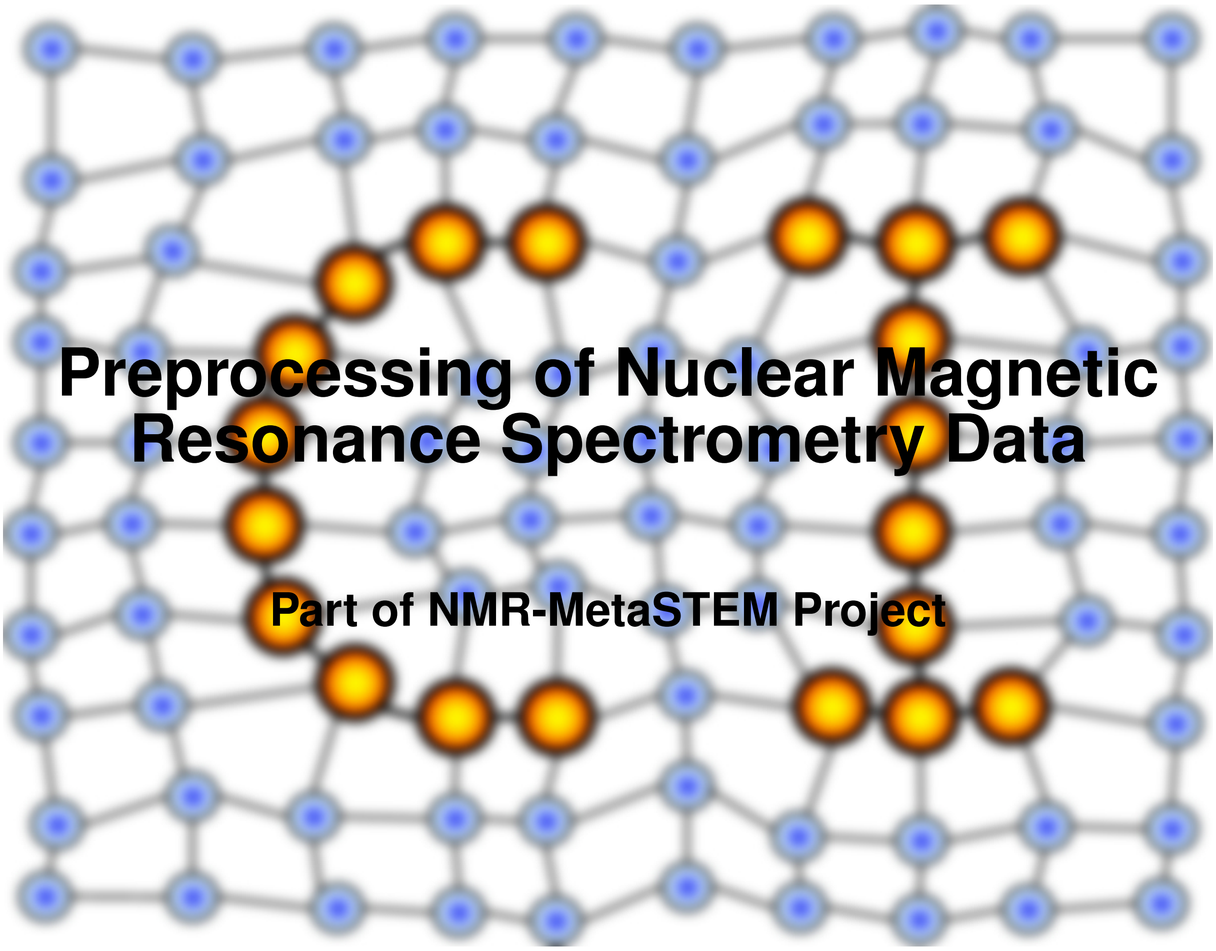
MACHINE LEARNING REPORTS



Preprocessing of Nuclear Magnetic Resonance Spectrometry Data

Part of NMR-MetaSTEM Project

Report 01/2007

Frank-Michael Schleif

- submitted 27th August 2007, published 7th September 2007 -

Machine Learning Reports,Research group on Computational Intelligence,
http://www.uni-leipzig.de/˜compint

Abstract

Nuclear magnetic resonance (NMR) is a physical phenomenon based upon the quantum mechanical magnetic properties of an atoms nucleus. Thereby NMR spectroscopy is a technique to obtain physical, chemical and structural information about molecules. Recent research is focused on the analysis of small molecules, e.g. metabolites by NMR or in a combination of NMR and electro spray ionization mass spectrometry. Thereby two main directions for the analysis are the identification of different metabolites and whose quantities in NMR spectra as well as the identification of group differences e.g. in metabolite mixtures of multiple culture solutions. Here we focus on the preprocessing of NMR spectra. An initial processing workflow for NMR spectra preprocessing is presented which makes use of NMR technique specific aspects and is compared to a workflow using a standard data analysis tool.

1 Introduction

Due to its relatively high sensitivity and ability to detect numerous tissue metabolites, proton NMR spectroscopy (MRS) has become well established as a non-invasive technique for studies of biological systems in vivo and in vitro. The quantification of the NMR-observable metabolites can provide considerable biochemical information, and can help clinical investigators in understanding the role of metabolites in normal and pathological conditions.

Thereby two main directions for the analysis are the identification of different metabolites and whose quantities in NMR spectra as well as the identification of group differences e.g. in metabolite mixtures of multiple culture solutions. Here we focus on the preprocessing of NMR spectra. Thereby the preprocessing workflow is developed such that it will be of use for both applications. The developed algorithmic processing will be applied on multiple data sets obtained from H-NMR measurements of different cultural solutions. The report starts with an introduction in the basic concepts of NMR measurements. Thereby an extract of the tutorial given in [Rze07] is given and modified in accordance to the current experimental settings.

2 NMR measurement procedure and specific aspects of NMR data analysis

NMR spectra are high dimensional functional data generated by NMR spectrometers. The nuclei of all elements carry a charge. When the spins of the protons and neutrons comprising these nuclei are not paired, the overall spin of the charged nucleus generates a magnetic dipole along the spin axis, and the intrinsic magnitude of this dipole is a fundamental nuclear property called the nuclear magnetic moment μ. The symmetry of the charge distibution in the nucleus is a function of its internal structure and if this is spherical it is said to have a corresponding spin angular momentum number of $I = \frac{1}{2}$, of which examples are $^1H,^{13}C, \ldots$. In quantum mechanical terms, the nuclear magnetic moment of a nucleus can align with an externally applied magnetic field of strength B_0 in only $2I + 1$ ways, either reinforcing or opposing B_0. The energetically preferred orientation has the magnetic moment aligned parallel with the applied field and is often given the notation α, whereas the higher energy anti-parallel oprientation is referred to as β. The rotational axis of the spinning nucleus cannot be orientated exactly parallel (or anti-paralled) with the direction of the applied field B_0 but must precess about this field at an angle with an angular velocity given by the expression:

$$\omega_0 = \gamma B_0 \qquad \text{Larmor frequency in Hz}$$

The constant γ is called the magnetogyric ratio and relates the magnetic moment μ and the spin number I for any specific nucleus. For a single nucleus with $I = \frac{1}{2}$ only one transition is possible between the two energy levels. NMR is all about how to interpret such transitions in terms of chemical structure. Skipping some further details on NMR we consider a sample which is placed into a magnetic field and for which most spins are oriented into one direction. We may now start with a so called pulsed NMR experiment. Thereby a transmitter oscillator placed perpendicular to B_0 generates a pulse of electromagnetic radiation producing a magnetic field B_1 along the x-axis. A

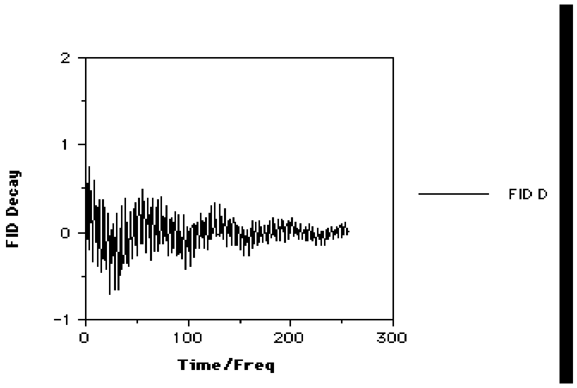


Figure 1: Pulsed NMR experiment and obtained sample signal.

range of frequencies $+\omega \pm 1/t_p$ (with t_p a duration) are produced, enabling resonance to be simultaneously established with all the Larmor frequencies of protons within $1/t_p$ of $+\omega$. The phase of each set of identical resonances is rendered coherent, which tips the macroscopic magnetisation vector for each resonance away from B_0, by the angle θ defined by;

$$\theta = \gamma B_1 t_p$$

A value of $\theta = 90°$ is often referred to as a $90°$ pulse because of the angle the magnetisation vector is changed by, i.e. from the z-axis into the xy-axis plane. Our particular pulse stops after t_p seconds, when all the various protons have an excited state population which is no longer in equilibrium with the ground state. Equilibrium is re-established via spin-lattice and spin-spin relaxation, a process which takes about 5-6 seconds for protons and which involves the return of the macroscopic vector M to the z-axis, i.e. the flip angle θ decays back to zero. M for each set of different protons will have a different Larmor precession frequency, which in this experiment will span the range 715Hz. As long as θ is not 0, the resultant non-zero vector component M in the direction of the receiver coil placed on the y axis induces a sinusoidal current in this receiver. Each set of distinct protons will produce a sine (or cosine) wave whose frequency matches their precession frequency and the intensity of which is related not only to the phase of the sine wave but also to the value of θ at any instant. The signal detected in the receiver therefore resembles a collection of exponentially decaying sine waves, and is called a Free Induction Decay or FID c.f. 1 (given in [Rze07]).

We now have a signal corresponding to our NMR spectrum which contains a set of sine/cosine waves measured as a function of time and decaying towards zero intensity at an exponential rate. This signal is said to be analogue i.e. it varies continuously with time and is described as a *time domain signal*. Thereby the signal is digitized and stored. Further the signal is processed by a fast Fourier transformation (FFT) as

$$F(\omega) = f(t) \exp^{-i\omega t} dt$$

Here the $FID = f(t)$ and $F(\omega)$ is the same data expressed as frequencies $(= 1/T)$ rather than as times and is called a *frequency domain spectrum*. This is the form we know for conventional NMR spectra, i.e. frequencies relative to an internal standard such as DSS. $F(\omega)$ can be rewritten in terms of a real and an imaginary part. Each component contains the same frequency data but with a different combination of phase and amplitude. Only the real part will be considered subsequently. Due to technical reasons for any NMR spectrometer a short delay between the end of t_p and the start of the measurement is required. The short delay means measurement begins when the sine waves are already out of phase. This so called *first order phase error* can be calculated from a suitable combination of the real and imaginary components of $F(\omega)$. Another phasing error due to technical imperfections in the spectrometer is often referred to as the *zero order phase* correction. These are both applied to $F(\omega)$ after the FT is complete. In our experiments we have a Bruker NMR device, thereby an addition digital filter is applied at the beginning of the signal which has to be removed prior any further high level processing.

To conclude the inital processing steps are given as:

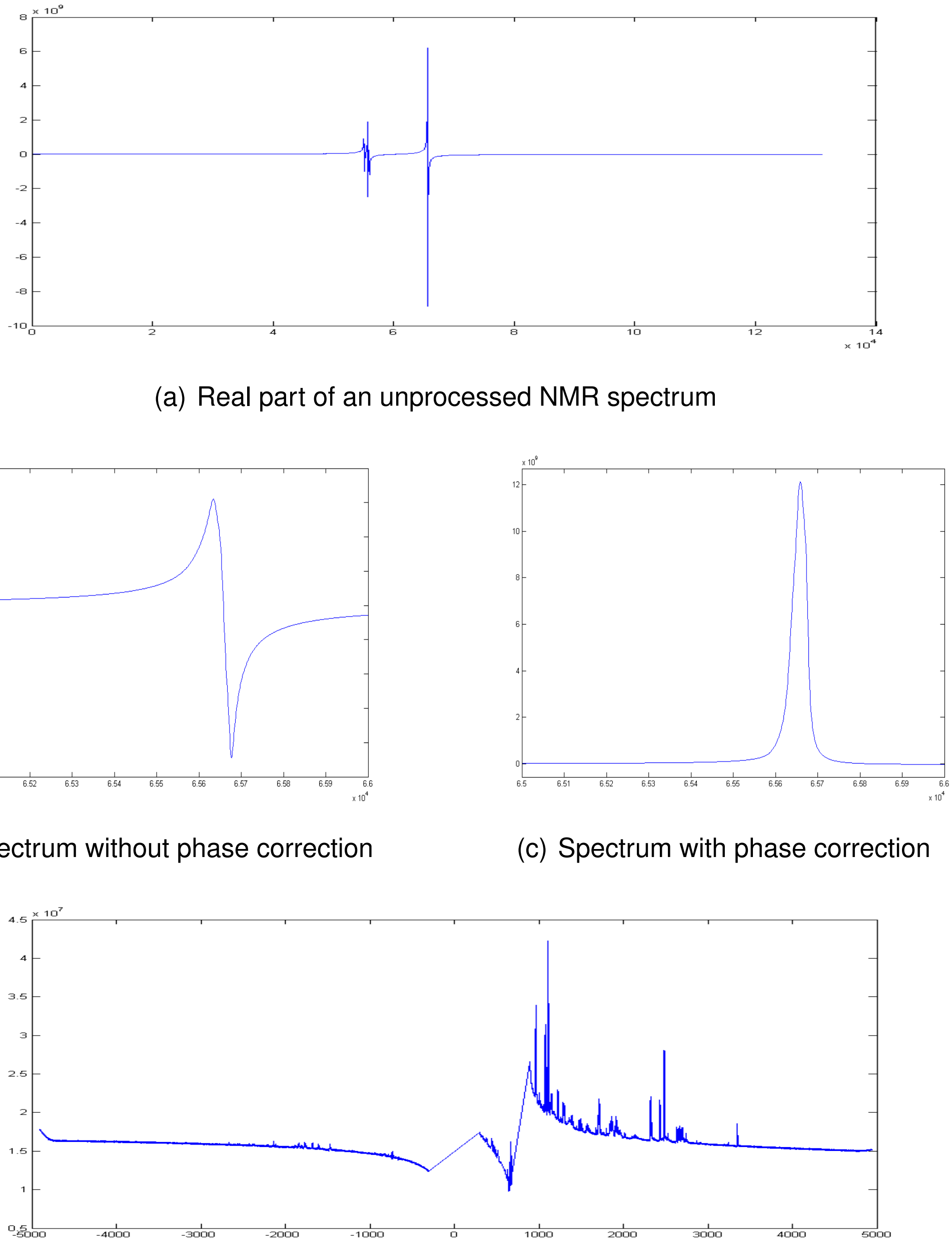
1. preparation of the biological sample
2. measuring a NMR spectrum of the sample
3. loading the spectra into the processing software
4. removing the digital filter
5. application of an FFT on the NMR signal
6. automatic phase correction

The loading and basic processing such as automatic phase correction can be done by use of the backend algorithm of the matNMR libary [Bee07]. Thereby the automatic phase correction can be done using the approach of Chen et al. as given in [CLGG02].

In our experiments all spectra contain a DSS signal of arround $1510^-6mol/lt$ which result in a peak at approximately 3350Hz. This signal can be used to align all measured spectra to each other which will be important for later group comparisons. As an additional step the measurement region responsible for the water signal will be removed by a cut of procedure taking the expect maximal signal intensity into account which is related to a multiple of the DSS signal intensity. The effect of this basic processing applied onto the signal is depicted in Figure 2. Now two possible approaches for the further processing may come into mind. On the one hand side we may stay on the profile spectra or on the other we may try to identify the peaks in the spectrum and tranform the profile spectra into line spectra of peak lists. Focusing on the later we will initally apply a baseline correction to each spectrum, followed by a peak picking algorithm. Thereby we finally obtain a list of peaks for each spectrum which is believed to be sufficient for the subsequently data analysis steps. Now we briefly describe the baseline correction and the peak picking procedure in two different settings.

2.1 Preprocessing by use of ClinProTools

First we will process the real part of the aligned NMR spectra using the commercial spectra processing tool ClinProTools 2.2 (Bruker Daltonik GmbH, Bremen, Germany) [KHT$^+$05]. ClinProTools (CPT) has been originally developed to process mass spectrometric data and has found wide acceptance in the field of clinical proteomics and MALDI-Imaging [KHT$^+$05, PFL$^+$03, DBD$^+$04, BCFa05, aC03, SCS$^+$07, ZLMS04] CPT allows the import of ascii spectra and the export of peak list information as well as the generation of classification models in case of multiple labeled sets of data. Thereby



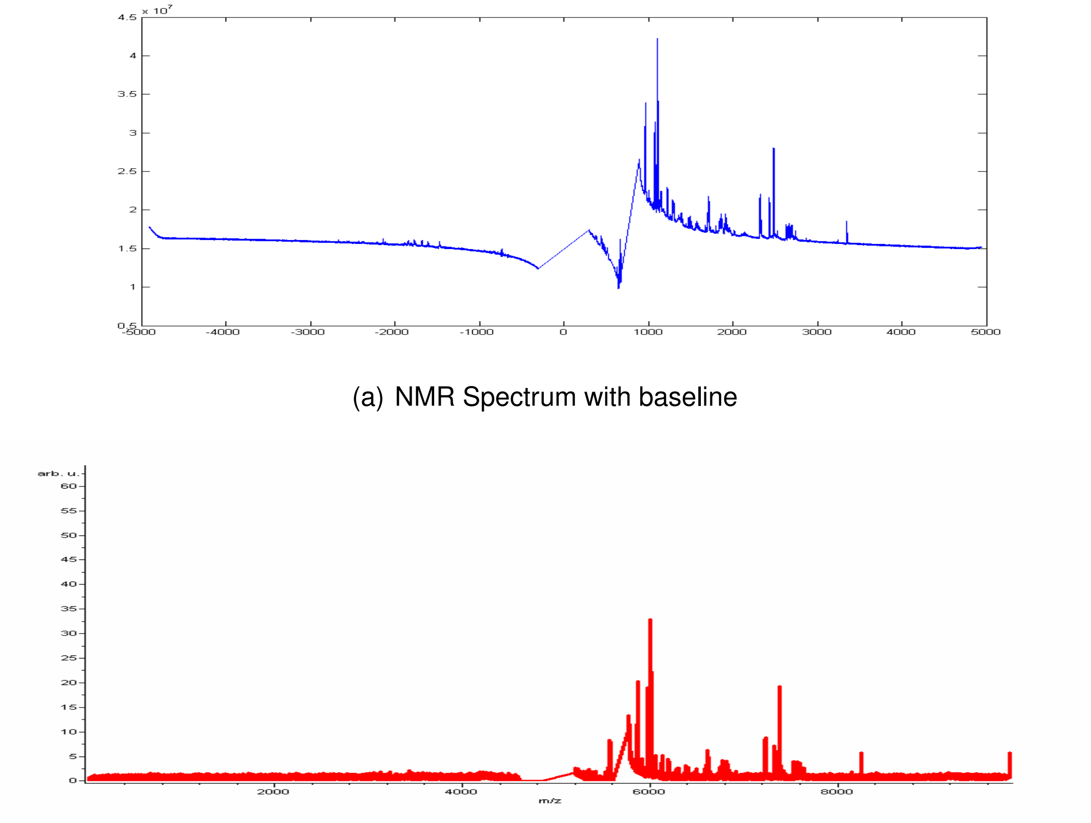
(a) Real part of an unprocessed NMR spectrum

(b) Spectrum without phase correction

(c) Spectrum with phase correction

(d) Spectrum without water peak

Figure 2: Illustration of the different basic NMR preprocessing steps. In the first plot the plain real part of the NMR spectrum is shown. In the two subsequent plots the effect of the phase correction is visible by considering the effect on the water peak region. The last plot shows the same spectrum but without very high peaks such as the water peak. The x-axis for plot $1-3$ is given in sample points, while for the last plot the x-axis is given in Hz. The y-axis gives intensities which are related to chemical concentrations in the sample.



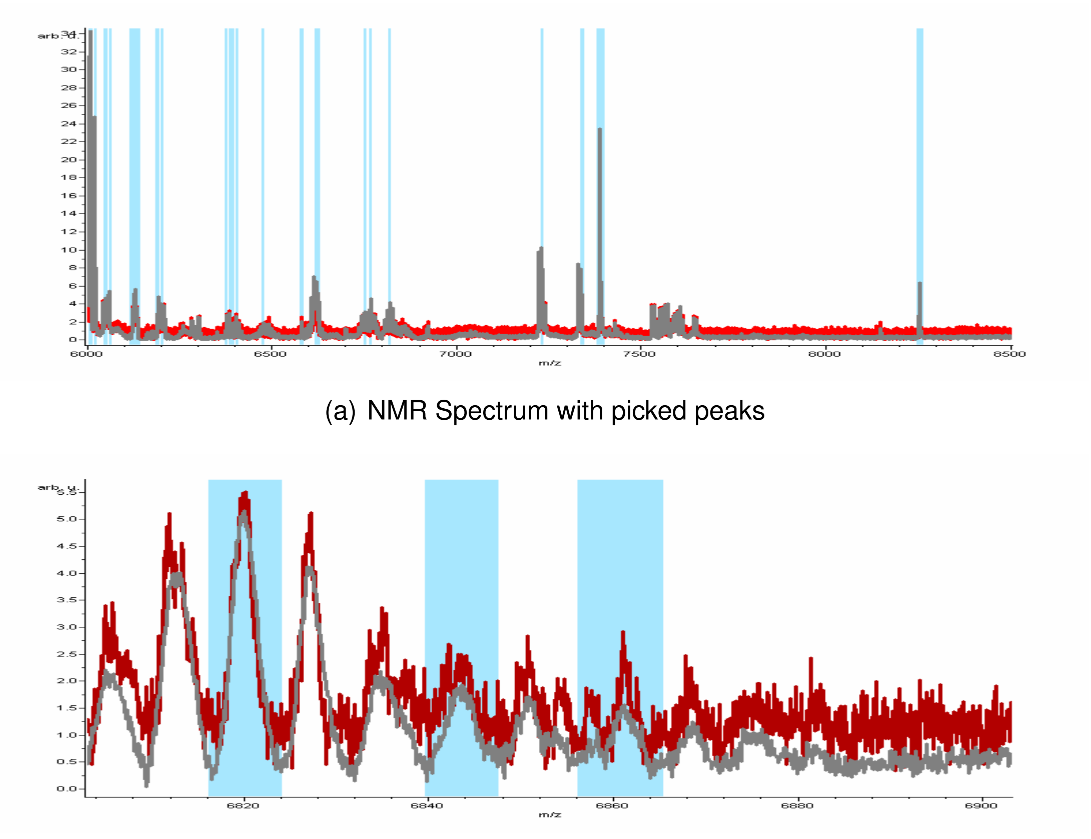
(a) NMR Spectrum with baseline

(b) Spectrum without baseline

Figure 3: Illustration of the effect of a baseline correction by use of the ClinProTools data analysis software. The x-axis labeling is with respect to mass spectrometry data but can be related back to the original frequency axis.

the underlying algorithms are already thoroughly tested and evaluated over multiple years. The import takes place by means of comma separated values (CSV) spectra in two columns. Thereby the first columns contains the x-axis and the second column the y-axis. In case of mass spectrometry data this is given by means of a mass axis in a range of typically $1kDa - 10kDa$ and an intensity axis with arbitrary units. Taking this into account the real part of the preprocessed NMR spectrum is transfered into a list of CSVs such that the frequencies are all positive by an offset addition and the intensities are simply taken as given.

Now an automatic baseline correction is applied. The overall effect of the baseline correction is shown in Figure 3. In the second the an automatic peak picking procedure by means of a so called sum-peak finder is applied onto the spectrum. Thereby a combined list of peaks generated from the single peak lists of each spectrum is generated. This effect is depicted in Figure 4. The obtained list of peaks by means of start and end-positions for the peaks can be used to obtained specific features of the single peaks such as maximal intensity or area under the peak (AUP). Thereby the AUP is a relative measure of the amount of material (e.g. chemical part of a specific metabolite, such as glutamat) present in the original sample, measured in the NMR device. The peak lists or peak feature can be exported and used for further processing steps such as the identification of metabolites employing corresponding library search engines. Alternatively in case of multiple labeled subsets the pattern recognition algo-



(a) NMR Spectrum with picked peaks

(b) NMR Spectrum with a region within 1900Hz-2000Hz

Figure 4: Peak picking by use of ClinProTools. The first plot shows a spectrum with annotate peak regions (underground boxes) and the second plots show a specific part of the spectrum with annotated peaks. Thereby one observes that part of the peaks are not picked. The smooth overlayed spectrum shows the average of the spectra.

rithms available in CPT such as Support Vector Machines (SVM) [**?**] or a Supervised Neural Network approach (SNN) related to [**?**] can be applied. Thereby the identified separating peaks or peak patterns can be further analyzed and potentially tracked back to differences in the expression of specific metabolites.

2.2 Preprocessing by use of Matlab-Algorithms

Beside the formerly mentioned approach a matlab based pre-processing has been developed. Thereby the focus mainly relies on a consistent handling of the data which avoids multiple ex- and import operations. In addition the applied operations can be easily logged, which is important for a standardized processing in the field of clinical sample analysis. The processing steps are the same as in the prior approach upto the step of baseline correction. The baseline correction is implemented in a very simple way using a windowed piecewise cubic interpolation method. The pseudo-code is given in 2.2. This approach has been found to give good results. The effect of the baseline correction can be observed in Figure 6.

Subsequently a peak picking algorithm is applied which makes use of the measurement chararacteristics. In a first step we take the DSS signal, to be exact, the peak region of the DSS signal and model this peak as a reference for the peaks we are look-

```
function [sProcessedNMR,vBaseline] = baseline_correction_matlab(sNMRData)
%% Baseline correction
vYValuesOrig = sNMRData(:,2);
ns = size(vYValuesOrig,1);          % number of points
w = 75; % window size
temp = zeros(w,ceil(ns/w))+NaN;
temp(1:ns)=vYValuesOrig;
[m,h]=min(temp);
g = h>1 & h<w;
h=w*[0:numel(h)-1]+h;
m = m(g);
h = h(g);
vBaseline = interp1(h,m,1:ns,’pchip’);
sProcessedNMR      = sNMRData;
sProcessedNMR(:,2) = [temp(1:ns)-interp1(h,m,1:ns,’pchip’)]’;
sProcessedNMR      = sProcessedNMR’;
return;
```

Figure 5: Matlab code of a simple window based baseline correction. The input is given by a structure which contains the signal intensities. The output is the baseline reduced spectrum.

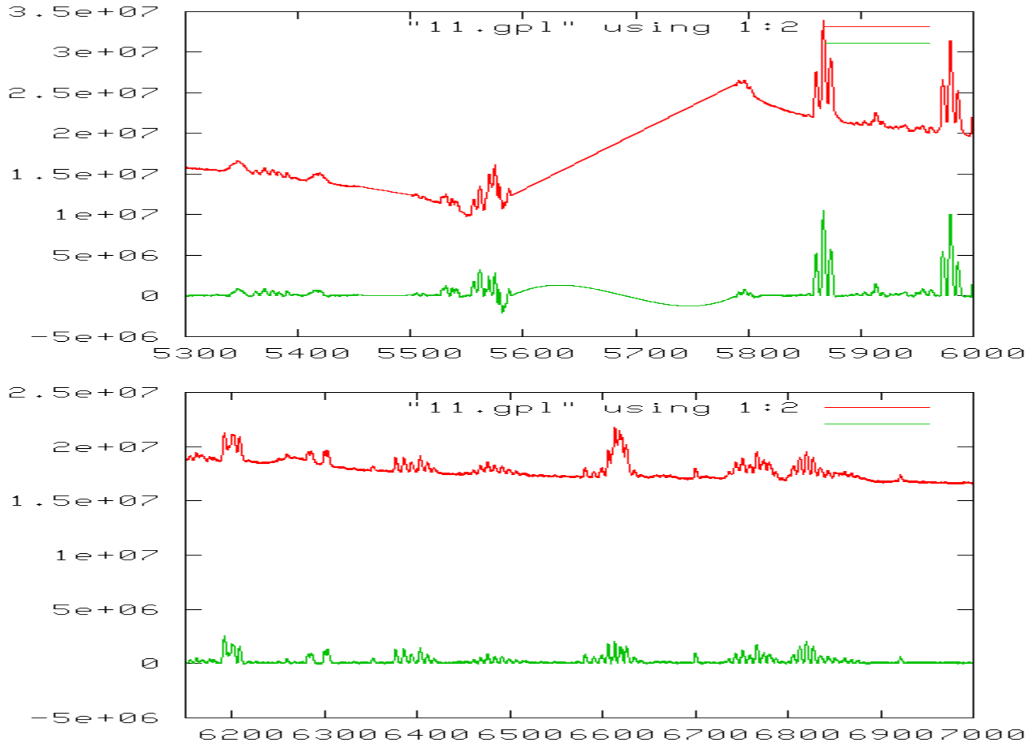


Figure 6: Illustration of the effect of a baseline correction by use of the matlab algorithm. The first plot shows the region around the water peak and the second plot shows the baseline in a normal NMR spectrum region. One observes that in parts of the data the substracted baseline maybe to close to the peak regions such that some broader peaks are negative effected.

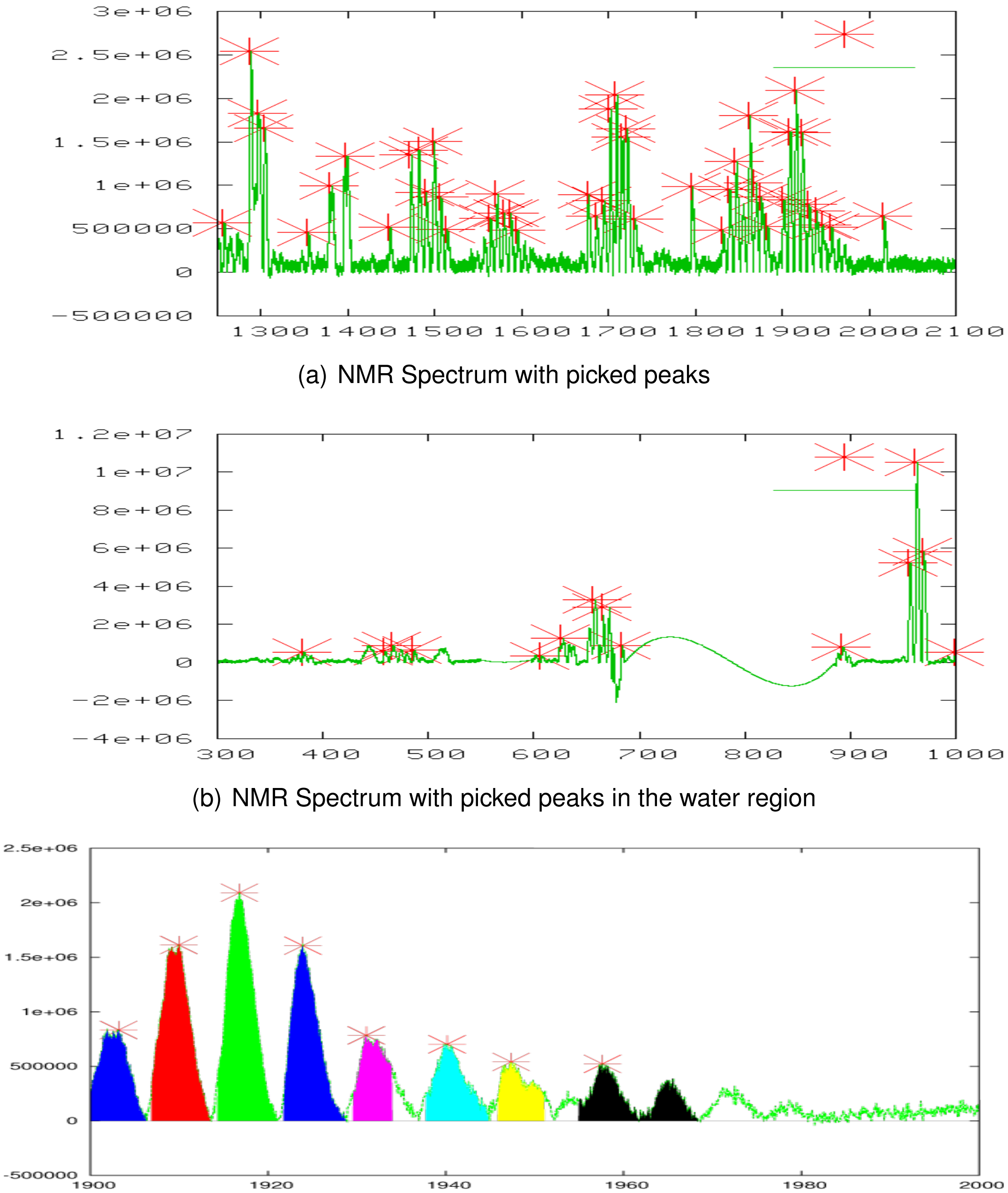
ing for. This peak is considered to be a pattern which is expected to be present in the signal multiple times. A cross correlation analysis is applied on the signal, whereby the baseline corrected spectrum is folded with the DSS pattern. In the obtained absolute, normalized correlation spectrum, high intensities indicate regions which are potential peak locations in the original signal. To get this regions we screen the correlation spectrum for intensities which are above a specific threshold, here 0.01. We calculate the first derivate of the signal which is f' and apply a median filter with a width of $\sigma = 10$ to smooth regions with small descent fluctuations. Here we scan f' for local extrema and log the kind of extremum for each potential peak region. Now a list of potential peak position with extrema attributes is available. Subsequently the peak positions are scanned to determine start and end-positions of the peaks and to screen out candidates which do not fit the criteria. Thereby a peak has a minimal width $mi_w = 0.2\Delta_w$ and a maximal width $ma_w = \Delta_w$ which is related to the peak width Δ_w of the DSS signal, further a peak should have a minimal height of $mi_h = 0.1\Delta_h$ with Δ_h as the height of the DSS signal. The peak candidates are processed and for each candidate the algorithm tries to identify a start and an end region by local minima and maxima searches incorporating the peak constraints defined above. The effect of the peak annotation is depicted in Figure 7.

3 Sample preparation and measurements

In the first initial experiments we applied the above mentionend preprocessing workflows on data sets generated by a Bruker DRX700 NMR device. Thereby a solution of an ultra filtrated FDCPMix with different amounts of glucose has been measured. The FDCPMix solution was solved in D_2O calcium-phosphate at $pH = 7.4$. In each solution a DSS reference is included. The DSS solution consists of $15\mu Ml$ DSS.

4 Conclusions

Two different workflows for the processing of NMR spectra have been presented and analyzed. Thereby we focused on a generic processing without user interactions. The preprocessed spectra and obtained peak list can be used to determine group differences between different sample cohorts or to identify metabolites in single spectra. The first workflow is using a commercial data processing package and gives similar results in comparison to the matlab-based approach but with the additional benefit of processing of the data in an integrated environment with available statistical and machine learning post-processing algorithms. Further a large number of visualization views is available as well as algorithms for outlier detection, variance analysis (PCA) a.s.o. Another point is that the list of detected peaks can be manually annotated such that missing peaks or inappropriate matches can be corrected. This in fact appears to be a necessary functionality, because during the inital experiments multiple visual observed peak were not automatically detected. The matlab based approach avoids multiple im/exports and is currently better adapted for the specific topic of NMR data processing with public available processing algorithms. If the analysis focus relies on the identification of metabolites in single spectra the matlab approach is favourable because it can be more easily adapted into this direction. In the initial experiments the



(a) NMR Spectrum with picked peaks

(b) NMR Spectrum with picked peaks in the water region

(c) NMR Spectrum with a region within 1900Hz-2000Hz

Figure 7: Peak picking by use of the matlab algorithm. The first plot shows the spectrum with annotated peak regions (stars at the top center of the peak), the second plot shows a specific part of the spectrum with annotated peaks around the water region. Thereby it should be mentioned that the water peak is not detected as a peak because its shape doesnt match to the DSS reference peak signal. The last plot shows a zoom into a specific region with peak annotations. Thereby one observes that some peaks are not picked and one is merged with a neighbor peak, but in general most of the peaks are picked in contrast to the standard CPT approach.

number of detected peaks was typically larger than by use of CPT but some of the peaks are merged to neighbored peaks giving errors in later quantification analyses. In conclusion it can not expected that the preprocessing will lead to a nearly perfect peak annotation but it may still be necessary to do some manual interactions or one has to take the fact of missing or invalid peaks into account in subsequently analyses. In case of the identification of group differences, some kind of supervised machine learning or statistical analysis is necessary which can be done very easily within the CPT tool package. So far both preprocessing queues show promising results but reveal still potential for optimization. In the next step the preprocessing should be integrated into an interactive tool environment and applied in metabolomic experiments focusing on identification of metabolites or the identification of differences in sample cohorts. Thereby for the identification of metabolites the usage of simulated spectra would be of high value. The integration of an appropriate simulation tool into the workflow is necessary.

*A*cknowledgement: This work was supported by the BMBF in the project MetaSTEM. The author would like to thank Dr. Thomas Riemer, IZKF, Leipzig, Germany for helpful discussions and support in interpreting the NMR data.

References

[aC03] ET AL., Sau-Mei L.; CLAUSS, Ute: A Novel Rapid Data Mining and Visualization Bioinformatics Tool for Cancer Biomarker Discovery from High Throughput Mass Spectrometry Based Protein Profiling ASCB, 2003. – 43rd Annual Meeting of the American Society for Cell Biology 2003

[BCFa05] BAUMANN, S.; CEGLAREK, U.; FIEDLER, G.M.; ET AL., J. L.: Standardized approach to proteomic profiling of human serum based magnetic bead separation and matrix-assisted laser esorption/ionization time-of flight mass spectrometry. In: *Clinical Chemistry* 51 (2005), S. 973–980

[Bee07] VAN BEEK, J.D.: matNMR. In: *Journal of Magnetic Resonance* 187 (2007), S. 19–26

[CLGG02] CHEN, Li; AN LAIYOONG GOH, Zhiqiang W.; GARLAND, Marc: An efficient algorithm for automatic phase correction of NMR spectra based on entropy minimization. In: *Journal of Magnetic Resonance* 158 (2002), Nr. 1-2, S. 164–168

[DBD$^+$04] DIAZ, M.; BREDIES, K.; DECKER, J.; SCHLEIF, F.-M.; CLAUS, U.; KUHN, M.; MAASS, P.; THIELE, H.; LAUKIEN, F.: Wavelet Transformation applied to the Analysis of Clinical Proteomics Mass Spectra. In: *52st ASMS Conference (ASMS) 2004*, 2004

[KHT$^+$05] KETTERLINUS, R.; HSIEH, S-Y.; TENG, S-H.; LEE, H.; PUSCH, W.: Fishing for biomarkers: analyzing mass spectrometry data with the new ClinProTools software. In: *Bio techniques* 38 (2005), Nr. 6, S. 37–40

[PFL$^+$03] PUSCH, W.; FLOCCO, M.; LEUNG, S.M.; THIELE, H.; KOSTRZEWA, M.: Mass spectrometry-based clinical proteomics. In: *Pharmacogenomics* 4 (2003), S. 463–476

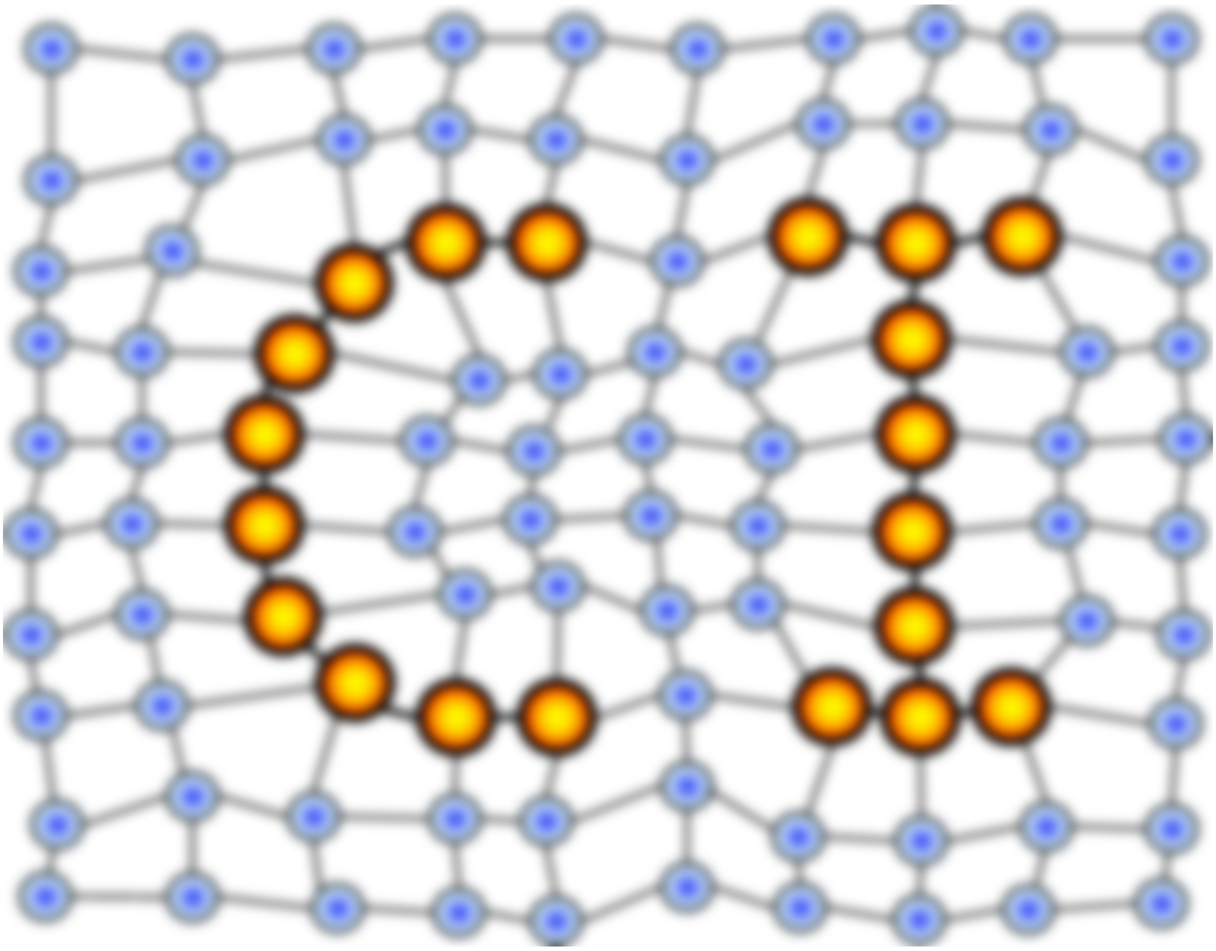
[Rze07] RZEPA, H. *NMR Spectroscopy. Principles and Applications*. http://www.ch.ic.ac.uk/local/organic/nmr.html (21.08.2007). 2007

[SCS+07] SHIN, S.; CAZARES, L.; SCHNEIDER, H.; MITCHELL, S.; LARONGA, C.; AN RR PERRY, OJ S.; DRAKE, RR: Serum biomarkers to differentiate benign and malignant mammographic lesions. In: *J. Am. Coll. Surgery* 204 (2007), Nr. 5, S. 1065–1071

[ZLMS04] ZHANG, X.; LEUNG, S.M.; MORRIS, C.R.; SHIGENAGA, M.K: Evaluation of a novel,integrated approach using functionalized magnetic beads bench-top MALDI-TOF-MS with prestuctured sample support and pattern recognition software for profiling potential biomarkers in human plasma. In: *Biomolecular Tech.* 15 (2004), Nr. 3, S. 167–175

MACHINE LEARNING REPORTS

Report 01/2007



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

PD. Dr. rer. nat. Thomas Villmann & Dr. rer. nat. Frank-Michael Schleif
Medical Department, University of Leipzig
Karl-Tauchnitz 25, D-04107 Leipzig, Germany •
http://www.uni-leipzig.de/˜compint

▽ Copyright & Licence

Copyright of the articles remains to the authors. Requests regarding the content of the articles should be addressed to the authors. All article are reviewed by at least two research in the respective field.

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.