic conformation, although the polymerization is somewhat enhanced.[24,25,27] The hydrophobicity of FBP in phosphate buffer of pH 7.4 was reflected in attenuated total reflection (ATR)-FTIR spectra (data not shown), which showed high amounts of intermolecular β-sheet, indicative of protein adsorbed to the ATR-crystal.

The amide I variation introduced by folate addition to BSA was small and non-significant (data not shown). As albumin binds folate in a low-affinity mode,[47] BSA acts as a control and reveals that the changes in the FBP spectra are related to specific binding of folate and not resulting from folate absorptions.

Condition	**A_{1660}/A_{1640}**	**A_{1615}/A_{1640}**
	(~α-helix/β-sheet)	**(~inter-/intra- β-sheet)**
Acetate pH 5.0, apo	**1.33**	**0.206**
Acetate pH 5.0, holo	2.35	0.478
PBS pH 5.0, apo	**0.78**	**0.056**
PBS pH 5.0, holo	2.35	0.221
PBS pH 7.4, apo	**1.7**	**0.507**
PBS pH 7.4, holo	3.1	0.521

Table 2. Relative intensities in the amide I spectra of apo and holo FBP in different buffer types.

Results from the transmission measurements are summarized in Table 3. The α-helix/β-sheet ratio is particularly low for apo FBP in phosphate buffer of pH 5.0. Ligation increases the ratio to similar values in phosphate and acetate buffer of pH 5.0, and to an even higher value at pH 7.4.

NIR results

PCA was applied to the NIR spectra with the purpose of getting an overview of the spectral differences between ligated and unligated forms of FBP in different buffer types.

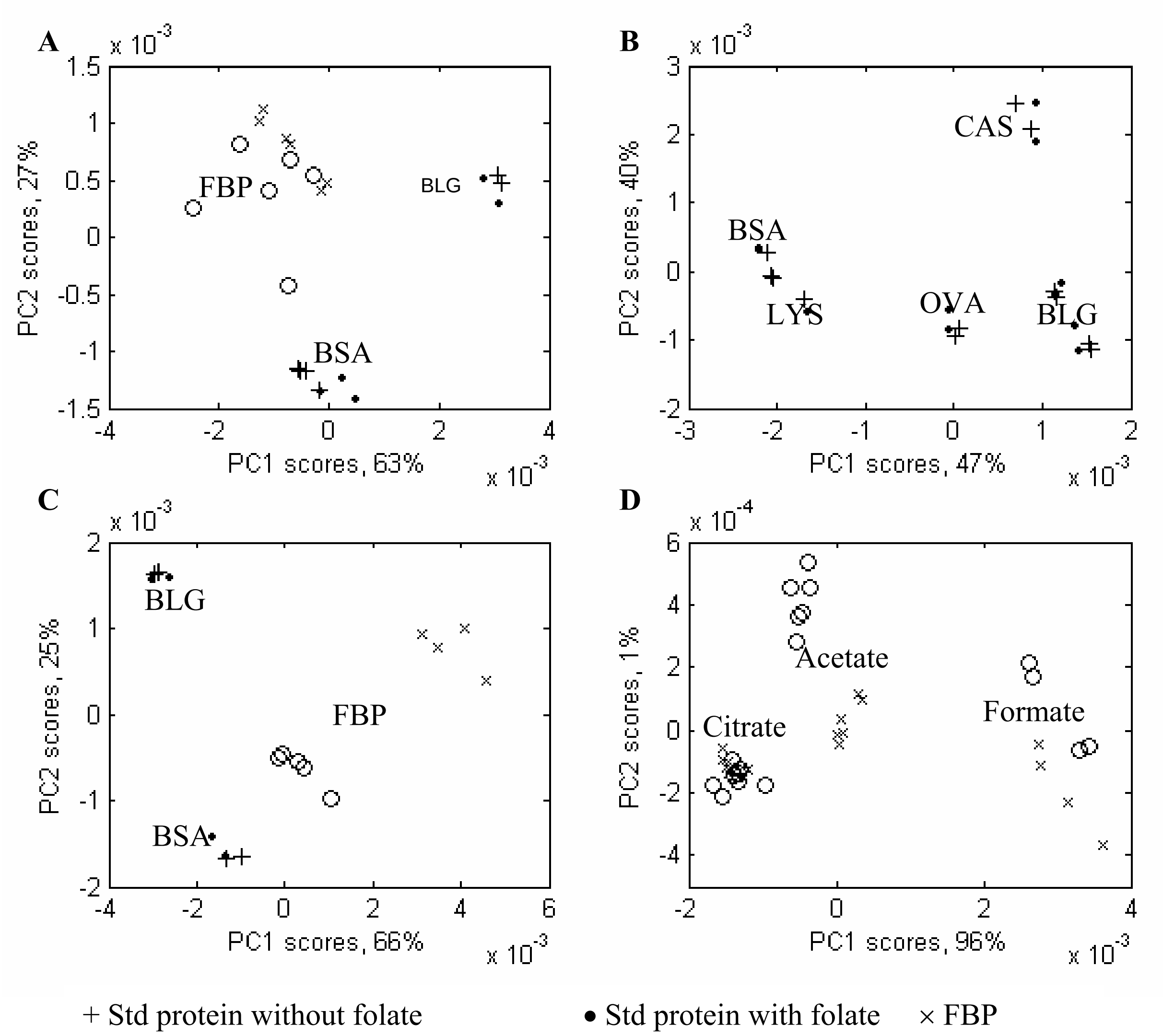


Fig. 5. PCA score plots (PC1 vs. PC2), showing the ability of NIR to discriminate between holo and apo FBP in different buffer types. Four PCA analyses based on the preprocessed NIR spectra (2100-2300 nm) were carried out. A) FBP in phosphate buffer of pH 7.4. B) Standard proteins with and without folate in phosphate buffer of pH 7.4. C) FBP in phosphate buffer of pH 5.0. D) FBP in different buffer types of pH 5.0. BSA and BLG with and without folate were included in most analyses for comparison. The percentages of explained variance accounted for in the different PCs are shown.

Several PCA analyses including different sub sets of the samples were carried out, and the resulting score plots (PC1 vs. PC2) are shown in Fig. 5. These plots reveal the main variations among the spectra.

Fig. 5A reveals a poor discrimination between holo and apo FBP in phosphate buffer of pH 7.4, as a high replicate-variability exists for the ligated samples. Similarly, folation of the standard proteins is not accounted for in any PCs (Fig. 5B). The distinction between holo and apo FBP is much clearer in phosphate buffer of pH 5.0 (Fig. 5C). For three other buffers of pH 5.0 (citrate, acetate and formate), the largest spectral difference between holo and apo FBP was observed in case of acetate buffer (Fig. 5D). Thus, while PCA results show only minor spectral changes of FBP in citrate and formate buffers of pH 5.0, spectral variations upon ligation are more significant in acetate and phosphate buffers of pH 5.0. This is not in agreement with the binding study (Fig. 3), which showed no binding in acetate and citrate buffer, but binding in phosphate and formate buffers. However, the much higher concentrations used here may cause a very different binding mechanism.

The information in the NIR spectra was further examined by use of 2nd derivative spectra. Since this calculation amplifies the noise, also the difference spectra (calculated from the non-transformed spectra) were considered. The difference spectra revealed no significant spectral variations in formate- and citrate buffers upon ligation (not shown). The small changes agree with results from the PCA, which showed that a discrimination was possible only in PC3 (data not shown). This is a spectral variance accounting for less than 0.5 % of the total variance. The small changes in citrate buffer can be ascribed to a rearrangement of the protein that does not involve significant secondary structure changes but perhaps small perturbations of the side chains.

Fig. 6A shows the buffer-subtracted 2nd derivative NIR spectra of holo and apo FBP in acetate buffer of pH 5.0. Spectral differences induced by the ligation are indicated by the arrows. The difference spectrum in Fig. 6B also reveals significant spectral differences between holo and apo FBP throughout the region from 2100 to 2300 nm, while addition of folate to BSA only causes a minor not significant change at ~2260 nm. Therefore, these variations in the FBP spectra are most likely to be ascribed to perturbation of the FBP structure rather than to folate absorptions. The ligation of FBP in acetate increases absorption at 2177 nm (α-helix), 2239 nm (α-helix), and 2289 nm (α-helix) and decreases absorption at 2151 nm (?), and 2264 nm (β-sheet and random). Thus, the NIR spectra support the formation of α-helix at the expense of β-sheet and/or random structures upon ligation of FBP in acetate.

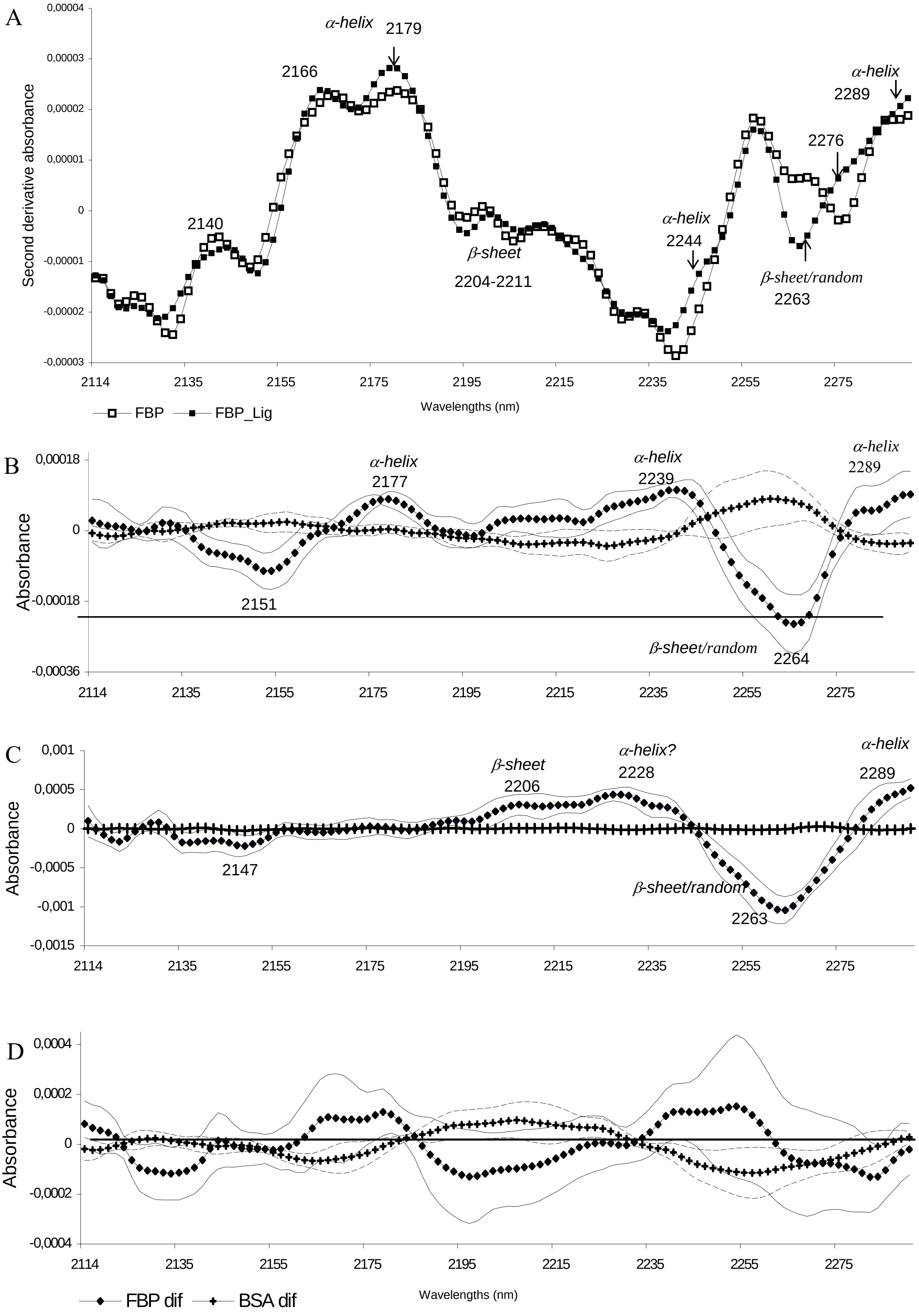

A
0,00004
0,00003
0,00002
0,00001
0
-0,00001
-0,00002
-0,00003
Second derivative absorbance
α-helix
2179
2166
2140
β-sheet
2204-2211
α-helix
2244
β-sheet/random
2263
2276
α-helix
2289
2114
2135
2155
2175
2195
2215
2235
2255
2275
Wavelengths (nm)
FBP
FBP_Lig
B
0,00018
0
-0,00018
-0,00036
Absorbance
α-helix
2177
2151
α-helix
2239
β-sheet/random
2264
α-helix
2289
2114
2135
2155
2175
2195
2215
2235
2255
2275
C
0,001
0,0005
0
-0,0005
-0,001
-0,0015
Absorbance
2147
β-sheet
2206
α-helix?
2228
β-sheet/random?
2263
α-helix
2289
2114
2135
2155
2175
2195
2215
2235
2255
2275
D
0,0004
0,0002
0
-0,0002
-0,0004
Absorbance
2114
2135
2155
2175
2195
2215
2235
2255
2275
Wavelengths (nm)
FBP dif
BSA dif

Fig. 6: NIR spectral differences in the 2100-2300 nm range between holo and apo FBP in different buffer types. Difference spectra (ligated-unligated) are shown for all buffers with standard deviations. The arrows point out regions where ligation induces an increase (downward pointing arrows) or a decrease (upward pointing arrows) in the FBP spectrum. BSA difference spectra are included for comparison. A) 2nd derivative of holo and apo FBP in acetate buffer of pH 5.0. B) Difference spectra for FBP and BSA in acetate buffer of pH 5.0. C) Difference spectra for FBP and BSA in phosphate buffer of pH 5.0. D) Difference spectra of FBP and BSA in phosphate buffer of pH 7.4.

The primary and secondary structures of FBP and the riboflavin-binding protein are closely related as evidenced by previous studies.[48,49] Release of riboflavin from the latter protein at low pH involves opening of a preexisting aromatic-rich cleft and increases molecular volume, protein surface and surface hydration.[50] This leads to segmental motion of aromatic side chains, most likely belonging to Trp and Tyr that participate in ligand binding.[49,50] Wasylevski et al.[51] saw no secondary structure changes upon riboflavin binding, and they suggested changed domain interactions or a ligand domain structure reorganization to take place instead. In a similar way, ligand-binding to FBP and the subsequent dimerisation could influence the solvation state and the packing of the side chains. As previously described, this could be the cause of the decreased intensity at 2260 nm.

The 2nd derivative spectra of FBP in phosphate pH 5.0 seemed more influenced by noise, especially in the low-wavelength region, and this is ascribed to a rather low protein concentration. Again, the difference spectrum in Fig. 6C reveals significant spectral changes between holo and apo FBP, whereas addition of folate to BSA causes no spectral changes. As being the case in acetate buffer of pH 5.0, there was a marked decrease in the absorbance band around 2263 nm (β-sheet/random?), a marked increase of the band at 2289 nm (α-helix?), and a decrease in the band at 2150 nm upon ligation. The increase at 2228 nm could also result from increased α-helix absorptions. By contrast to acetate buffer, there was no increase at 2177 nm (α-helix), and the absorbance at 2206 nm (β-sheet) was increased. Therefore, although the FTIR analyses show similar changes in phosphate and acetate buffers, the NIR analysis indicates that the changes in phosphate buffer of pH 5.0 are somewhat different from the β-sheet/random coil to α-helix transformation seen in acetate buffer. The discrepancy of the two methods possibly results from the contribution of side chain absorptions to the NIR region. The band most affected by the ligation in phosphate buffer of pH 5.0 was the one at 2260 nm. The decrease of this band upon ligation caused the disappearance of the large deviation seen in Fig. 2C between buffer types.

The NIR difference spectra of FBP in phosphate buffer of pH 7.4 are shown in Fig. 6D. The large replicate variability in the FBP spectra made it impossible to identify any spectral change that with certainty resulted from ligation of FBP. However, the indicated increase at 2167-2177 nm (α-helix) is similar to that seen in acetate buffer (pH 5.0) and supports the observations from the amide I spectra. A different trend in the combination band region compared to the other buffers could possibly stem from the different oligomerisation states of FBP at the different pHs, since the monomer to dimer-transition upon ligation only occurs at pH 5.0.[24,28]

An overview of the ligation phenomena in phosphate buffers of pH 5.0 and pH 7.4 is seen from the score plot in Fig. 7. The PCA was again based on the NIR region 2100-2300 nm. The ligation of FBP at pH 7.4 causes only small variations, whereas the ligation at pH 5.0 gives rise to groupings in the score plot. At pH 5.0, apo FBP is placed in the upper part of the score plot, whereas after ligation samples have moved to the lower part. This change could reflect a β-sheet to α-helix transition, when comparing to the included standard proteins (see Table 3), and it results in a close similarity between FBP spectra at the two pH values. A higher similarity between secondary structures after ligation accords with the FTIR results presented in Table 2. Also in the PC1 direction (related to the variation in the 2260 nm peak) a higher similarity between FBP at the two pH values is seen after ligation, as was previously described.

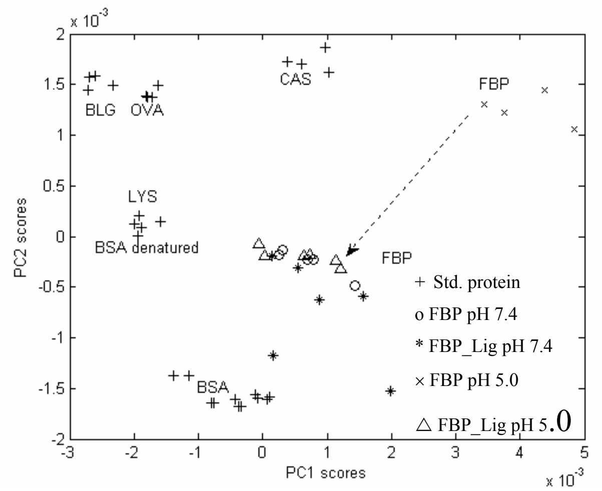


Fig. 7. PCA results, showing the effects of pH and ligation on the NIR spectra. Score plot (PC1 vs. PC2) shows differences and similarities of ligated and unligated FBP in phosphate buffers of pH 7.4 and pH 5.0.

Protein	α-helix (%)	random (%)	β-sheet (%)
Bovine serum albumin (BSA)	66	31	03
Lysozyme (LYS)	46	35	19
Ovalbumin (OVA)	25	49	26
β-Lactoglobulin (BLG)	06	48	46
Casein (CAS)	7-10	56-70	20-37

Table 3. Secondary structure contents of standard proteins.[13]

CONCLUSION

The ligand-induced structure changes in FBP in different buffer types have been studied by use of FTIR and NIR. We observed significant spectral changes in the NIR region upon ligation in acetate and phosphate buffers of pH 5.0, which could be interpreted as a decrease of β-sheet/random coil content in accordance with the FTIR and previous CD studies.[23] In addition, FTIR showed an increase in α-helix content and a higher content of intermolecular β-sheets after ligation in both buffer types. The latter finding is consistent with the ligand-induced dimerisation of FBP observed at pH 5.0.[24,28] With NIR, the increase in α-helix content was only observed in acetate buffer. This discrepancy is possibly due to some absorptions associated with the amino acid side chains that dominate the spectral variations in the NIR region upon ligation.

The buffer effects observed at low FBP concentrations (Fig. 3),[28] i.e. no binding of folate at pH 5.0 in acetate and citrate buffers but binding in phosphate and formate buffers, were reflected as increased NIR side chain absorptions at 2260 nm for FBP in phosphate and formate buffers. The distinctive side chain absorption for apo FBP in phosphate buffer of pH 5.0 was associated with a particular secondary structure, as seen from FTIR. However, the side chain absorptions in NIR may also reflect the tertiary and quaternary structure of FBP.

At pH 7.4, FTIR studies of FBP in phosphate buffer showed a transition from β-sheet to α-helix upon ligation similar to that at pH 5.0. However, with NIR, we observed a great difference between pH 5.0 and 7.4 (phosphate buffer) as to spectral changes of the side chains after ligation, and this could be related to the different polymerization events at the two pH values.

Therefore, we find that the two spectroscopic techniques complement each others. In the NIR, changes in the side chain absorptions reflect protein tertiary and quaternary changes, whereas secondary structure changes may give less significant spectral alterations. The interpretation of the NIR spectral changes demands the comparison to a more well-established method for secondary structure analysis, such as FTIR. The advantage of NIR over FTIR is that problems experienced in FTIR analyses (water vapor absorptions and adsorption of protein to the cuvette) are less serious in NIR analyses. In addition, the NIR has the ability to be applied as a non-invasive method in on line control of e.g. pharmaceutical products.

ACKNOWLEDGEMENT

The authors are grateful to M. Van der Weert for technical assistance in the FTIR measurements.

REFERENCES

1. A. Barth and C. Zscherp, Q. Rev. Biophys. **35**, 369 (2002).
2. C.M. Cheatum, A. Tokmakoff, and J. Knoester, J. Chem. Phys. **120**, 8201 (2004).
3. M. Jackson and H.H. Mantsch, Crit. Rev. Biochem. Mol. Biol. **30**, 95 (1995).
4. A. Elliot and E.J. Ambrose, Nature. **165,** 921 (1950).
5. C. Perez and K. Griebenow, Biotechnol. Lett. **22**, 1899 (2000).
6. A.C. Dong, B. Kendrick, L. Kreilgard, J. Matsuura, M.C. Manning, and J.F. Carpenter, Arch. Biochem. Biophys. **347**, 213 (1997).
7. A.C. Dong, T.W. Randolph, and J.F. Carpenter, J. Biol. Chem. **275**, 27689 (2000).
8. B.K. Mohney, E.T. Petri, V. Uvarova, and G.C. Walker, Appl. Spectrosc. **54**, 9 (2000).
9. T. Lefevre and M. Subirade, Food Hydrocolloid. **15**, 365 (2001).
10. K. Murayama, Y.Q. Wu, B. Czarnik-Matusewicz, and Y. Ozaki, J. Phys. Chem. B. **105**, 4763 (2001).
11. A.C. Dong, J.D. Meyer, J.L. Brown, M.C. Manning, and J.F Carpenter, Arch. Biochem. Biophys. **383**, 148 (2000).
12. J.F. Neault and H.A. Tajmir-Riahi, BBA-Protein Struct. M. **1384**, 153 (1998).
13. P. Robert, M.F. Devaux, N. Mouhous, and E. Dufour, Appl. Spectrosc. **53**, 226 (1999).
14. M. Miyazawa, J. Near Infrared Spec. **6**, 253 (1998).
15. H. Kamishikiryo-Yamashita, M. Tatara, H. Takamura, and T. Matoba, J. Jpn. Soc. Food Sci. **41**, 65 (1994).
16. Y. Wang, K. Murayama, Y. Myojo, R. Tsenkova, N. Hayashi, and Y. Ozaki, J. Phys. Chem. B. **102**, 6655 (1998).
17. Y.Q. Wu, B. Czarnik-Matusewicz, K. Murayama, and Y. Ozaki, J. Phys. Chem. B. **104**, 5840 (2000).
18. S. Navea, A. de Juan, and R. Tauler, Anal. Chem. **75**, 5592 (2003).
19. A.C. Antony, Annnu. Rev. Nutr. **16**, 501 (1996).
20. A.C. Antony, Blood. **79**, 2807 (1992).
21. G.B. Henderson, Annu. Rev. Nutr. **10**, 319 (1990).
22. H. Elnakat and M. Ratnam, Front. Biosci. **11**, 506 (2006).
23. N.C. Kaarsholm, A.M. Kolstrup, S.E. Danielsen, J. Holm, and S.I. Hansen, Biochem. J. **292**, 921 (1993).
24. T.G. Pedersen, I. Svendsen, S.I. Hansen, J. Holm, and J. Lyngbye, Carlsberg. Res. Commun. **45**, 161 (1980).

25. S.I. Hansen, J. Holm, J. Lyngbye, T.G. Pedersen and I. Svendsen, Arch. Biochem. Biophys. **226**, 636 (1983).

26. I. Svendsen, B. Martin, T.G. Pedersen, S.I. Hansen, J. Holm, and J. Lyngbye, Carlsberg. Res. Commun. **44**, 89 (1979).

27. J. Holm, S.I. Hansen, and M. Høier-Madsen, Biosci. Rep. **21**, 305 (2001).

28. J. Holm and S.I. Hansen, Biosci. Rep. **21**, 745 (2001).

29. K. Sasaki, M. Natsuhori, M. Shimoda, Y. Saima, and E. Kokue, Am. J. Physiol. **270**, 105 (1996).

30. S. Sabharanjak and S. Mayor, Adv. Drug Deliv. Rev. **56**, 1099 (2004).

31. R.J. Lee, S. Wang, and P.S. Low, BBA-Mol. Cell. Res. **1312**, 237 (1996).

32. B. Czarnik-Matusewicz, K. Murayama, R. Tsenkova, and Y. Ozaki, Appl. Spectrosc. **53**, 1582 (1999).

33. K. Murayama, B. Czarnik-Matusewicz, Y.Wu, R. Tsenkova, and Y. Ozaki, Appl. Spectrosc. **54**, 978 (2000).

34. K. Murayama and Y. Ozaki, Biopolymers. **67**, 394 (2002).

35. B. Yuan, K. Murayama, Y. Wu, R. Tsenkova, X. Dou, S. Era, and Y. Ozaki, Appl. Spectrosc. **57**, 1223 (2003).

36. K. H. Esbensen, "Principal component analysis (PCA)-in practice". In Multivariate Data Analysis –in practice. An introduction to multivariate data analysis and experimental design, K.H. Esbensen, Ed. (Camo Process, Oslo, 2001), Chap. 4, p. 75.

37. I. Svendsen, S.I. Hansen, J. Holm, and J. Lyngbye, Carlsberg Res. Commun. **49**, 123 (1984).

38. B.W. Sigurskjold, T. Christensen, S.I. Hansen, and J. Holm, "Thermal stability of folate binding protein from cow's milk with and without complexation with ligands". In Chemistry and Biology of Pteridines and Folates 1997. Proceedings of the Eleventh International Symposium on Pteridines and Folates, W. Pfleiderer and H. Rokos, Eds. (Blackwell Science, Berlin, 1997), p. 345.

39. H. Martens, J.P. Nielsen, and S.B. Engelsen, Anal. Chem. **75**, 394 (2003).

40. S.T.R. Walsh, R.P. Cheng, W.W. Wright, D.O.V. Alonso, V. Daggett, J.M. Vanderkooi, and W.F. DeGrado, Protein Sci. **12**, 520 (2003).

41. H. Torii and M. Tasumi, J. Chem. Phys. **95**, 3379 (1992).

42. K. W. Surewicz and H.H. Mantsch, Biochim. Biophys. Acta. **952**, 115 (1988).

43. Y.M. Dou, P.F. Baisnee, Y. Pecout, J. Nowick, and P. Baldi, Bioinformatics. **20**, 2767 (2004).

44. J. Wang, M. Sowa, K. Ahmad, and H.H. Mantsch, J. Phys. Chem. **98**, 4748 (1994).

45. R. L. Baldwin, Biophys. J. **71**, 2056 (1996).

46. N.L. Sefara, N.P. Magtoto, and H.H. Richardson, Appl. Spectrosc. **51**, 536 (1997).

47. J. Holm, S.I. Hansen, and J. Lyngbye, Biochim. Biophys. Acta. **529**, 539 (1980).

48. D. B. Zheng, H.M. Lim, J.J Pene, and H.B. III White, J. Biol. Chem. **263**, 11126 (1988).

49. H.L. Monaco, EMBO J. **16**, 1475 (1997).

50. T.F. Kumosinski, H. Pessen, and H.M. Farrell, Arch. Biochem. Biophys. **214**, 714 (1982).

51. M. Wasylewski, J. Protein. Chem. **19**, 523 (2000).

Appendix IV-1

NIR assignments to lipid and starch

Lipid absorptions in the 1400-2400 nm region

Wavelength [nm]	Assignment
1648	: 2*CH-str.
1696	: 2* CH_3 antisym-str.
1722	: 2* CH_2 antisym-str.
1736	: CH_2 antisym-str.+CH_2 sym-str.
1760	: 2*CH_3/CH_2 sym str.
1816	: 2*C=C-CH str.+ CH_2 bend.
1856	:
2144	: CH cis str.+ C=C-str.
2304	: CH_3 antisym-str. +CH_3 antisym-bend.
2336	: CH_3 antisym-str. +CH_3 sym-bend.
2384	: CH_3 sym-str. +CH_3 -bend.

Starch absorptions in the 1400-2400 nm region.

Wavelength [nm]	Assignment
1400-1600	: 2*OH str.
1702	: 2*CH_2/CH-str.
1748	: 2*CH_2/CH-str
1770	: 2*CH_2/CH-str.
1900	: OH-str.+2*CO-str.
1920	: OH-str.+ bend
2090-2130	: OH-str.+bend+CO-str.
2260	: OH str. +bend or C-O-C combination
2272	: OH str.+C-C str.
>2280	: CH str.+bend

References:

Barton II, F. E., Himmelsbach, D.S., Archibald, D.D. (1996). Two-dimensional vibration spectroscopy. V: Correlation of mid- and near infrared of hard red winter and spring wheats. *J. NearInfrared Spec*. 4, 139-152.

Chung, H., Arnold, M.A. (2000). Near-infrared spectroscopy for monitoring starch hydrolysis *Appl. Spectrosc*. 54, 277-283.

Gouti, N., Rutledge, D.N., Feinberg, M.H. (1998). Factorial correspondence regression applied to multi-way spectral data. *Analysis*. 123, 1783-1790.

Rodriguez-Saona, L.E., Khambaty, F.M., Fry, F.S., Calvey, E.M. (2001). Rapid detection and identification of bacterial strains by Fourier Transform Near-Infrared spectroscopy. *J. Agric. Food. Chem*. 49, 574-579.

Appendix IV-2

Preprocessing method for NIR in Experiment IV

Method: A 2nd derivative water spectrum is used as “bad spectrum” in an EMSC correction of the 2nd derivative gluten spectra (in the region: 1200-2340 nm). This process reduces ligth scattering and water content variations in NIR spectra of moistened gluten (from Experiment IV). See Fig. IV-A and IV-B.

IV-A

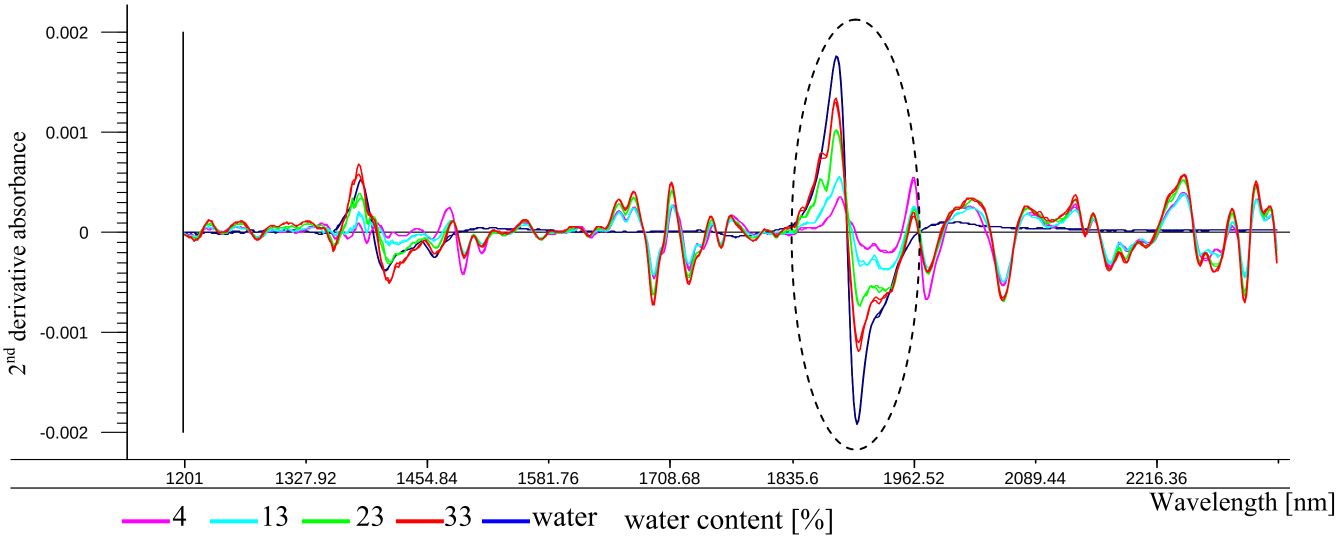


Figure IV-A: 2nd derivative transformed spectra of gluten (obtained in reflectance mode) and a water (obtained in transflectance mode).

IV-B

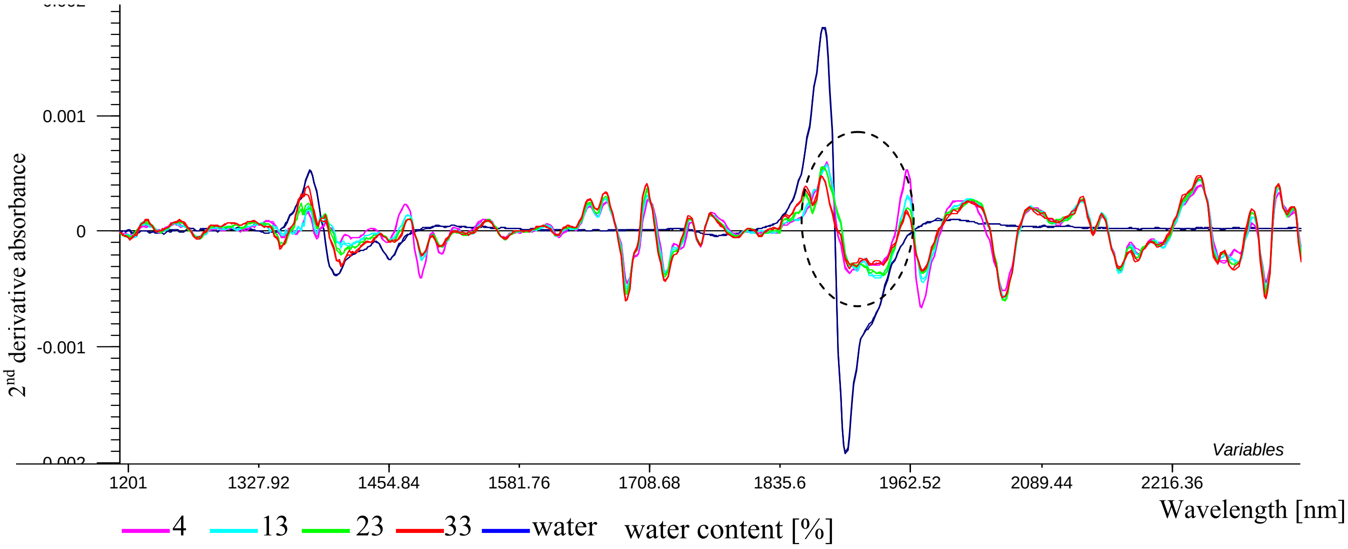


Figure IV-B: 2nd derivative transformed spectra of gluten (obtained in reflectance mode) and a water (obtained in transflectance mode) after EMSC correction with the water spectrum used as “bad spectrum”.

Appendix V

Additional figures to Experiment V

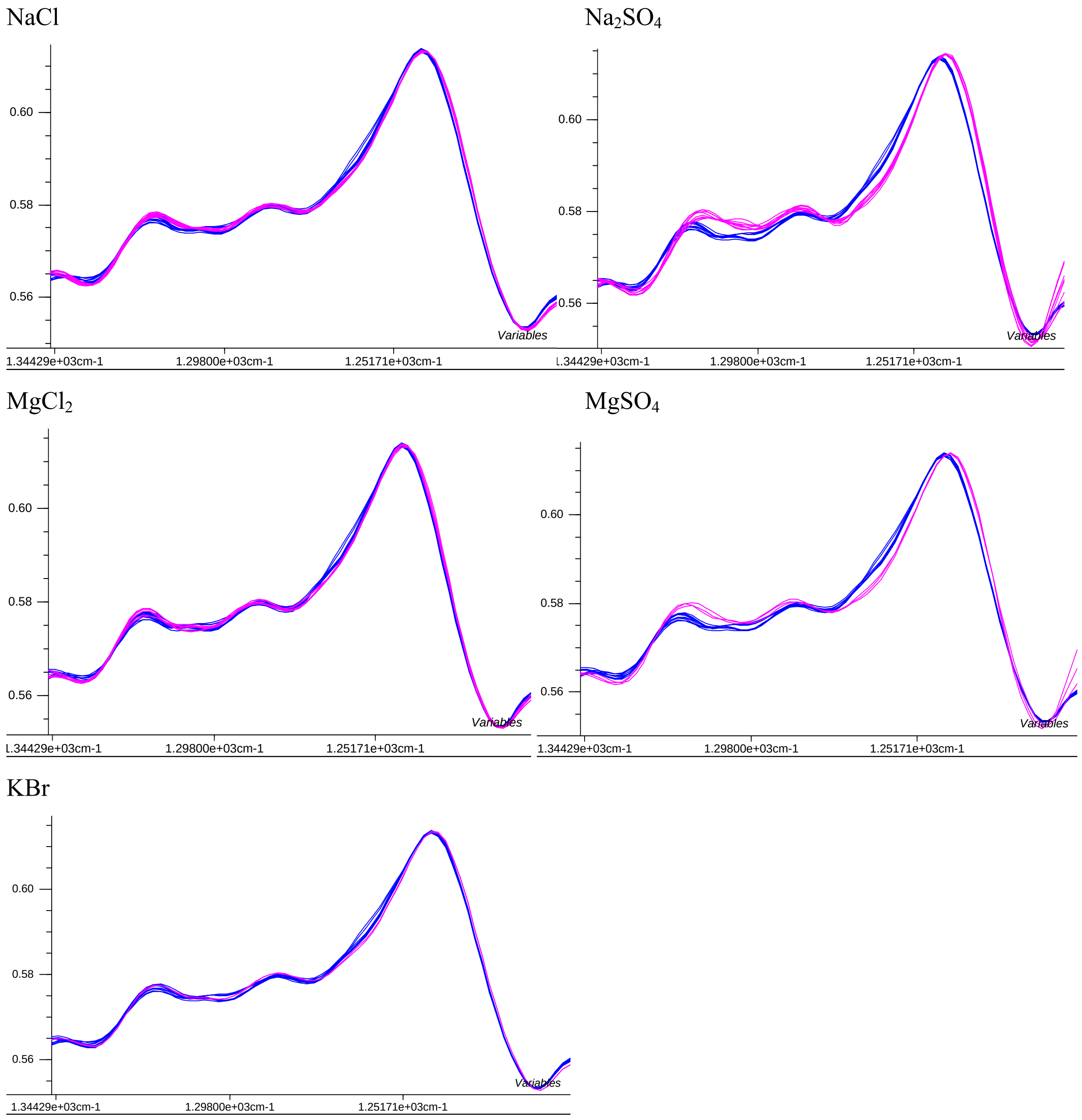


Fig. V-A: Amide III band region after EMSC correction with bad-spectrum subtraction, for correction of the water spectrum variations due to the various salts. Spectra of gluten hydrated in water (blue) or in different 1.0 M salt solutions (pink) are compared.

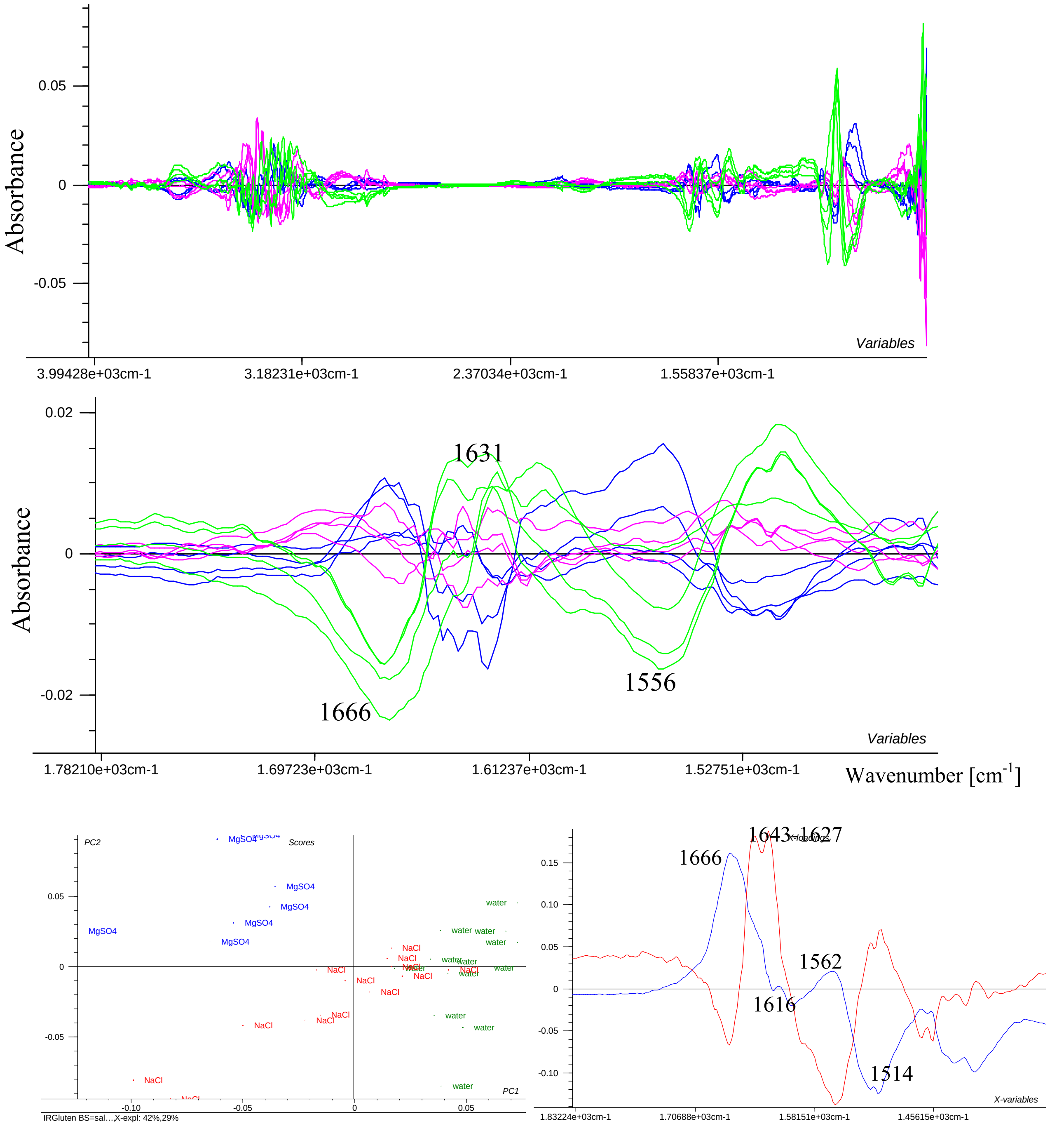


Fig. V-B. Mean-centred EMSC-residuals of ATR-FTIR spectra of gluten in water, NaCl and $MgSO_4$ (after EMSC#2 as described in Paper III in section 2.5). Upper figure: Full wavenumber region. Water (blue), 1M NaCl (pink), 1M $MgSO_4$ (green). Middle figure: Amide I and II band region. Water (blue), 1M NaCl (pink), 1M $MgSO_4$ (green). Lowest figures: PCA score plot PC1 vs. PC2 (left) from analysis of the amide I and II region and corresponding loading plots (right). PC1 explains 42 % and PC2 explains 29 % of the variation.

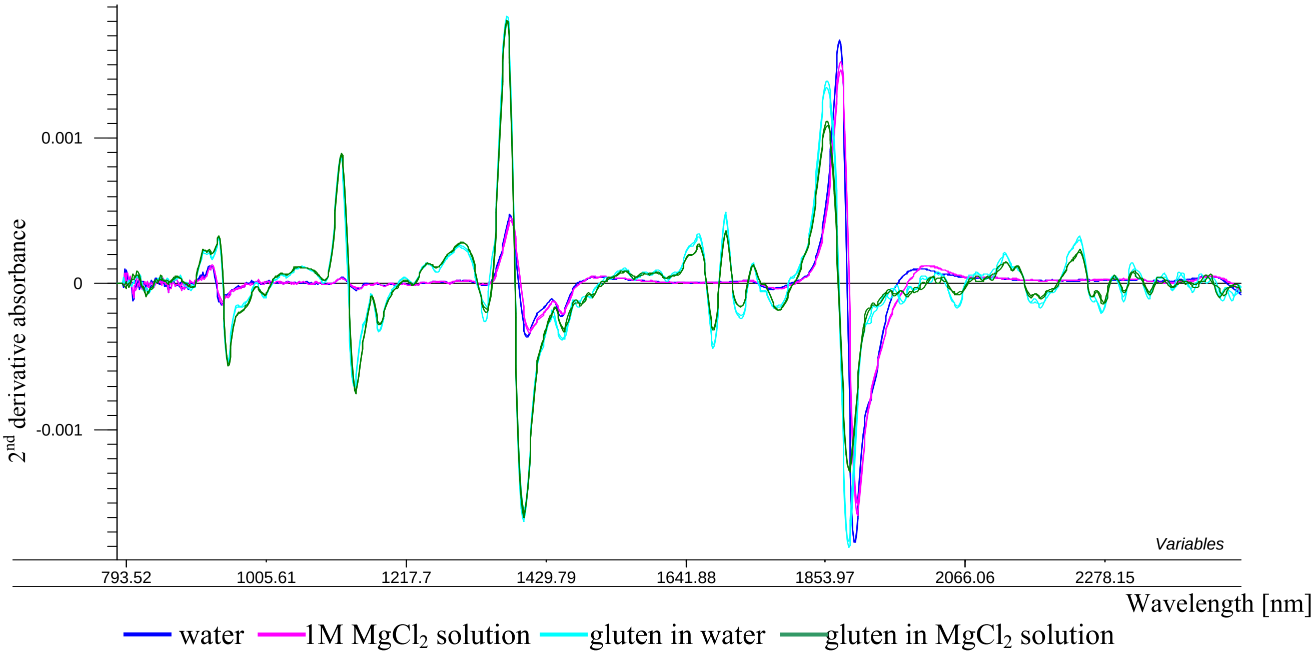


Fig. V-C: 2^{nd} derivative NIR spectra of gluten hydrated in water or 1M $MgCl_2$ solutions. Also the salt solution spectra are shown. The water peak at 1930 nm is shifted for the gluten samples compared to the water samples. Therefore, it is not possible to use the PCA loadings from the salt-solution spectra to correct for water variations in the sample spectra.

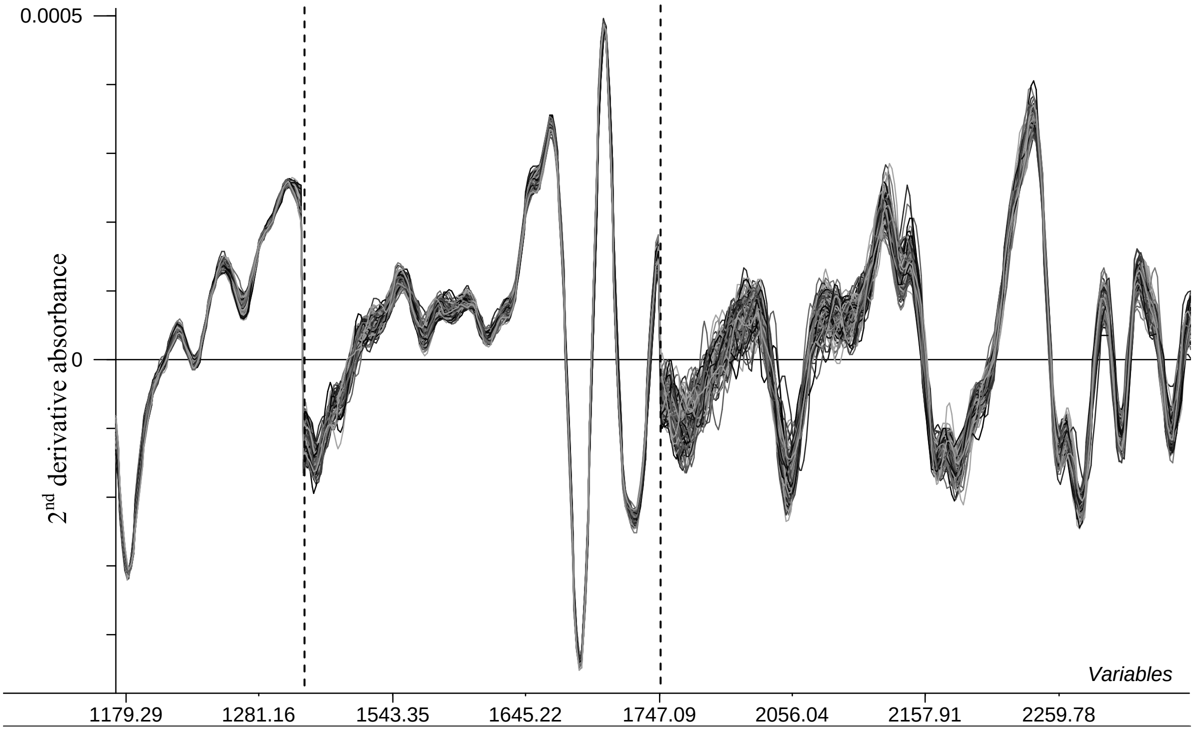


Fig. V-D. All pretreated NIR spectra of gluten hydrated in water and different salt solutions (in Experiment V). The spectra were 2^{nd} derivative transformed and subsequently EMSC corrected in selected regions (1175-1320 nm, 1480-1750 nm and 1960-2360 nm). The three regions are shown together.

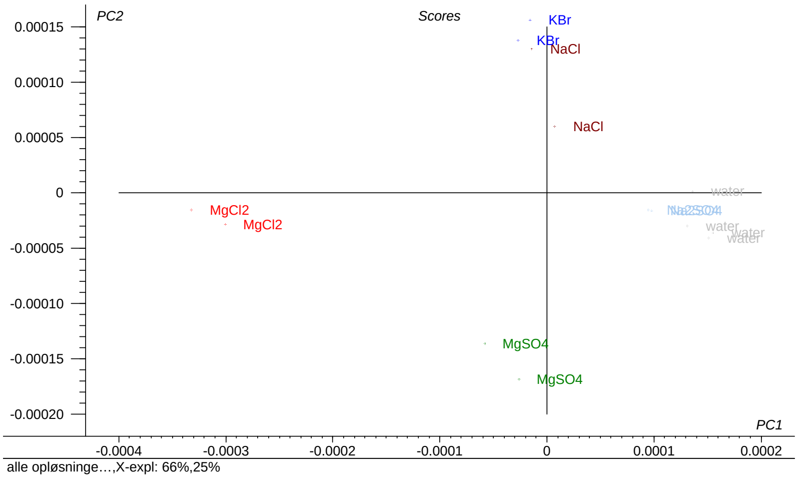


Fig. V-E. PCA score plot (PC1 vs. PC2) of some salt solution spectra, pretreated as the gluten spectra in Fig. V-C.

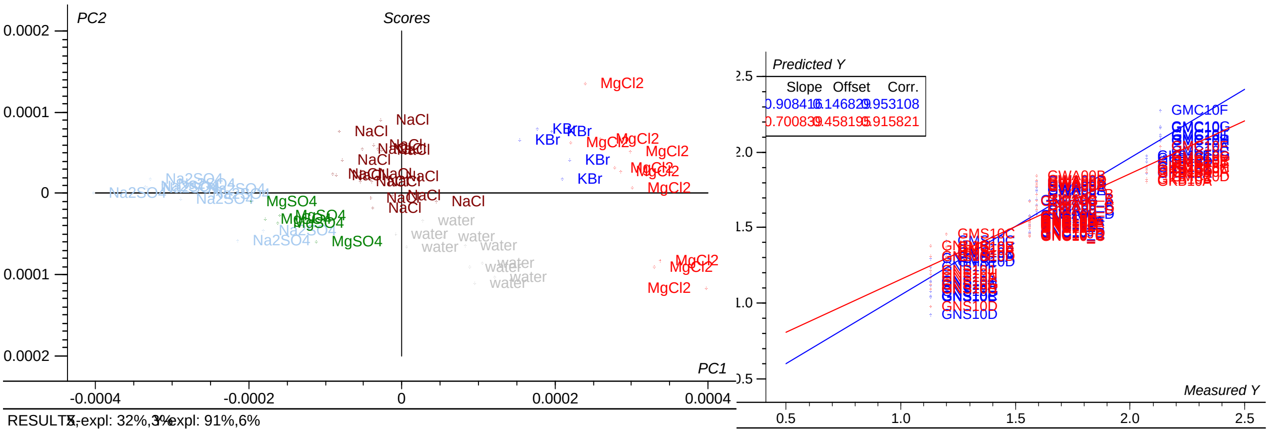


Fig. V-F. Results from a PLSR model (X=preprocessed NIR spectra, Y=water content). Samples =gluten hydrated in 1.0 M salt solutions. Left: Score plot (PC1 vs. PC2). Right: Predicted vs. measured plot.