EFFICIENT IDENTIFICATION AND QUANTIFICATION OF METABOLITES IN ^{1}H NMR MEASUREMENTS BY A NOVEL DATA ENCODING APPROACH

F.-M. Schleif[1], T. Riemer[2], U. Boerner[2], L. Schnapka-Hille[3] and M. Cross[3]

[1]University of Bielefeld, CITEC, Universitätsstrasse 2123, Bielefeld 33615, Germany
[2]University of Leipzig, Medical Physics and Biophysics, Härtelstrasse 16-18, 04107 Leipzig, Germany
[3]University Clinics, Hematology, Liebigstrasse 21, 04103 Leipzig, Germany
schleif@informatik.uni-leipzig.de,{riemer,uboerner}@uni-leipzig.de,
{michael.cross,lydia.schnapka}@medizin.uni-leipzig.de

ABSTRACT

The analysis of metabolic processes is becoming increasingly important to our understanding of complex biological systems and disease states. Nuclear magnetic resonance (NMR) spectroscopy is a particularly relevant technology in this respect, since the NMR signals provide a quantitative measure of metabolite concentrations. However, due to the complexity of the spectra typical of biological samples, the demands of clinical and high throughput analysis will only be fully met by a system capable of reliable, automatic processing of the spectra. We present here a novel data representation strategy for the measured spectra which simplifies the pre-processing of the data and supports the automatic identication and quantification of metabolites. The approach is combined with an extended targeted profiling strategy to allow the highly automated processing of ^{1}H NMR spectra, generating readouts suitable for the derivation of system biological models. The parallel application of both manual expert analysis and the automated approach to ^{1}H NMR spectra obtained from stem cell extracts shows that the results obtained are highly comparable. Use of the automated system therefore significantly reduces the effort normally associated with manual processing and paves the way for reliable, high throughput analysis of complex NMR spectra.

1. INTRODUCTION

The quantitative profiling of metabolites and the mathematical modelling of metabolic networks is expected to make a major contribution to our understanding of complex biological systems, including the processes underlying development and tissue homeostasis [1].

The most commonly used methods for metabolite detection are mass spectrometry (MS) and NMR spectroscopy, while each has its specific advantages, the inherently quantitative nature of NMR makes it most attractive for providing data for the development of mathematical models. However, the current challenge is to extract reliably quantitative data from experimental spectra which are often complex and subject to background variability. The general strategy involves pre-processing steps such as phase- and baseline-correction, smoothing and data reduction [2, 3], followed by the identification of distinct metabolite signatures in the signal and the estimation of metabolite concentration with respect to the original biological samples. A number of approaches have been reported to help solve these problems [4, 5, 6]. However, none of the methods currently available is sufficient to be applied in a reliable, automated fashion necessary for the high-throughput processing of complex biological samples [7]. We present here an approach designed to improve this situation by semi-automatic analysis of the spectra such that only minor, simple interaction steps are necessary to allow the processing of large data sets. We first provide a basic introduction to NMR spectra analysis, and then review the recently published approach of Targeted Profiling (TP) [6], which will be extended in this work. Initial results from stem/progenitor cell extracts are provided to document the improved performance of the present approach compared to manual expert analysis.

2. METABOLIC PROFILING BY NMR

We focus on the analysis of ^{1}H liquid NMR spectra obtained from stem/progenitor cell extracts. For the purpose of the analysis it is assumed that the chemical preprocessing and experimental design follows roughly the guidelines in [6]. The obtained spectra consist of $2^{16} = 65536$ measurement points per spectrum spanning a frequency range of about 8389.2 Hz corresponding to 11.982 ppm on a chemical shift axis for a ^{1}H resonance frequency of 700.15 MHz. Hence the spectral resolution is approximately 0.0002 ppm. A preprocessed sample spectrum is depicted in Figure 1. High resolution ^{1}H NMR spectra consist of a large number of relevant signals. Metabolite signatures are represented in general by multiple narrow peaks located on top of a wide underlying complex baseline. The NMR spectrum $s(\nu)$ can be approximated as a super composition of Lorentzians [8] but also Gaussian functions or mixtures thereof are common. However, this setting is highly idealized and in practical measurements the line shape of the peaks is much more complex and inhomogeneous due to measurement imperfections. The unknown line shape generates multiple challenges in the analysis. Almost all relevant signals in the NMR mea-

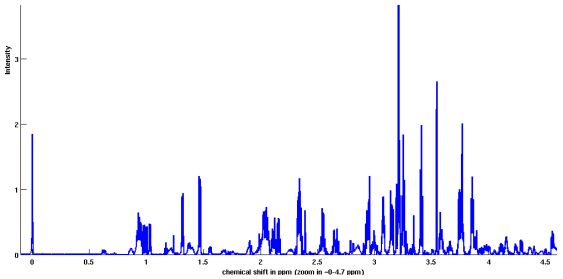

Figure 1. ^{1}H NMR spectrum from a stem cell extract. It was phased, baseline corrected, the water peak removed and global shift corrected with respect to a CSI.

surement show strong overlapping components. Without an appropriate model of the signal structure a deconvolution is extremely complicated. This is especially true for signal components at low concentrations which may otherwise be easily overlooked. The Targeted Profiling (TP) approach [6] assumes not only an ideal situation but also that the number of candidate signatures in the mixture $s(\nu)$ is small and restricted to a specific subset of known metabolites. For the set of known metabolites (targets; e.g. the metabolite Alanine (Ala), HOOC-CHNH_2-CH_3) the peak sequence of a plain measurement is known beforehand and constructed (manually) by adding appropriate peaks at the correct chemical shift position given in ppm.

A more formal definition for the target signal Alanine $f(\nu)$ is $f(\nu) = \sum_j^G g_j(\nu)$ with $g_j(\nu)$ as a peak pattern (e.g. a quartet) with appropriate settings for $g_j(\nu)$ as described below. An alternative compact description of Alanine is given by its ^{1}H NMR spin system classification as A_3X spin system with the associated values for the chemical shifts of $A = 1.46$ ppm, $X = 3.76$ ppm and $A - X$ coupling constant of 7.2 Hz. The known peak sequence information (signature) for the metabolites can be used to analyze the signal with respect to these signatures employing a simple least squares fit. While this approach is quite promising, fast and efficient [6] it suffers from multiple underestimated problems. The main problem comes with the target itself. In TP it is assumed that the signature of the target is perfectly known and can be observed in the signal. This, however, is very often not the case: (1) Due to variations in the measurement, (e.g. temperature, pH) the positions of the sub patterns in a target (groups of peaks) may shift in a non-linear manner. (2) A specific line shape has to be chosen for the fitting of the targets against the signal. In general it will be a Lorentzian or a Gaussian line shape. This however is a more or less good assumption which leads to further problems especially for strongly overlapping signals as depicted in Figure 2. (3) The simple fit of individual targets against the signal $s(\nu)$ may fail for strong overlapping structures and also incorrect identifications of targets are common because the fit is not constrained enough.

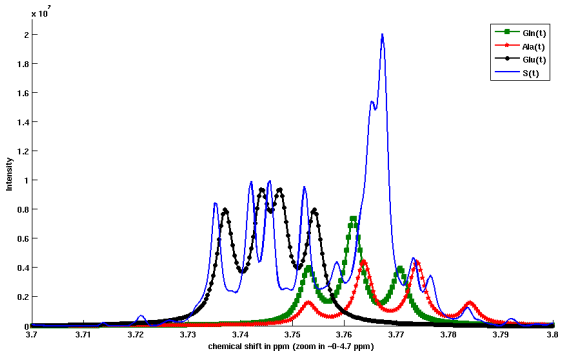

Figure 2. Overlapping effect within a preprocessed ^{1}H NMR spectrum with multiple metabolite signatures. It is obvious, that the assumption of the Lorentzian fails in parts to provide a sufficient approximation. This can lead to wrong estimates of target heights and its concentrations.

3. EXTENDED TARGETED PROFILING

The phased and baseline corrected signal is better approximated by Eq. (1).

$$s(\nu) = (\sum_j^J \alpha_j f_j(\nu - o)) + \epsilon \quad (1)$$

$$f_j(\nu) = \sum_i^G g_i(\nu - \Delta_i) \quad (2)$$

$$g_i(\nu) = \sum_k^K \Theta_k(\nu) \otimes \wp(\nu) \quad \text{e.g. } \wp = \exp(\ldots) \quad (3)$$

We employ a non-negative Least Squares Fit over all identified targets $f_j(\nu)$ with respect to the signal $s(\nu)$ using the functional encoding and the subsequently generated peak information. $s(\nu)$ is expected to be a linear combination of the the targets $f_j(\nu)$ with J as the number of targets. Here, o can be considered as a global shift which can be compensated by a reference shift correction and ϵ representing noise. The target f_j can be approximated as a super composition of its component functions g_i defined by the number G of chemical shifts in the molecules spin system. Thereby for each chemical shift and peak group g_i within the spin-system a small local shift $-\gamma \leq \Delta_i \leq +\gamma$ within a range of typically $|\gamma| \leq 0.005$ ppm can be expected. Each component $\Theta_k(\nu)$ of $g_i(\nu)$ can be considered as a delta function, contributing to a line spectrum with non vanishing amplitude for one peak position only. K is the multiplicity of a component function g_i. The origin of the chemical shift group components $\Theta_k(\nu)$ lies in the spin-spin interaction characterized by the so called scalar coupling constant and can be deduced from quantum mechanical calculations for the spin system parameters describing the targeted metabolite. Subsequently this line spectrum is folded $\otimes$ by a line shape function $\wp$ to mimic the real measurement's lineshape.

In NMR the position of the sub patterns or peaks are known as chemical shifts. The estimates of these shift positions need to be as accurate as possible and are the

main error-source in the TP approach. An appropriate peak shape estimate is the key to get a suitable subtraction of signal components from $s(\nu)$ in order to reveal potentially hidden components. In an initial step our approach takes the shape of the chemical shift identifier (CSI), a reference compound added to the sample, as a template. This shape is used to estimate the expected peak width present in the signal and is used to quantify identified targets.

To overcome the shift problem we estimate values for the disturbances Δ shown in Eq. (1) and present an initial solution to optimize the sub pattern positions in potential targets using a grid search strategy. This approach leads to in general improved position estimates for the *true* chemical shifts of the sub-patterns g_i of potential targets f_j and hence to more accurate identification and quantification estimates as shown later on.

As pointed out before, standard TP identifies signatures in NMR mixtures by employing known database references of (manually) specified peak patterns. In our extended Targeted Profiling approach (ETP) we modify this concept such that the targets are modeled on the theoretical spin-system model [9]. This model provides the peak information as transition tables. The target's spin system description acts as a highly accurate physical model that provides very accurate peak lists and can deal with varying NMR spectrometer field strengths easily. The parameters of the targets are optimized with respect to the measurement at hand.

Each target description T (generating a signal $f_j(\nu)$) is characterized by a set of spin-system descriptors $T_d \in S$. S describes the theoretical aspects of the spin system of T and can be used in combination with a model of the measurement system (NMR system) to simulate the spectrum f_j for T. A spectrum representation of T can be divided into multiple parts, one for each spin-system descriptor T_d, the peak group (g). A peak group may consist of multiple or a single peak and is potentially overlapping e.g due to the measurement resolutions. For each group a potential (limited) shifting error Δ_i can be expected. We now detail the three steps of ETP (line spectrum representation, peak assignment and shift estimation and non negative least squares fit) to obtain an optimized fit on this new encoding of the spectrum.

3.1. Line representation of a NMR spectrum

NMR spectra can be described by means of a set of overlapping peaks. To generate such a list of peaks, an appropriate model of the peak shape is necessary. In general the peak shape is assumed to be Gaussian such that a single peak can be represented by the following equation $\wp(\nu) = \exp(-(\frac{t-\mu}{\sigma})^2/2)$ with μ as a center position and σ as the line width. Also a Lorenzian peak shape is commonly used as provided before. A further implicit assumption is that the peak shape is symmetric and that the model is sufficient, e.g. is no super composition of gaussians or Lorentzians. In real measurements these assumptions are only partially fullfilled and a more complex peak shape is observed. This makes the peak picking rather complicated and so far different heuristic approaches have been proposed [8, 10]. Here we focus on a simple parametric hill-climbing approach. We further assume that for each measurement a known reference signal (CSI) is available, in our case this is the Tri-Methyl-Silyl-singulett from DSS signal[1]. This signal has a known position of 0 ppm, which is used to compensate the global shift offset o of the spectrum. At the expected DSS position we search for a maximum within a window of $0.05ppm$. From this position we go down (to lower intensities) on the left and the right flank of the peak as long as the signal is monotone decreasing. At a predefined maximal width the peak is truncated. For this peak its center position is calculated and the peak width at half maximum (PWHM). The PWHM is used as a rough estimate of the peak width. Due to effects such as imperfect phasing, shimming or baseline correction a direct inverse deconvolution of $s(\nu)$ with the CSI reference is in general not possible. Instead we employ a hill-climbing algorithm and look for local maxima in the whole signal which are above a predefined threshold (expected noise level). Additional criteria are signal flanks that are sufficiently steep and a sufficient peak width. By application of this algorithm we obtain a list of peaks in the spectrum. This list is subtracted from $s(\nu)$ and the algorithm is repeated until no further peaks are detected. Using this approach also overlapping peaks can be detected. As an alternative strategy the approach in [8] can be used with an underlying Lorentzian support. The list of peaks is subsequently denoted as $\mathcal{P}$. These peak lists are compared to the potential targets and their peak lists. If a sufficient amount of peaks (e.g. 30%) in a target (with a tolerance of $0.01ppm$) can be matched to the peaks $\mathcal{P}$ we consider the target as identified and proceed with the analysis steps for this target. Using the target description as mentioned before we can generate a simulated peak list for this target employing the gamma simulation environment [9]. We now have the target as a functional line spectrum $f_j(\nu)$ with $\wp$ as the fitted line function mentioned above.

3.2. Peak assignment and shift estimation

In a first step the peak list $\mathcal{P}$ can be filtered such that only those peak positions remain in $\mathcal{P}$ which are part of the peak lists of the targets using a rough shift tolerance of e.g. $0.05ppm$. Now an assignment matrix $M = n \times m$ is generated with n as the number of peaks over all target peaks and m as the number of peaks in $\mathcal{P}$. Thereby multiple assignments are possible and the shift-error of the peak with respect to the expected peak position is stored. Further only such assignments take place which are within a predefined tolerance $0.01ppm$. After this step a voting scheme is applied to M such that a maximal coverage of the target peaks with a minimal error with respect to the shifts is obtained. Hereby it is also ensured that a shift Δ_i applies only to a whole group g_i. The distance between two peaks within g_i is rather stable and determined by the quantum mechanical calculation of the spin systems coupling constants. Subsequently one obtains shifts ≥ 0 for

[1] 2,2-Dimethyl-2-silapentane-5-sulfonic acid. Alternative choices for the CSI such as TSP or ETH are possible as well.

each target T and each group g within a target. The optimized target simulations and peak descriptions can now be used in the fitting approach.

3.3. Non negative linear Least Squares Fit

The targets $f_j(\nu)$ are now given in the functional description of (2) with optimized Δ_i, using known Θ_k and our shape estimation for all $g_i(\nu)$. We can generate a reduced representation of each target and define a design matrix for the non negative linear Least Squares Fit (NNLSQ). The function to fit, is our spectrum $s(\nu)$ reduced to the position Θ_k. We add constraints for non negative α_i and allow user defined fixed α_j on some target f_j. Solving the optimization problem provides the α_j estimates used to calculate the concentration estimates as shown in [3].

4. EXPERIMENTS AND RESULTS

We analyzed our approach using different measurements of metabolites in growing media. Here, the focus is not on a specific biochemical question rather then to show that the identified and quantified metabolite concentrations are very close to the expert findings using ETP. Details about the data are shown in Table 1 and Figure 3. One clearly observes that the optimized approach provides results which are much closer to the expert analysis. In parts TP was not able at all to provide a concentration estimate because the shift error lead to unidentified metabolites.

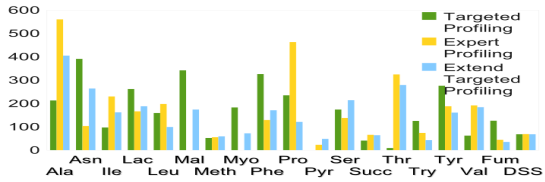

Figure 3. Concentrations for different metabolites using ETP compared to TP and an expert analysis (Spec_1).

Spec.	Err. TP	Err. ETP	Spec.	Err. TP	Err. ETP
1	49.65	37.57	4	87.53	46.01
2	68.68	57.55	5	64.04	44.06
3	30.92	31.41	6	111.09	86.94

Table 1. Mean errors in μ-mol of TP and ETP with respect to the expert concentration estimates. The expert concentration is assumed to be optimal (0 error), the values for TP and ETP are compared with the expert using the mean square error, normalized by J. One observed that the new approach strongly improves the concentration estimates.

In Figure 4 a reconstruction of a signal part is shown with respect to the original signal to illustrate the effect of the shift correction.

5. CONCLUSION

We presented a semi-automatic approach, called, *Extended Targeted Profiling*, for the identification and quantification of metabolites in NMR measurements by excessive use of a physical simulation model and functional description. Initial results are already quite promising and it could be shown that the new approach is beneficial with respect to traditional techniques. It can simplify the metabolite profiling task and is more flexible due to the formal model. For very high overlapping signals our approach still shows potential for improvements e.g. by optimizing the Δ and $\wp$ estimates to further reduce manual interactions.

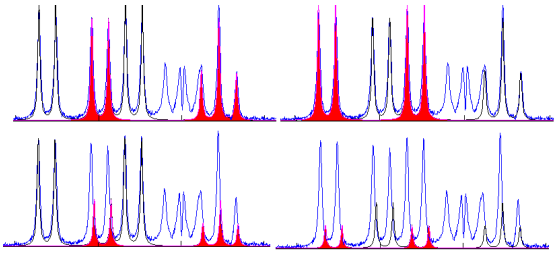

Figure 4. Spectrum in the region of Valine (Val) and Iso-Leucine (Ile). Two top figures show the ETP fit (filled), left Ile, right Val. Below the same region but fitted by TP.

6. ACKNOWLEDGMENTS

This work was supported by the Fed. Ministry of Edu. and Res.:FZ:0313833 A, (NMR Metabolic Profiling of the Stem Cell Niche, METASTEM) and the German Res. Fund. (DFG), HA2719/4-1 (Relevance Learning for Temporal Neural Maps). We would like to thank Prof. Thomas Villmann (Univ. of Appl. Sc. Mittweida) for discussions on sparse approximation and functional signal processing and the whole METASTEM team.

7. REFERENCES

[1] M. Cross, R. Alt, and D. Niederwieser, "The case for a metabolic stem cell niche," *Cells Tissue Organs*, vol. 188, no. 1-2, pp. 150–159, 2008.

[2] Y. Xi and D. M. Rocke, "Baseline correction for nmr spectroscopic metabolomics data analysis," *BMC Bioinf.*, vol. 9, pp. 324–333, 2008.

[3] F.-M. Schleif, T. Riemer, M. Cross, and T. Villmann, "Automatic identification and quantification of metabolites in h-nmr measurements," in *WCSB 2008*, 2008, pp. 165–168.

[4] J. Xia, T. C. B. abd P. Tang, and D. S. Wishart, "Metabominer semi-automated identification of metabolites from 2d nmr spectra of complex biofluids," *BMC Bioinf.*, vol. 9, pp. 507–522, 2008.

[5] Q. Zhao, R. Stoyanova, S. Du, P. Sajda, and T. R. Brown, "Hires a tool for comprehensive assessment and interpretation of metabolomic data," *Bioinf.*, vol. 22, no. 20, pp. 2562–2564, 2006.

[6] A. M. Weljie, J. Newton, P. Mercier, E. Carlson, and C. M. Slupsky, "Targeted profiling: Quantitative analysis of 1h nmr metabolomics data," *Analytical Chemistry*, vol. 78, pp. 4430–4442, 2006.

[7] S. Moco, R. J. Bino, R. C. D. Vos, and J. Vervoort, "Metabolomics technologies and metabolite identification," *Trends in Analytical Chemistry*, vol. 26, no. 9, pp. 855–866, 2007.

[8] H.-W. Koh, J. Lambert, S. Maddula, R. Hergenrder, and L. Hildbrand, "Feature selection by lorentzian peak reconstruction for 1-h nmr post processing," in *Proc. of CBMS 2008*. 2008, pp. 608–613, IEEE Press.

[9] 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.

[10] G. Brelstaff, M. Bicego, N. Culeddu, and M. Chessa, "Bag of peaks: interpretation of nmr spectrometry," *Bioinformatics*, vol. 25, no. 2, pp. 258–264, 2009.