

AUTOMATIC IDENTIFICATION AND QUANTIFICATION OF METABOLITES IN ^1H -NMR MEASUREMENTS

F.-M. Schleif¹, T. Riemer², M. Cross² and T. Villmann¹

¹ Medical Department, Leipzig University, Semmelweisstrasse 22, 04103, Germany

² Interdisciplinary Center for Clinical Res., Hematology, Leipzig University, Inselstrasse, 04103, Germany
{schleif,villmann}@informatik.uni-leipzig.de,crossm@medizin.uni-leipzig.de,riemer@uni-leipzig.de

ABSTRACT

Stem cells therapy is currently at the frontier of biomedical research. A better understanding of the metabolism of stem cells is necessary to improve and extend initial promising results. Nuclear Magnetic Resonance Spectroscopy (NMR) allows for a precise measurement of metabolites in cell extracts. The identification and quantification of these metabolites is essential to model cellular metabolic network activity. To meet clinical standards and high throughput demands a full automatic evaluation is required. A NMR signal processing system is presented and initial results for the identification and quantification of metabolites from murine hematopoietic progenitor cell extracts (FDCPmix cells) and simulated spectra are given.

1. INTRODUCTION

Nuclear Magnetic Resonance Spectroscopy (NMR) is one of the most promising techniques for the analyses of complex substances such as cell extracts. One prominent NMR application is metabolite profiling in stem cell biology. NMR spectra are high dimensional functional signals consisting of a multitude of peaks. The peak positions describe the presence of specific chemical compounds in the analysed material while the area of the peaks are quantitative with respect to the amount of this analyte in the substance. To meet clinical standards and high throughput demands a full automatic evaluation is recommended. Here we present the basic methodology of such a system. The paper is organized as follows. First we briefly explain the bio-chemical aspect of the considered studies. Subsequently key elements of the automatic analysis system are described. Thereafter the system behavior is shown in the analysis of synthetic and real metabolite data. The paper is closed by the discussion of the results.

2. MATERIAL AND DATASETS

We consider real and simulated ^1H NMR spectra recorded at 700.15 MHz with 65K complex data points. The simulations are done using the gamma-library[1]. It is assumed that each sample is solved in D_2O with DSS added as reference standard set to 0.0 ppm. The simulated data are generated from known spin systems [2] and calibrated by own reference measurements. Thereby we consider the following metabolites Alanine (Ala), Citric Acid (Cit), Glycine (Gly), Lactate (Lac), Malate (Mal), Myo-Inositol (Myo), Serine (Ser) and Succinate (Suc). The biological data are taken from FDCPmix cells cultivated in three different levels of glucose concentration (1 *m*-mol, 5 *m*-mol and 25 *m*-mol) in growth medium as specified in [3]. For each concentration level at least 4 ^1H NMR spectra have been recorded.

3. AN AUTOMATIC SYSTEM FOR ^1H -NMR MEASUREMENTS

The NMR data of modern spectrometers are usually obtained in the time domain and thereby given as a set of sine/cosine waves measured as a function of time and decaying toward zero intensity at an exponential rate (free induction decay). Under ideal conditions we can write the signal as:

$$s(t) = \sum_j^J A_j e^{i(w_j t + \phi_j) - t/T_{2j}^*}$$

with A_j as the amplitude, w_j as the frequency, ϕ_j as the phase and T_{2j}^* as the effective decay time of all spectral components $j \in J$. Assuming the exponential decay of $s(t)$ we obtain the signal as a sum of Lorentzian lines after application of an FFT.

$$S(w) = \sum_j^J e^{i\phi_j} (a_j(w) + d_j(w))$$

with $a_j(w) = \frac{A_j T_{2j}^*}{1+(w-w_j)^2 T_{2j}^{*2}}$ as the so called absorption signals and $d_j(w) = \frac{-A_j T_{2j}^{*2} (w-w_j)}{1+(w-w_j)^2 T_{2j}^{*2}}$ as the dispersive signal. In case of perfect phasing $\phi_j = 0, \forall j \in J$, $S(w)$ becomes:

$$S(w) = \sum_j^{|J|} \frac{A_j T_{2j}^*}{1+(w-w_j)^2 T_{2j}^{*2}}$$

For real, biological spectra these assumptions are not fulfilled in general. Multiple preprocessing steps are needed to get interpretable data from the Fourier transformed NMR spectra as depicted in Figure 1.

Due to technical reasons for each NMR spectrometer a short delay between the end and the start of the measurement occurs. This implies that the sine-waves being out of phase, called the first order phase error (see Fig. 1, plot 1). Due to further imperfections a second error called zero order phase error occurs. Therefore a phase correction is desired, which can be done using the approach given in [4]. As for most spectral data a baseline correction is necessary to remove broad, baseline distorting components from the narrow metabolite NMR signals. The baseline correction is done using the a cubic interpolation approach as shown in Figure 2 (simplified).

As another (optional) step the spectra can be deconvolved see e.g [5], In the deconvolution one tries

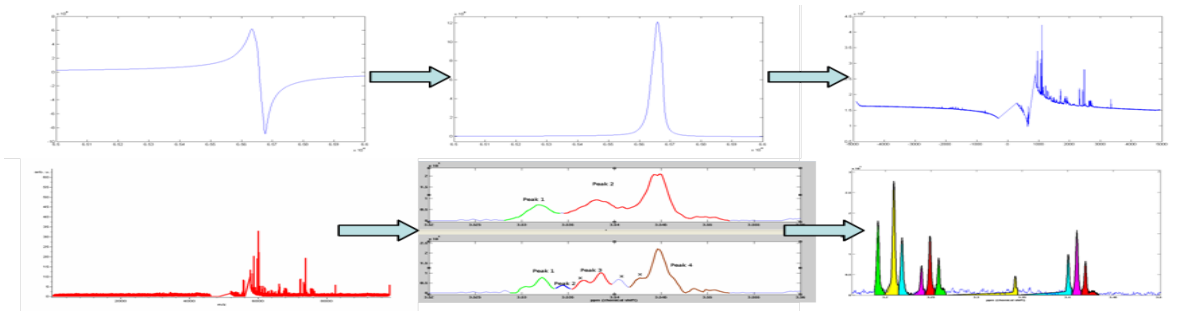


Figure 1. Workflow of the preprocessing steps applied in our NMR system. From top-left to bottom-right: We begin with a fourier transformed signal (out of phase), this error is corrected by phase correction as shown in plot 2, subsequently the water peak (with known position) is removed (interpolated by a cubic spline) and a baseline correction is applied plot 3 – 4, as an optional step a deconvolution can be tried - plot 5 followed by a peak picking algorithm plot 6.

```
function [vBL] = base_det(vSignal, dDSSPoints)
% Signal size, Window width
ns = size(vSignal,1); w = dDSSPoints;
% Empty windows whole signal
temp = zeros(w,ceil(ns/w))+NaN;
% Fill in - over windows, min's per window
temp(1:ns) = vSignal; [m,h] = min(temp);
g = h>1 & h<w; % mins, not at borders
% calc minima positions with resp. to x-axis
h = w*(0: numel(h)-1)/h;
% get valid minima and intensities
m = m(g); h = h(g);
% interpolate
vBaseline = interp1(h,m,1:ns,'pchip');
```

Figure 2. Matlab code for baseline correction by piecewise cubic interpolation using a problem adequat segment width.

to remove disturbances by an inverse filtering process. Thereby we take the DSS signal - detected by the peak picking mentioned later on - as a reference $s_{ref}(t)$. The reference is considered to be a known signal, disturbed by some transformation (not considering noise effects). The signal $s(t)$ is deconvolved using the reference. An ideal reconstruction $s_{ideal}(t)$ is convolved (*) with the modified $s(t)$, subsequently. Under ideal conditions this leads to an improvement of signal resolution as shown in [5]. This procedure can be summarized very briefly as:

$$s_{comp}(t) = \frac{s(t) \cdot s_{ideal}(t)}{s_{ref}(t)}$$

the signal $s_{comp}(t)$ is the ideal spectrum of interest $s(t)$ i.e. without disturbances. Here we use the convolution theorem $FFT(f * g) = FFT(f) \cdot FFT(g)$ and add appropriate zero-padding procedures in the deconvolution and the convolution step. The deconvolution can be helpful to identify signals which are not completely resolved in the original measurement. The assumption of $s(t)$ being free of noise, is a critical point, which results in an increase of the noise level for real signals in general, therefore the deconvolution has not been used in this experiments. However it may be desirable if a large number of measurements of the same experiment are available to compensate artificial - noise related - peaks.

In standard NMR the further steps of metabolite identification and quantification are done (in general) manually by fitting a known metabolite spectrum against the signal, subtracting this pattern from the signal and repeating the prior steps until the given signal can be reliably reconstructed from the fitted patterns. This approach is very time consuming and subjective. Alternatively the data are binned, leading to a data reduction, the areas in the bins are calculated and these features of the bins are

fed into a Principal Component Analysis or another analysis method. Binning in general leads to a very strong data reduction, is difficult to parametrize adequately (e.g. width of the bins) and removes a lot of the resolution of the measurement system. To overcome this a curve fitting approach (called targeted profiling) was proposed recently in [6]. Thereby a set of Lorentzians with known positions and intensity proportions are fitted against the signal and a subsequent analysis is carried out on the coefficients of the superpositions of these Lorentzians only. This greatly improves the former approach of binning leading to a compact, reliable encoding of the NMR spectra. A critical point of this approach is the complex fitting procedure the complexity of which is linearly increasing with the number of tested patterns. Further the assumption of a Lorentzian peak model for a real NMR peak is subject of discussion in the NMR community. Typical - *real measurements* - show a non Lorentzian peak shape and a method which would be able to deal with the real shape would be more desirable and accurate with respect to experimental practice. Taking this into account we focus on a peak picking approach which explicitly looks for peaks having a shape similar to the DSS signal and combine this approach with the targeted profiling suggested in [6]. The approach is sketched in Figure 3 and can be summarized as follows: On the prepared signal, as mentioned before, we apply a hill climbing search [7] for potential local maxima. Thereby only those maxima are kept which have a comparable height with respect to the DSS signal. Further constraints such as minimal/maximal peak width/height can be added to reduce the number of false hits. A line spectrum is generated at the identified peak positions and intensities which is convolved with the reference signal (here the DSS peak). This signal is subtracted from the original signal and the process is iterated until no further peaks can be detected. The obtained final peak list consists of multiple, potentially overlapping, peaks which can be considered as an information preserving reduced representation of the original NMR signal. These peak lists are subsequently compared with respect to known simulated or real metabolite spectra as depicted in Figure 4. We calculate a list of peak center positions for the measurement, considering the middle of the half peak height for each peak. This list is compared to the detected peaks in the patterns. Here a tolerance of $0.005ppm$ is chosen. In each case a matching peak must be in a (size limited) range of the start/end positions of the metabolite peak pattern. A high match of the number of peaks from the test pattern (metabolite) with respect to the peak list in the measured

```

function [peakBin, peakMag] = hillClimbing(x, negMagThresh, posMagThresh)
% container with peak positions and intensities
peakBin = []; peakMag = [];
xLen = length(x); % signal length
minPeakMag = min(x); % minimal intensity in the signal
tempPeakMag = minPeakMag; tempPosMagThreshOffset = 0.0;

foundPeak = 0; % indicator for a starting peak
peakCount = 1; % peak counter
i = 1; % current index in the signal
slope = x(i+1)-x(i); % slope of the currently investigated region
while i < xLen-1 % while terminate
% scan positive slope (slope, position, indicator, temporary height)
[slope, i, foundPeak, tempPeakMag] = positive_slope_start
% temporarily store peak candidate
if x(i) > tempPeakMag % new potential peak maximum?
tempPeakBin = i; % position
tempPeakMag = x(i); % local maximum to compare with
end
% scan negative slope
[slope, i, bAddPeak] = negative_slope_start
if (bAddPeak) % negative slope search successful
peakBin(peakCount) = tempPeakBin; % store position
peakMag(peakCount) = tempPeakMag; % store magnitude
peakCount = peakCount+1;
foundPeak = 1;
tempPosMagThreshOffset = x(i);
end
end
return

```

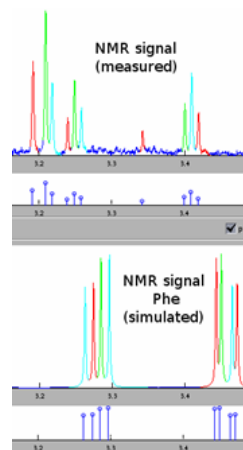


Figure 3. Pseudo code (simplified) for peak picking - hill climbing part - using problem adequate negative (0.0) and positive (90% of minimal peak height) magnitude thresholds (in acc. to DSS).

spectrum is an indicator that the pattern may be present in the signal. However further checks are needed to support this hypothesis, e.g. the intensity proportions between associated peaks (e.g. in a quartet) must be checked. For all identified patterns a quantification can be tried. Thereby the area under the matching peaks is calculated and associated to the area of the DSS signal, further a scaling by the number of protons of the DSS (9) with respect to the number of protons in the metabolite e.g. Ala (4) is done to obtain a concentration (c) in *m* mol:

$$c(\text{Ala}) = \frac{\text{area}(\text{Ala} - ^1\text{H}) \cdot c(\text{DSS}) \cdot 9}{\text{area}(\text{DSS} - ^1\text{H}) \cdot 4}$$

The identified and quantified metabolites are stored with further meta informations (e.g. preprocessing parameters) in an XML file, which can be considered as a metabolite model for the analysed data set. This metabolite model will be further subject of a pathway analysis to determine models for the chemical pathways of the cell estimated by the observed metabolite concentrations with respect to the growing medium conditions.

4. EXPERIMENTS

To test the methodology, we start with simulated spectra of the considered metabolites. Thereby for each metabolite a simulation was generated with an intensity value (*I*) of *I* = 30 for the spin system of the metabolite, with *I* = 1 for the DSS signal and *I* = 100 for the water signal. These data form the theoretical model for the metabolite spectra database in our experiments. The application of this model against itself (the simulations which build this model) results in a perfect recognition (100% peak match) and good intensity quantifications. In a next step a mixture of all considered metabolites was simulated at different concentration levels. Again a recognition of 100% was found. The results are good for the obtained quantifications but also some under/overestimations can be observed. A closer inspection reveals these effects caused by some overlapping of peaks. For example this effect can be observed for Myo-Inositol and Glycine which do not exactly share a common peak, but the glycine peak is very close to one of the triplets of Myo-Inositol. The same argumentation applies for Lactate which is close with its quartet to another Myo-Inositol-Triplet.

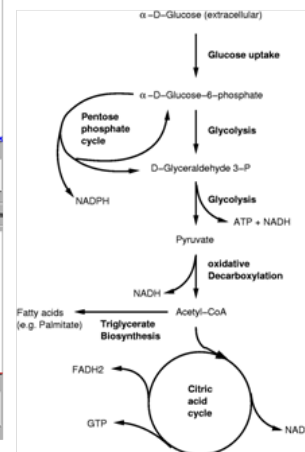


Figure 4. Identification of metabolites is based on comparison of the peak list (colored and as stems) obtained from the measured signal with respect to the peak list of a simulated metabolite (here Phenylalanine [Phe])(left). In general multiple metabolites can be detected in a spectrum. The area of the matching peaks can be used to calculate an estimate of the concentration of the metabolite. Results are serialized to an XML model which is further used to model a metabolic network (right).

After this initial experiments of pure synthetic data we take real measurements of the considered metabolite spectra. The results are depicted in Table 1. One observes that the metabolites could be identified in general, with the exception of Lac and Mal which were given in very low concentrations. To limit the effect of noise in the identifications only peaks with a minimal height of 1% of the DSS size are allowed, relaxing this criterion slightly solves this problem. However to avoid a large number of false hits and to keep detected signal intensities sufficiently above the noise level we will keep the minimal peak height of 1% of the DSS size for subsequently analyses.

In a next step data from the glucose experiment have been analysed, which are real biological data generated by metabolic process of FDCPmix cells on three types of growing media. Some results are depicted in Table 2. These results are very preliminary due to multiple reasons. For example the number of spectra for each condition is very small. Under this light the experiments should only be considered as an illustrative real life example how to use the presented system for such types of experiments. The results are also shown in Figure 5. As already mentioned the data support is very limited so it's hard to give any interpretation of the results, but considering the different graphs one may get the impression that for Alanine the third condition will cause a decreased expression. For succinate one may conclude that the conditions do not have any effect and for lactate a small increase of the concentration can be seen. These results have been checked by manual inspection and appear to be correct. However, as already mentioned, there are only few spectra supporting the data such that new measurements are necessary to prove this initial hypotheses. We also compared our findings with respect to an alternative method using the data analysis package Chenomx 5.0 [6]. Thereby we found a perfect agreement of the mean concentrations as depicted in Figure 5. However for some of the metabolites the esti-

Metabo (S)	Metabo (I)	PM	EC	QC
Ser	Ser	100%	0.50	0.73
Myo	Myo	93%	0.34	0.45
Lac	(Lac)	(100%)	0.07	(0.2)
Gly	Gly	100%	0.19	0.43
Suc	Suc	100%	0.25	0.41
Cit	Cit	100%	0.20	0.34
Mal	(Mal)	(100%)	0.27	(0.39)
Ala	Ala	71%	0.38	0.52

Table 1. Analysis of real measured pure metabolite data (S) with respect to a synthetic metabolite database (identification - I). At standard conditions (minimal peak height 1% of DSS almost all metabolites could be identified successfully. The experimental protocols note for *Lac* bad solving conditions. A closer (manual) inspection of the spectrum reveals, that the doublet of *Lac* has been detected but the quartet is very small and could not be detected. Without the quartet a concentration of 0.19 is quantified. A similar situation occurred for malate, lowering the minimal peak height of the peak picking algorithm to 0.5% of the DSS signal both metabolites can be detected (results in brackets). The percentage of matched peaks is given in the column (PM), the expected concentration in (EC) and the quantified concentration in (QC).

mations of Chenomx appear to be unlikely due to artificial fittings, not sufficiently supported by the analysed data. These effects have not been found using a peak based approach, because fits are only tried for identified peaks.

Condition	Metabolite	PM	QC (mean/std)
1	Ala	71 – 85%	0.8/0.47
1	Gly	100%	1.25/0.49
1	Lac	71%	0.2/0.28
1	Suc	100%	0.07/0.1
2	Ala	71 – 100%	0.89/0.7
2	Gly	100%	1.8/0.65
2	Lac	85%	0.6/0.48
2	Suc	100%	0.17/0.09
3	Ala	71 – 85%	0.54/0.37
3	Gly	100%	0.74/0.51
3	Lac	71 – 100%	0.73/0.08
3	Suc	100%	0.09/0.1

Table 2. Analysis of real measured extracts of growing media with FDCPmix cells. Only those metabolites are shown which are frequent within the specific conditions (Ala,Gly,Lac,Suc). Concentrations are given as mean concentration values over multiple spectra for a metabolite in a condition. Condition 1 accounts for a glucose level of 1mM (5 spectra), condition 2 accounts for glucose of 5mM (6 spectra) and the last condition for a glucose level of 25mM (4 spectra). If a metabolite has not been detected its concentration is assumed as 0.0 in the calculations (this is in general correct - verified by manual inspection). PM and QC like in Table 1

5. DISCUSSION AND CONCLUSIONS

We presented a system for the automatic identification and quantification of metabolites from $^1\text{H-NMR}$ -measurements. The approach is based on peak lists gener-

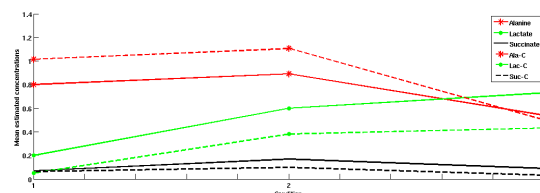


Figure 5. Mean estimated concentration values calculated for the three conditions using the prominent metabolites Ala (*), Lac (o) and Suc. The results are compared with an analysis done using Chenomx 5.0 (dashed lines - scaled)

ated by a hill climbing peak picker combined with a measurement specific shape subtraction. This allows a sensitive and measurement specific detection of peaks, which is in general more appropriate than a modeling with a Lorentzian (only) peak assumption. Another advantage of our method is the automatic reliability estimation of the identifications such that false positives are reduced, because it is only using the identified peaks and does not fit arbitrary metabolites against the signal.

We have shown that the method can be successfully applied on simulated spectra, real pure metabolite spectra and real experimental NMR spectra obtained from growing medium experiments. Beside of these positive aspects there are also some remaining challenges. First, the shape modeling, which is currently based on the DSS signal could be made more general by use of e.g. a wavelet based fitting procedure applied on multiple peaks, this would reduce the effect of noise or artifacts which may be present on the DSS reference and interfere the subsequently peak detection. Further the quantified concentrations are strongly affected by overlapping peaks. A rule based, knowledge driven, correction of plain area calculations may be desirable. Constraints on the accounted peak areas with respect to concurrent metabolites with potential fuzzy peak sets could be an interesting option to get more reliable estimates. In a next step the initial results must be verified by a larger amount of measurements combined with a modeling of the chemical reactions which also may be helpful to improve the metabolite model¹.

6. REFERENCES

- [1] S. Smith, T. Levante, B. Meier, and R. Ernst, "Computer simulations in magnetic resonance. an object oriented programming approach," *J. Magn. Reson.*, vol. 106a, pp. 75–105, 1994.
- [2] V. Govindaraju, K. Young, and A. A. Maudsley, "Proton NMR chemical shifts and coupling constants for brain metabolism," *NMR in Biomedicine*, vol. 13, pp. 129–153, 2000.
- [3] O. Kan, A. D. Whetton, and C. Heyworth, "Development of haemopoietic cells in liquid culture," in *Haemopoiesis: a Practical Approach*, N. Testa and G. Molineaux, Eds., pp. 123–137. IRL Press, Oxford, 1993.
- [4] L. Chen, Z. W. an Laiyong Goh, and M. Garland, "An efficient algorithm for automatic phase correction of nmr spectra based on entropy minimization," *Journal of Magnetic Resonance*, vol. 158, no. 1-2, pp. 164–168, 2002.
- [5] K. R. Metz, M. M. Lam, and A. G. Webb, "Reference deconvolution: A Simple and Effective Method for Resolution Enhancement in Nuclear Magnetic Resonance Spectroscopy," *NMR in Biomedicine*, vol. 13, pp. 129–153, 2000.
- [6] A. M. Weljie, J. Newton, P. Mercier, E. Carlson, and C. M. Slupsky, "Targeted profiling: Quantitative analysis of ^1H nmr metabolomics data," *Anal. Chem.*, vol. in press, pp. 4430–4442, 2006.
- [7] T. H. Park, *Towards Automatic Musical Instrument Timbre Recognition*, Ph.D. thesis, Princeton University, 2004.

¹ACKNOWLEDGMENT: We are grateful to C. Wierling at MPI f. Molecular Genetics and the whole MetaSTEM team. This work was supported by the Federal Ministry of Education and Research under PNR:934000-545, in the Project NMR Metabolic Profiling of the Stem Cell Niche (MetaSTEM).