BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–10

Genetic algorithm for shift-uncertainty correction in 1-D NMR based metabolite identifications and quantifications

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

[1]Univ. of Bielefeld, Dept. of CS, CITEC, Universitätsstrasse 21-23, 33615 Bielefeld, Germany
[2]Univ. of Leipzig, Dept. of Med. Phys. and Biophys., Härtelstrasse 16-18, 04107 Leipzig, Germany
[3]Univ. of Leipzig, Dept. of Hematology, Liebigstrasse 21, 04103 Leipzig, Germany

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: The analysis of metabolic processes is becoming increasingly important to our understanding of complex biological systems and disease states. Nuclear magnetic resonance spectroscopy (NMR) 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. An initial step in this direction has been taken by *Targeted Profiling* (TP), employing a set of known and predicted metabolite signatures fitted against the signal. However, an accurate fitting procedure for ^{1}H NMR data is complicated by shift uncertainties in the peak systems caused by measurement imperfections. These uncertainties have a large impact on the accuracy of identification and quantification and currently require compensation by very time consuming manual interactions. Here, we present an approach, termed *Extended Targeted Profiling* (ETP), that estimates shift uncertainties based on a genetic algorithm (GA) combined with a least squares optimization (LSQO). The estimated shifts are used to correct the known metabolite signatures leading to significantly improved identification and quantification. In this way, use of the automated system significantly reduces the effort normally associated with manual processing and paves the way for reliable, high throughput analysis of complex NMR spectra.

Results: The results indicate that using simultaneous shift uncertainty correction and least squares fitting significantly improves the identification and quantification results for ^{1}H NMR data in comparison to the standard targeted profiling approach and compares favorably with the results obtained by manual expert analysis. Preservation of the functional structure of the NMR spectra makes this approach more realistic than simple binning strategies.

Availability: The simulation descriptions and scripts employed are available under: http://139.18.218.40/metastemwww /bioinf/bioinf_suppl_nmr_ga_opt_schleif_et_al.tgz

Contact: schleif@informatik.uni-leipzig.de

1 INTRODUCTION

The quantitative profiling of metabolites and the mathematical modeling of metabolic networks is set to make a major contribution to our understanding of complex biological systems, including the processes underlying development and tissue homeostasis (Weckwerth (2003)). The most commonly used methods for metabolite detection are mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR). 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. Here we focus on the exact extraction of metabolite information from ^{1}H NMR measurements. The general strategy involves pre-processing steps such as phase- and baseline correction, smoothing and data reduction (Xi and Rocke (2008); Chang *et al.* (2007)), followed by the identification of distinct metabolite signatures in the signal and the estimation of metabolite concentrations with respect to the original biological samples. Details of the basic pre-processing used in this work are provided in (Schleif (2007); Schleif *et al.* (2008)). A number of approaches have been reported to help in the subsequent identification and quantification of individual metabolites from preprocessed data (Böcker *et al.* (2009); Xia *et al.* (2008); Zhao *et al.* (2006); Weljie *et al.* (2006)). However, none of the methods currently available can be applied in the reliable, automated fashion necessary for the high-throughput processing of complex biological samples (Moco *et al.* (2007); Mendes (2006)). As an initial step towards automatic processing, *targeted profiling* (TP) (Weljie *et al.* (2006)), employs a set of known and predicted metabolite signatures (targets) fitted against the signal. However, an accurate fitting procedure for ^{1}H NMR data is complicated by small but significant shift uncertainties in the peak systems, caused by even minor variations in parameters such as temperature and pH (Defernez and Colquhoun (2003)).

These uncertainties have a large impact on the accuracy of identification and quantification and currently need to be compensated by very time consuming manual interactions. Independent correction of the shift followed by fitting of the corrected target descriptions against the signals is not generally feasible because of the strong overlaps typical of ^{1}H NMR spectra.

Generic methods for the compensation of peak shifts are typically based on a specific or average reference signal taken from the data (Forshed *et al.* (2005)). If such a reference is available, then the NMR spectra are locally aligned to it such that the final set of spectra is reasonable aligned and corresponding peaks match. The used optimization techniques commonly employed include partial least squares approaches (Vogels *et al.* (1996)), genetic algorithms (Forshed *et al.* (2003)) and procedures based on the fourier transformation (Savorani *et al.* (2010)). This type of alignment

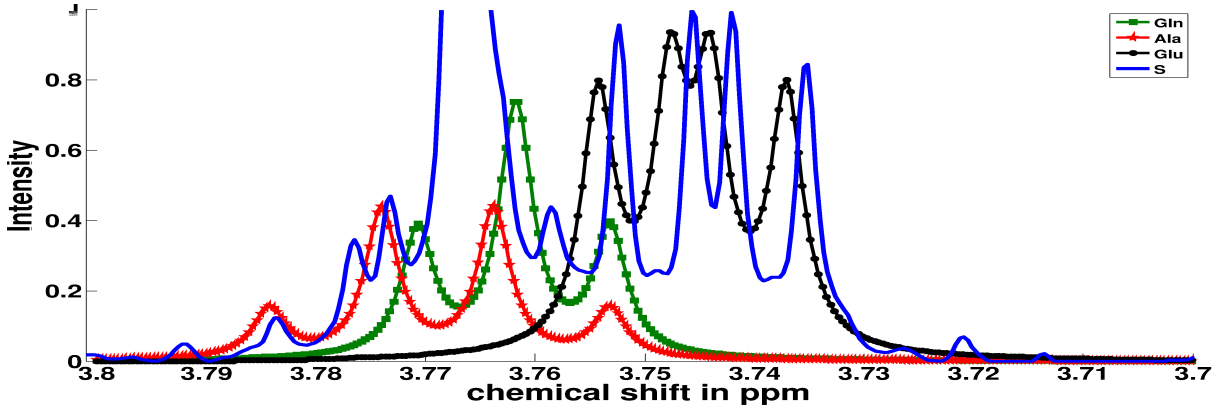


Fig. 1: Overlapping effect in a ^{1}H NMR spectrum of multiple metabolites. It can clearly be seen, that the assumption of the Lorentzian fails to provide an accurate approximation in some regions. This can lead to incorrect estimates of target heights and hence wrong concentration estimates.

problem is relevant not only to NMR but also to other data, including mass spectrometry (Pierce *et al.* (2007); Schleif (2006)). While the proposed approaches are promising and reasonable fast, they assume the availability of a reference spectrum to be used as the objective goal. Sometimes it is merely assumed that a set of common reference peaks is available so that an alignment function can be estimated based on these data (Schleif (2006)). However, this is often not realistic and in the setting considered here we do not assume the existence of a (global) reference spectrum. Furthermore, even for the aligned spectra one can not ensure that the peaks are aligned to their *true* position, only they are aligned to one another. If the chosen reference is not an undisturbed signal then there is no guarantee that the aligned spectra show correct ppm or mass positions for the peaks. In the case of metabolic profiling, this leaves the problem of correct identification and quantification of the metabolites in a spectrum with potential peak shifts. Our approach focuses on this special problem. The prior mentioned alignment methods can be used as a potential preprocessing only if the analyzed spectra are reasonable similar, as it should be the case for replicates. In this case it is possible to align the spectra first before using the approach, presented below.

The targets consist of a set of parametrized peak models showing uncertainties in their positions with respect to a true measurement, as described in more detail below. A typical NMR signal from a biological sample containing a variety of targets contains around 100 erroneous shift parameters. Local shift uncertainties need to be corrected within a given tolerance for all these parameters and often within the context of overlapping targets. Furthermore, NMR data show very spiked peaks so that both the correct peak positions and accurate target height estimates are decisive to the accuracy of metabolite concentration estimates. This makes a complete evaluation of all possible solutions unfeasible and the problem is ill posed.

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 developed an approach estimating shift uncertainties based on a genetic algorithm (GA) (Goldberg (1989); Mitchell (1995)) combined with a least squares optimization (LSQO) (Fletcher (2000)). Genetic algorithms are known to be very effective in finding local optimal solutions for ill-posed problems and have already been applied to spectroscopic data (Jarvis and Goodacre (2005); Hastie *et al.* (2001)). The estimated shifts are used to correct the known metabolite signatures, leading to significantly improved identification and quantification results. The shift uncertainties are generally corrected with sufficient accuracy that little or no subsequent manual interaction is necessary to generate the final quantifications. The method has been tested on a range of NMR spectra obtained from cell culture experiments. We have evaluated the models obtained in comparison to a standard targeted profiling approach as well as to the defacto standard of a careful manual analysis. We have also studied the observed shift uncertainties with respect to their influence on the concentration estimates during the multiple steps of the GA.

2 APPROACH AND METHODS

2.1 NMR Spectroscopy

All ^{1}H NMR-spectra were acquired on an AVANCE 700 MHz NMR-spectrometer (Bruker, Rheinstetten, D) equipped with a 5 mm cryo-probe. A pulse acquire sequence was used with 512 accumulations, 65536 complex points, 8389.2 Hz sweep width corresponding to 11.982 ppm on the chemical shift axis (0.002 ppm , 0.13 Hz nominal spectral resolution, respectively) and a repetition time of 20 seconds ($>$ five times the T1 of the reference and metabolites) ensuring fully relaxed, quantifiable signals. NMR samples were prepared by re-suspending lyophilised cell extracts in $500\mu l$ D_2O (99.9 atom %, Sigma Aldrich, Steinheim, D) potassium phosphate-buffer (0.05M, pH 7.4) containing a known concentration ($60 - 120\mu$M) of $2,2$'dimethylsilapentane-5-sulfonate (DSS, 99.0%, Fluka, Taufkirchen, Germany) as a reference for chemical shift and quantification. Each extract was then mixed vigorously by vortexing and centrifuged for 4 min at 10.000g. The supernatants (approx. 500μl) were transferred to 5 mm NMR-tubes (Wilmad, Vineland NJ USA). All samples were subject to NMR analysis at 298 K within 12 h.

2.2 Data pre-processing

We focus on the analysis of ^{1}H liquid NMR spectra obtained from extracts of cultured stem/progenitor cells, detailed subsequently. Each spectrum was preprocessed using in-house

Matlab (Mathworks Inc (2008)) routines. Spectra were phased, baseline corrected and referenced using DSS as a chemical shift and shape indicator (CSI) [1]. Furthermore, the region around $(4.5 - 5.9ppm)$ was set to zero for each spectrum to remove the water resonance contributions. Further details on the basic pre-processing are given in (Schleif (2007); Schleif *et al.* (2008)).

2.3 Data set description

We employed a set of 6 NMR spectra from cells cultured under a range of conditions to provide biologically realistic degrees of sample complexity and variation. The expected metabolites in the signal (subsequently referenced as targets) were: Alanine - (Ala), Asparagine - (Asn), Aspartate - (Asp), Citric Acid - (Cit), Cysteine - (Cys), Glutamate - (Glu), Glutamine - (Gln), Glycine - (Gly), Histidine - (His), Iso-Leucine - (Ile), Lactate - (Lac), Leucine - (Leu), Malate - (Mal), Methionine - (Meth), Myo-Inositol - (Myo), Phenyl-Alanine - (Phe), Proline - (Pro), Pyruvate - (Pyr), Serine - (Ser), Succinate - (Succ), Threonine - (Thr), Tryptophan - (Trp), Tyrosine - (Tyr), Valine - (Val), Fumarate - (Fum) and DSS as the standard reference. The signal is also expected to contain some unspecified metabolites.

The murine multipotent hematopoietic progenitor cell line FDCPmix (Factor Dependent Cells Paterson mixed potential) was grown in IMDM supplemented with 5 mM D-glucose, 2 mM L-glutamine, 1 mM sodium pyruvate, 20% horse serum and 10 u/ml IL-3. Six independent cultures were analysed, generated separately over a period of 18 months under the same culture conditions. The cells were maintained at $37\,^{\circ}\mathrm{C}$ in 5% CO_2 in air at densities between 6×10^4 and 5×10^5 cells per ml by passaging every $2 - 3$ days. At the final passage, the cells were transferred to fresh medium and cultured for 3 days. Between 1×10^8 and 2×10^8 cells from each experiment were harvested by centrifugation and washed four times with ice cold phosphate buffered saline (PBS) to remove medium constituents. The cell pellets were shock frozen in liquid nitrogen and extracts prepared by addition of $800\mu l$ ice cold methanol:acetonitrile:water $1 : 1 : 1$ mixture. To ensure efficient cell disruption the cells were subjected to 2×1 minute bursts of ultrasound in an ice cold ultrasonic bath. The samples were then transferred to a $70\,^{\circ}\mathrm{C}$ water bath for 10 minutes to denature the proteins before being diluted $1 : 7$ with water and lyophilized.

Additionally we analyzed a set of 4 spectra of wet-lab mixtures of the 5 metabolites (Ile,Leu,Glu,Val,Meth) and DSS as a standard with known concentrations.

2.4 Manual NMR expert analysis

The metabolites of interest were first measured individually by NMR to provide reference-spectra. A known concentration of the metabolite (1 - 20 mM) together with DSS ($0.1 - 2$ mM) was prepared in $500\mu l$ buffered D_2O solute (see 2.1) and measured under the same conditions as those used for the cell extracts. This allowed the determination of all chemical shifts (σ) and coupling constants (J) of each signal-generating metabolite proton as a basis for the reliable identification of metabolites in subsequent experiments.

Metabolite identification and quantification was achieved using purpose-developed NMR software (NMRj,Schlumm and Riemer (2001)) allowing for the interactive subtraction of a simulated from a measured NMR- spectrum. The chemical shifts and coupling constants from the simulation were carefully adjusted within a range of < 0.01 ppm to enable stringent fitting of the frequency pattern of the individual spin systems to the cell extract-spectrum. The criteria for an acceptable fit were firstly that all of the simulated peaks be present in the measured NMR-spectrum (i.e. identification of the metabolite) and secondly that the difference spectrum resulting from subtraction of the simulation from the measurement exhibited a smooth baseline at the position of metabolite frequencies. The latter step requires that the simulated signal is folded by a line broadening function that is as close as possible to that of the measured spectrum. This was achieved by using up to three exponential broadening functions, independent in amplitude, damping and frequency offset, for folding the simulated spectral time signal. Metabolite concentrations were calculated from the identified metabolite's NMR time-signal amplitude relative to the time signal amplitude of the known DSS reference concentration taking into account the relative number of contributing protons.

[1] Other choices for the CSI e.g. trimethylsilyl propionate (TSP) are also possible. The ideal CSI is only one peak with no overlap to other peaks.

2.5 NMR and targeted profiling

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 signal $s(\nu)$ can be approximated as a super composition of Lorentzians (Koh *et al.* (2008)), Gaussian functions or mixtures thereof. However, such assumptions are highly idealized. In practical measurements the line shape of the peaks is much more complex and inhomogeneous due to measurement imperfections. This poses multiple challenges in the analysis because almost all relevant signals in the NMR measurement show strong overlapping components. Without an appropriate model of the signal structure and line shape a deconvolution is extremely complicated. This is especially true for signal components at low concentrations which may otherwise be easily overlooked.

The TP approach (Weljie *et al.* (2006)) analyses metabolites by referencing to a set of known signatures. Taking some relatively strong assumptions concerning the line shape and knowledge about the structure of the targets, TP tries to identify and quantify these target metabolites in the complex NMR spectrum.

The TP approach assumes an almost perfect knowledge of the peak or line shape, which is typically modeled as a Lorentzian or a Gaussian function. It is also assumed, that the number of candidate signatures in the mixture $s(\nu)$ is small and restricted to a specific subset of known metabolites, the targets. Furthermore, it is assumed that for all targets, their peak sequence, i.e. the signal signature defined by the position and height of the peaks, is known perfectly beforehand. In practice it is often very difficult to provide such a description analytically for complex mixtures with extensive overlaps. For this reason the peak system is constructed (manually) by adding appropriate peaks at the correct ppm position and height. The targets are subsequently fitted against the measurement.

TP is being adopted as a standard technique in metabolite analysis and has already been employed in a number of studies see e.g. (Tiziani *et al.* (2009); Swire *et al.* (2009); Son *et al.* (2009)). While TP has been found to be very effective in a range of applications it remains suboptimal in many cases: (1) Due to variations in the measurement conditions (e.g. temperature, pH) the position of the g_i 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 candidate targets against the signal. Since the actual line shape may deviate from the chosen forms, this assumption can lead to further

problems especially for strongly overlapping signals as depicted in Figure 1. (3) The simple fit of individual targets against the signal $s(\nu)$ may fail for strongly overlapping structures, while the use of lower constraints on the fitting commonly leads to incorrect identifications of targets. In the later case it can happen that lines are fitted into regions without signal.

The TP approach also lacks the formal and mathematical derivation and modeling basis which would simplify adaptations, for instance to accommodate moderate changes in the device settings such as alternative measurement frequencies, or to incorporate alternative peak shape models.

In the following section we formalize targeted profiling and detail our extension thereof. We provide an appropriate mathematical modeling for the fitting and parameter estimation approach, taking the functional characteristic of the measurements into account.

3 EXTENDED TARGETED PROFILING

An arbitrary metabolite may formally be given by a *functional description* $f(\nu)$ for a target signal as $f(\nu) = \sum_j^G g_j(\nu)$ with $g_j(\nu)$ as a peak pattern or a function of delta functions with non-zero entries only on the appropriate peak positions as detailed below and G as the number of such peak patterns. Using the TP approach $f(\nu)$ may be folded with an appropriate line shape e.g. a Gaussian. A reconstruction of alanine using the functional description is given in Figure 2.

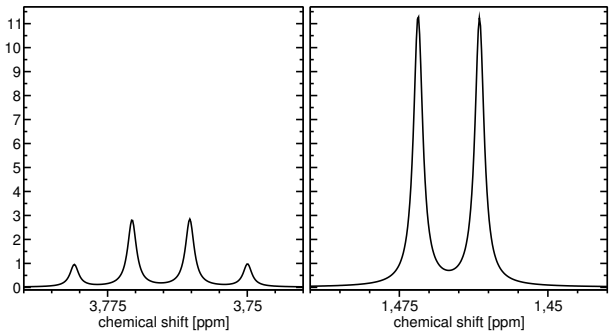


Fig. 2: Reconstruction of L-alanine using the functional description. The x-axis is given in ppm and the y-axis shows the intensities. (a): the quartet generated by the H_x proton with a shift parameter $\sigma(H_x)$ and (b): the doublet caused by the three magnetically equivalent H_A protons with shift parameter $\sigma(H_A)$.

An alternative compact description of a target e.g. alanine is given by its ^{1}H NMR *spin system classification* as A_3X spin system (see e.g. Levitt (2008)), with the associated values for the chemical shifts of $\sigma(H_A) = 1.46$ ppm, $\sigma(H_X) = 3.76$ ppm and an A-X coupling constant of $J_{AX} = 7.2$ Hz see Figure 3.

Using the above spin system classification, we can employ a NMR simulation environment (Smith *et al.* (1994)) to simulate the alanine spectrum whilst taking the physical properties of our measurement system (such as device frequency) into account.

This simulation yields transition tables providing information on the peak positions and heights of each peak for the target. A transition table for L-alanine is shown in Table 1.

From this line spectrum we can generate a profile spectrum, similar to a true measurement by folding the line spectrum with an

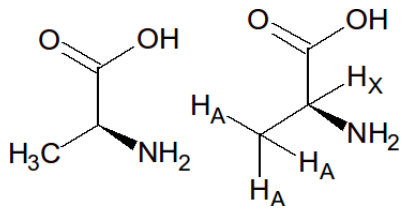


Fig. 3: Structure of L-alanine (left) and in A_3X notation (right).

Index	PPM	Intensity	Group index	A_3X
1	1.4596	11.2246	1	A_3
2	1.4700	11.2752	1	A_3
3	3.7499	0.9438	2	X
4	3.7603	2.8187	2	X
5	3.7706	2.8061	2	X
6	3.7809	0.9311	2	X

Table 1. Transition table providing the information for a line spectrum reconstruction of L-alanine. The table was generated using standard settings for a 700.153 MHz NMR system ^{1}H channel as specified before.

assumed line shape, leading to our functional description $f(\nu)$ of a given target (see Figure 2). Taking this approach we can model the, phased and baseline corrected signal $s(\nu)$ as

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

$$f_j(\nu) = \sum_i^{G_j} g_i(\nu - \Delta_i) \quad (2)$$

$$g_i(\nu) = \sum_k^{K_{j,i}} \Theta_k(\nu) \otimes \wp(\nu) \quad (3)$$

$$\wp = \text{e.g.}\exp(\ldots) \quad \text{line shape} \quad (4)$$

We employ a non-negative Least Squares Fit over all J identified targets $f_j(\nu)$ using the functional description and the subsequently generated peak information. Thereby o represents a global shift which can be compensated by a reference shift correction and ϵ represents noise. The target f_j can be approximated as a super composition of its component functions or peak groups g_i defined by the number G_j of chemical shifts in the molecule's spin system.

A small local shift $-\gamma \leq \Delta_i \leq +\gamma$ typically within a range of $|\gamma| \leq 0.005$ ppm can be expected for each peak group. 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. We denote such a single position ν as $\nu_{j,i,k}$ to specify peak k caused by group i in metabolite j. 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 scalar coupling constant J_{AX} and can be deduced from the quantum mechanical calculations for the spin system parameters describing the target metabolite. Subsequently this line spectrum is folded $\otimes$ by a line shape function $\wp$ to mimic the line shape of the real measurement. In the following we will use G for G_j and K for $K_{j,i}$ if the indices are known from the context. In NMR the position of the g_i 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 accurate peak shape estimate is the key to an appropriate subtraction of signal components from $s(\nu)$ in order to reveal potentially hidden components. We approach this issue by taking the shape of the DSS reference signal added to the sample, as a template for $\wp$. This shape is used to estimate the expected peak width present in the signal.

To tackle the shift-uncertainty problem, we estimate values for the disturbances Δ shown in Eq. (2) and present an initial solution to optimize the g_i positions in potential targets using a grid search strategy. This approach leads to a general improvement in position estimates for the *true* chemical shifts of the sub-patterns g_i of potential targets f_i and hence to more accurate identification and quantification estimates as shown below.

Whereas standard TP identifies signatures in NMR mixtures by employing known database references of (manually) specified peak patterns the Extended Targeted Profiling approach (ETP) described here modifies this concept by modeling the targets based on their theoretical spin-system model (see (Smith *et al.* (1994)). This model provides the peak information (transition tables). The physical model easily deals with measurement variables such as different device frequencies and is known to provide very accurate peak lists. The parameters of the targets are optimized with respect to the measurements 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, known as the peak group (g). A peak group may consist of multiple or single peaks and is potentially overlapping. For each group a potential (limited) shifting uncertainty Δ_i can be expected. New targets can be added to the ETP approach very easily by specifying the spin-system model, outlined above, based either on knowledge available in the literature or by own measurements of the pure target substance under the previously defined measurement conditions. In the latter case the obtained spectrum is analyzed manually to define the spin-system model. Hereby an NMR expert constructs a spin-system model such that the reconstructed spectrum, based on this model, fits best to the observed data. The three steps of ETP required to obtain an optimized fit based on this new encoding strategy are detailed below.

3.1 Line representation of a NMR spectrum

NMR spectra can be described by means of a set of overlapping peaks, which provides a compact representation of the signal and can also reveal quickly whether or not an expected target is likely to be present in $s(\nu)$, since all simulated target peaks must also be present in the peak list of $s(\nu)$. The peak picking process is rather complicated, and a number of heuristic approaches have been proposed to improve the situation (Koh *et al.* (2008); Brelstaff *et al.* (2009)). Here we focus on a simple parametric hill-climbing approach (Schleif *et al.* (2008)). We further assume that for each measurement a known CSI signal is available, in our case this is DSS. This signal has a known position of 0 ppm, which can be used to compensate the global shift offset of the spectrum. We look for a maximum within a window of $0.05ppm$ at the expected CSI position. From this position we then go down (to lower intensities) on the left and the right flank of the peak as long as the signal is a descending monotone. The peak is then truncated at a predefined maximal width. The center position and the peak width at half maximum (PWHM) are then calculated for this peak. 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 not generally feasible. Instead, we employ a hill-climbing algorithm and look above a predefined threshold (the expected noise level) through the whole signal for local maxima, whose flanks are sufficiently steep and for which the obtained peak has a sufficient width. By application of this algorithm we obtain a list of peaks in a spectrum. This list is subtracted from $s(\nu)$ and the algorithm is repeated until no further peaks are detected. This approach can also resolve peaks in an overlap, although not in every case. Alternatively, the strategy described in (Koh *et al.* (2008)) can be used with an underlying Lorentzian support, the particular peak picking algorithm is not of much relevance here as long as it discovers the peaks in the spectrum to a sufficient degree of accuracy. The list of peaks is subsequently denoted as $\mathcal{P}$. These peak lists are compared to those of the potential targets. If a sufficient number of peaks (e.g. 30%) in a target can be matched within a tolerance of $0.01ppm$ to the peaks $\mathcal{P}$ we consider the target to be identified and proceed with the analysis steps for this target. We now have the target as a functional line spectrum $f_j(\nu)$ with $\wp$ as the fitted line function.

3.2 Genetic algorithm for shift uncertainty estimation

A major feature of our approach is the shift uncertainty correction performed by means of a Genetic Algorithm (GA). The genetic algorithm software was written in-house in Matlab running on an Intel Xeon multiprocessor system with 8 3.20 GHz processors and 16 GB memory using the parallel processing, signal processing and optimization toolbox with Matlab 2008b. We made use of the GA implementation in the optimization toolbox but replaced some of the core methods with our own purpose-developed implementations. Specifically, we replaced the methods used to generate the initial population and the mutation function and provided a specific fitness function as described below. The basic algorithm and parameters for the GA are shown in Table 2 and the overall workflow is depicted in Figure 4. Briefly, we generate a large number of chromosomes P, each chromosome has the same length $Z = \sum_j^J G_j$, equal to the number of groups over all analyzed targets, and each of which contains the currently estimated, or randomly determined shift values for the Δ_i. These Δ_i are optimized by the GA. Furthermore the smallest shift is limited by the ppm-axis resolution. We have found that a reduction of the original spectrum to $16k$ points corresponding to 0.5 Hz spectral resolution, is possible, whilst maintaining structural information of sufficient quality for the subsequent identification task. In this case up to 25 valid shift positions are possible for a given shift uncertainty of $\pm 0.01ppm$ [2].

3.2.1 Fitness function and evaluation measures The fitness function is the core element of the GA and is evaluated for each single chromosome separately. It consists of three procedures: (1) The spin-system classifications of all identified metabolites are used to generate the corresponding spectral representation. Thereby, the shifts given by the chromosome are applied to the corresponding groups g_i, and the reconstructions folded with the prior estimated line shape. We denote the matrix of all reconstructions $f_j^*(\nu)$, given

[2] If we assume a spectral resolution of $SR = 0.5$ Hz, a device frequency of $F = 700$ MHz and an error of PPM-E$= \pm 0.01$ ppm the number of valid positions V is $V \approx 2 \cdot \text{PPM-E}/(SR/F)$.

Parameter	Description	Value
C	single chromosome	$c \in \mathbb{R}^Z$ $c_i \in [\text{-PPM-E,PPM-E}]$
M	set of chromosomes	$M = \{C_1, \ldots, C_P\}$
K	Number of generations	200
P	Number of chromosomes	900; $\|M\|$
p_i	permutation probability	0.1
Z	Length of the chromosomes	$\sum G_j$
PPM-E	PPM uncertainty of the Δ_i	0.01
SR	Down-sampling rate	4
d	distance measure	{0,1}

Table 2. Basic parameters of the genetic algorithm. The distance measure is either euclidean - 0, or a functional distance - 1.

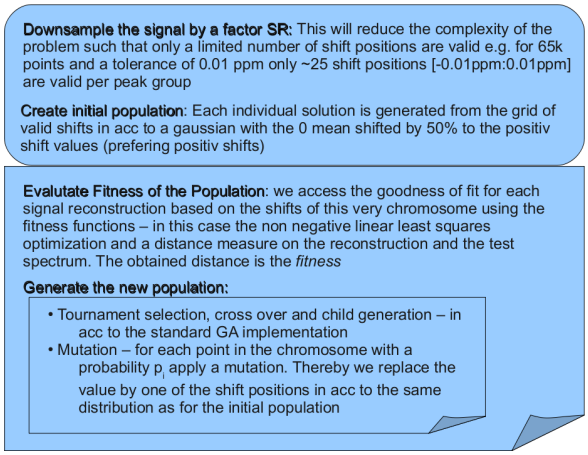


Fig. 4: Workflow of the genetic algorithm used to obtain optimal Δ_i. The first (outer) folded corner is repeated until K generations are analyzed or one of the alternative standard stopping criteria is met. The inner folded corner is repeated until a new population of the same size as before is generated.

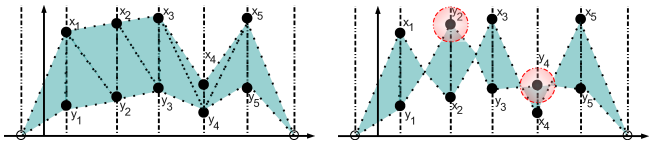
as row vectors, as the matrix R

$$R = \begin{pmatrix} f_1^*(\nu) \\ \ldots \\ f_J^*(\nu) \end{pmatrix} \quad R' = \begin{pmatrix} f_1^*(\bowtie) \\ \ldots \\ f_J^*(\bowtie) \end{pmatrix}.$$

(2) These reconstructions are reduced to a range representation such that a compact form of R denoted as R' is obtained. In R' not all values for ν are used but only a limited set of ν in form of potentially overlapping range vectors $\bowtie_l = [\nu_l - (2 \cdot PPM - E) : \nu_l + (2 \cdot PPM - E)]$ with l as an index of a peak positions in Υ.

We collect all peak center positions of the metabolites denoted as $\Upsilon = \{\Upsilon_1, \ldots, \Upsilon_J\}$ with $\Upsilon_j = \{\nu_{j,i,k}\}_{i,k}^{G \times K}$ and $\nu_l \in \Upsilon$. Here we also incorporate the Δ_i and take the peak positions from the transition tables extended to a range of twice the assumed ppm-uncertainty PPM-E for each peak. A reduced test spectrum $s'(\bowtie)$ is constructed, accordingly.

(3) The matrix R' and the vector $s'(\bowtie)$ are subsequently used in the LSQO, as the third step, to calculate the α_i for all targets. The reduction to a range based representation is useful to avoid very large and extremely sparse matrices, which would complicate the subsequent α estimations. This step has no detrimental effects on the $\alpha-$estimates.



(a) Two functions: Euc = L^p-norm (b) Two functions: Euc $\neq L^p$-norm

Fig. 5: Illustration of the L^p-norm. Plot (a) indicates the case in which the distance between two functions is equal, both for Euclidean or L^p-norm. In plot (b) parts of the functions are interchanging (crossing). The distance using Euc is still the same as in plot (a) but for the L^p-norm the distance is changed, giving a more realistic measure of the distance of the two functions.

3.3 Non negative least squares fitting

The targets $f_j'(\bowtie)$ are now given in the functional description of (2) with optimized Δ_i, using the known Θ_k and our functional shape estimation for all peak groups. The function to fit is our reduced spectrum $s'(\bowtie)$. We add constraints for non negative α_i and allow for user definition of α_j fixed on a target f_j by employing standard optimization modeling techniques. Solving the optimization problem by use of a standard constrained linear least squares algorithm we obtain the α_j in a column vector α, which can subsequently be used to calculate the concentration estimates. To this end, the area under the α-scaled target is calculated and associated to the area of the α-scaled reference signal (here DSS). A scaling step is then performed, based on the number of protons ^{1}H present in the reference, nine for DSS, compared to the number of protons present in the metabolite e.g. four for Ala. This leads to the following equation for the concentration c in mol: $c(Ala) = \frac{area(Ala) \cdot c(DSS) \cdot 9}{area(DSS) \cdot 4}$ with *area* as an appropriate estimation function for the area under the curve. One can also calculate estimates of the lower concentration limits by scaling the target intensities of f_j' to the noise level and repeating the procedure. The reconstruction s^* is obtained as:

$$s^* = R^\top \cdot \alpha \tag{5}$$

To judge the fitness of this solution we may now either use the quality of fit provided by the LSQO algorithm or evaluate the reconstructed spectrum $s^*(\nu)$ with respect to $s(\nu)$ using a problem specific distance measure. Here we use either the standard Euclidean distance (EUC) or a functional distance measure as an extension of the L^p norm proposed in (Lee and Verleysen (2005)) (FUNC). The functional distance measure has the advantage of taking the functional nature of the spectra into account. The standard Euclidean distance considers the individual features of the NMR spectrum to be independent, so that a change in the order of the ppm positions does not affect the calculated distance. However, the features or measurement points in NMR spectra are not independent, so that a distance taking this aspect into account can be considered to be more appropriate for this type of data. Lee proposed a distance measure taking the functional structure into account by involving the previous and next values of a signal v_i in the i-th term of the sum, instead of v_i alone. Assuming a constant sampling period τ, the proposed norm (FUNC) is:

$$\mathcal{L}_p^{fc}(\mathbf{v}) = \left(\sum_{k=1}^{D} \left(A_k(\mathbf{v}) + B_k(\mathbf{v}) \right)^p \right)^{\frac{1}{p}} \tag{6}$$

	Ile			Leu			Val			Glu			Meth		
	EXP	TP	ETP	EXP	TP	ETP	EXP	TP	ETP	EXP	TP	ETP	EXP	TP	ETP
M-DS$_1$	113.23	31.07	62.32	32.35	17.29	23.53	36.27	17.88	18.59	52.37	26.29	63.40	79.61	56.52	49.50
M-DS$_1$-C		72.64			37.77			21.85			38.65			57.57	
M-DS$_2$	36.32	26.00	73.98	62.96	15.00	21.43	32.78	15.98	28.52	30.92	22.30	35.07	71.97	47.94	52.41
M-DS$_2$-C		60.54			50.37			43.71			30.92			43.18	
M-DS$_3$	51.92	19.62	27.10	60.30	13.52	21.44	51.45	16.29	32.00	37.68	21.59	33.95	94.22	48.50	75.11
M-DS$_3$-C		36.32			62.96			32.78			30.92			71.97	
M-DS$_4$	57.98	21.27	36.01	58.85	16.55	25.94	34.33	14.50	11.46	31.58	30.73	50.48	115.01	61.03	84.51
M-DS$_4$-C		48.43			62.96			21.85			23.14			86.36	
$\otimes$		0.63	**0.46**		0.74	**0.60**		0.55	**0.43**		0.35	**0.17**		0.40	**0.27**

Table 3. Concentrations of metabolites in the synthetic wet-lab study. The weighted sample concentration is given in the -*C* rows each. The estimate of the expert EXP, TP and ETP are given in the columns. All concentrations are given in μ mol. Considering the median relative error ($\otimes$) of the concentration estimates the ETP approach is best in all cases. In a case by case comparison ETP is almost always the best, with three exceptions $\{(Spec_2, Val), (Spec_4, Glu), (Spec_1, Ile)\}$.

with

$$A_k(\mathbf{v}) = \begin{cases} \frac{\tau}{2}|v_k| & \text{if } 0 \leq v_k v_{k-1} \\ \frac{\tau}{2}\frac{v_k^2}{|v_k|+|v_{k-1}|} & \text{if } 0 > v_k v_{k-1} \end{cases} \quad (7)$$

$$B_k(\mathbf{v}) = \begin{cases} \frac{\tau}{2}|v_k| & \text{if } 0 \leq v_k v_{k+1} \\ \frac{\tau}{2}\frac{v_k^2}{|v_k|+|v_{k+1}|} & \text{if } 0 > v_k v_{k+1} \end{cases} \quad (8)$$

representing the triangles on the left and right sides of v_i and D being the data dimensionality. For the data considered in this paper v takes the position of ν. As for L_p, the value of p is assumed to be a positive integer. At the left and right extremes of the sequence, v_0 and v_D are assumed to be equal to zero. The concept of the L^p-norm is shown in Figure 5. The calculation of this norm is slightly more complex than that of the standard Euclidean but, as is shown below, significantly improves the fitting results as well as the convergence speed of the GA[3].

4 RESULTS AND DISCUSSION

4.1 Identification and quantification

We have tested our approach using measurements of metabolites in lysates of cultured cells as well as a small test set with known concentrations of defined metabolites. Rather than focusing on a specific biochemical question we aim to compare the range and concentrations of metabolites detected using TP and ETP with those obtained by manual expert profiling.

4.1.1 Wet lab metabolite mixture experiment The wet-lab mixture data sets (M-DS) can be considered to be an artificial data set with known concentrations. In Table 3 the known concentrations (weighted sample) and the concentration estimates as obtained by the expert, TP and ETP using the functional distance measure are given. We observe that the ETP approach is closer to the expert estimation than is TP. The DSS concentration was given with 77.05 μ mol for all spectra.

4.1.2 Cell culture experiment Details of the analyzed cell extract data are shown in Figure 6. The optimized approach provides results which are much closer to the expert analysis, for 16 of the 21 targets. Eth, Cit, His, Myo and Mal were noted as being absent by the expert but by ETP and TP with very low concentrations.

Test spectrum	Error TP	Error ETP (FUNC)	Error ETP (EUC)
Spec$_1$	49.65	31.95 (97)	32.05 (121)
Spec$_2$	68.68	45.89 (106)	68.71 (122)
Spec$_3$	30.92	28.40 (112)	29.87 (147)
Spec$_4$	87.53	55.70 (118)	56.01 (132)
Spec$_5$	64.04	47.30 (97)	46.39 (121)
Spec$_6$	111.09	87.60 (81)	93.77 (137).

Table 4. 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 then compared with the expert using the mean square error, normalized by the number of metabolites. It can be seen that the new approach clearly improves the concentration estimates. The number of generations until convergence is shown in brackets.

Figure 9 shows a reconstruction of a signal part with respect to the original signal to illustrate the effect of the shift correction.

From Table 4 we observe that the EUC measure in the fitness function is indeed less effective then the FUNC measure, consistent with our expectation that the FUNC norm is more appropriate to data which are themselves functions [4]. We subsequently restrict our analysis to the FUNC norm and the standard TP approach. In Figure 7 we show Box-Whisker Plots of the relative concentration errors with respect to the expert of the metabolite concentrations using the standard TP and the ETP (FUNC) approaches. It can be seen that the relative error of ETP is much smaller than that of TP in the large majority of the metabolites. Also the variance of the results is smaller. Median errors of TP vs ETP are shown in Table 5. Significance of an improvement is indicated by a $+$ using a t-test on a 5% level, with significances of $p \leq 0.03$. We observe that $\approx 25\%$ of the differences are significant and all of these are positive.

[3] The additional effort in the calculations is almost negligible - the time to calculate a generation is changing only minor, by a few seconds.

[4] Using the median in Figure 6 gives similar results with respect to TP but comparing FUNC and EUC the results are less pronounced due to the dominance of the (many) small metabolites

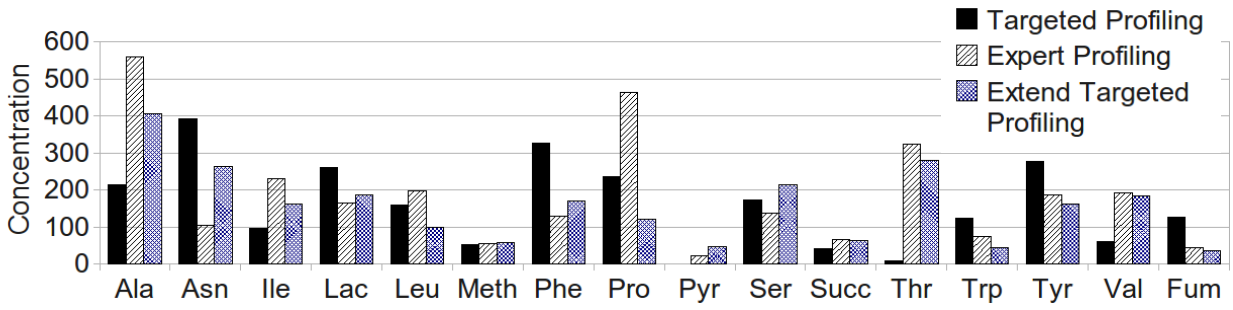


Fig. 6: Concentration estimates for some of the different metabolites using ETP in comparison to TP and an expert analysis ($Spec_1$). The x-axis denotes the metabolites and the y-axis the intensities in μ-mol.

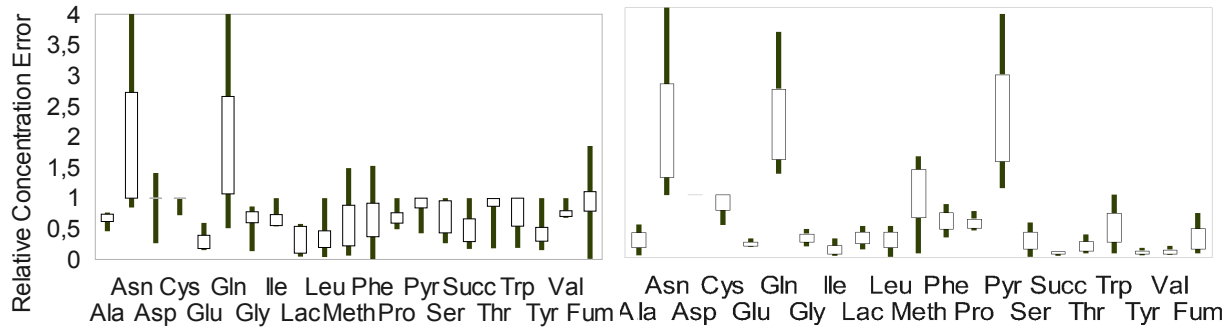


Fig. 7: Relative error estimates for different metabolites using TP (a) and ETP (b). The y-axis encodes the relative error and is limited to [0,4]. The x-axis lists the different metabolites.

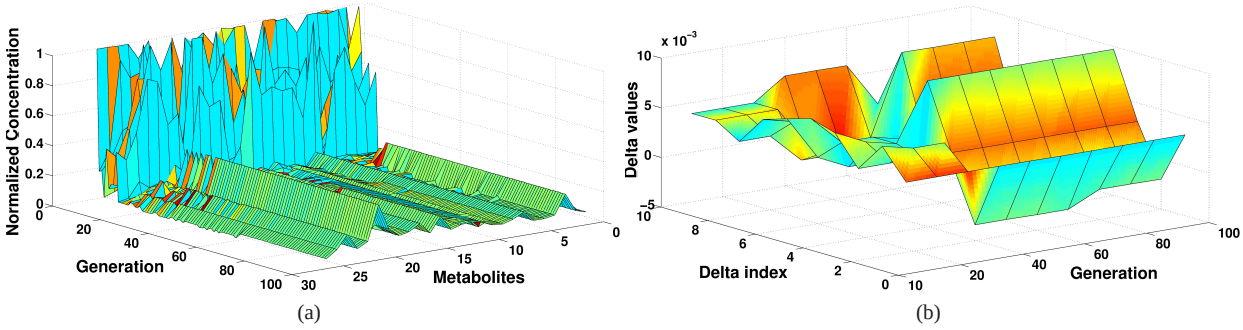


Fig. 8: (a): Concentrations (normalized) for metabolites over time (without CSI) for $Spec_1$. (b): parameter changes over time for $\Delta_1 \ldots \Delta_{10}$ using the median over 10 generations.

4.2 Shift uncertainty properties and influence

The shift uncertainty estimates Δ change over time with respect to the GA evolution and the underlying constraints. The GA can only determine a local optimal solution, which is expected to correspond to the global optimum in only a very few cases. An analysis of the number of updates per shift uncertainty estimate reveals parameters which are likely to be incorrect either because they have not been updated at all or because they have been updated very frequently. An example for $Spec_1$ is shown in Figure 11. Taking this statistic into account the expert can be assisted by an indicator that highlights peak group shifts, that are likely to be incorrect. Considering the values of the shift uncertainties for the analyzed spectra we found around $3-4\%$ of the Δ_i to be 0 after convergence. Analyzing the shift updates also provides information about potentially unreliable modeled regions, indicated by either very few or very many Δ updates, as shown in Figure 10. There the relative number of Δ

	Ala	Asn	Asp	Cys	Glu	Gln	Gly	Ile
TP	0.66	1.77	1	1	0.26	1.83	0.69	0.61
ETP	0.28	1.56	1	1	0.19	1.78	0.3	0.08
Imp	1	1	0	0	1	1	1	1
Sig	+	o	o	o	o	o	+	+
	Lac	Leu	Meth	Phe	Pro	Pyr	Ser	Succ
TP	0.34	0.27	0.49	0.56	0.64	1	0.64	0.44
ETP	0.32	0.31	1.21	0.58	0.5	1.96	0.26	0.09
Imp	1	1	2	2	1	2	1	1
Sig	o	o	o	o	o	o	o	o
	Thr	Trp	Tyr	Val	Fum	Mean		
TP	0.97	0.85	0.44	0.71	1	0.81		
ETP	0.14	0.42	0.07	0.06	0.21	0.66		
Imp	1	1	1	1	1	-		
Sig	+	o	o	+	o	0		

Table 5. Relative median metabolite error. Change judged in the row labeled by *Imp*: 1 (improvement/optimal), 2 (worse estimate), 0 (no improvement). The rows labeled with *Sig* indicate if the change was significant by use of a $t-$test.

changes by the GA with respect to the total number of generations until convergence is shown over all spectra. The various metabolites are indicated by different symbol shapes and shadings.

One can clearly see that for most Δ (indicated by the symbols) around 40% of the GA generations are sufficient to obtain a stable solution. Even if this solution may not be a global optimum, it can still be considered as a stable local optimum. For some of the Δ (e.g. those for pyruvate and aspartate) a (much) larger number of updates is necessary and it can be expected that these shifts are not well optimized, but that no better solution could be found by the GA. For some of the other metabolites one can also see that only a single group is optimized very frequently as is the case for the group of Val (valine) around 0.98 ppm or Ser (serine) around 3.83 ppm. The concentration estimates for these two amino acids compare quite well with the expert estimates. Very few updates can be observed e.g. for Succ (succinate) around 2.39 ppm, Gly (glycine) around 3.55 ppm or Glu (glutamate) around 2ppm and 3.75ppm. Interestingly Succ, Gly and Glu are optimized very well and the estimated concentrations correspond reasonably to those obtained by the expert. However it should be borne in mind that the concentration estimate is not equally split over the groups.

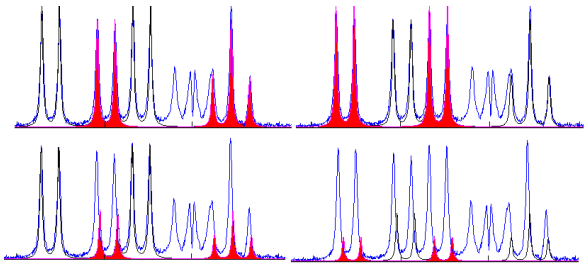


Fig. 9: Spectrum in the region of valine and iso-leucine. The two sub figures on the top show the fit with ETP, left for iso-leucine (filled), right for valine (filled). Below the same but in the original TP fit.

The plot in Figure 10 provides an initial indication of which metabolites are most likely to have been poorly optimized and should therefore be manually corrected by the expert. This provides a basis for the focused and guided manual interaction avoiding the inspection of all metabolites. In the example shown, the optimizations appear to have been reasonably effective and correct for those metabolites for which the number of updates for the corresponding Δ lies within a range of $20-40\%$. The plots in Figure 8 show the effect of the genetic evolution with respect to the concentration of the metabolites and the parameter modifications. One can see that most of the optimization of GA parameters and hence of concentration changes occurs in the first 10-20 generations. The plot also shows that even relatively small errors in the Δ_i may have large impact on the concentration estimates, with very high values at the beginning of the optimization and comparatively small values at the end for some metabolites.

Median relative error estimates of single metabolites using TP and ETP (FUNC) are shown in the Box-Whisker plot 7. The relative error is calculated as the absolute concentration error compared to the expert value.In terms of the median errors, we find that ETP provides a clear improvement over TP but has still problems with some metabolites such as Leu, Meth, Phe and Pyr. In these cases, however, we note that even the manual fit by the expert is challenging. On average the median error improved from 0.78 to 0.64 with 0 as the perfect agreement. These findings show that ETP is superior to TP in providing reasonable estimates for metabolites on a magnitude level. However, the accuracy attainable from single measurements is still low. This highlights the need both for the use of experimental replicates and for the analysis of multiple spectra of the same sample.

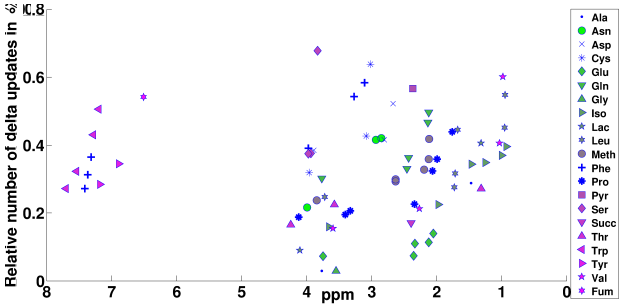


Fig. 10: Relative #Δ updates in % (y) with respect to the ppm position (x).

5 CONCLUSION

In summary, this work has shown that an approach combining GAs with LSQO leads to highly effective error estimates for the shift uncertainties in ^{1}H NMR measurements. The simultaneous fit outperforms the standard TP approach with respect to identification and quantification accuracy and compares favorably to the expert analysis. We have further shown that the usage of a data specific (functional) distance measure to calculate the fitness values is preferable to a standard Euclidean measure. It also significantly improved the convergence rate of the GA. The interpretation of the obtained shifts over time with the best model allows an in depth analysis of the optimization, revealing potentially unreliable fits. This provides initial guidance for the expert to focus further manual improvement of the obtained fit where necessary, reducing the

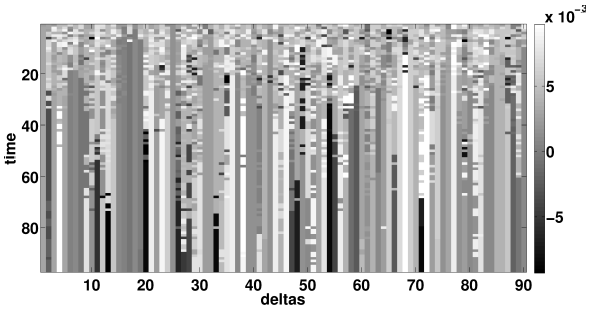


Fig. 11: A typical evolution of the Δ (columns) over time (rows) for 97 generations. The gray levels indicate the shift values. Some Δ converge early to stable values, some (few) need more updates.

demand for extensive shift corrections in order to generate correct uncertainty estimates. Furthermore, the approach also allows the manual, specification of concentration values in the fit for known concentrations, by additional constraints. Overall the combined approach can improve the identification and quantification accuracy of NMR based targeted profiling to allow a semi-automatic high throughput analysis. Further improvements are to be expected from improved preprocessing of the spectra. Variations in the baseline and slightly incorrect lineshapes being the main sources of error in the automatic identification and quantification of metabolites in NMR measurements.

ACKNOWLEDGMENT

We thank Prof. Thomas Villmann (Univ. of Appl. Sc. Mittweida) for discussions about functional signal processing, the METASTEM team and Peter Tino, Univ. of Birmingham for a very effective research stay during the preparation of this manuscript.

Funding: This work was supported by the Fed. Ministry of Edu. and Res.:FZ:0313833 A, (NMR Metabolic Profiling of the Stem Cell Niche, METASTEM), the German Res. Fund. (DFG), HA2719/4-1 (Relevance Learning for Temporal Neural Maps) and by the Cluster of Excellence 277 Cognitive Interaction Technology funded in the framework of the German Excellence Initiative.

REFERENCES

Böcker, S., Letze, M. C., Liptak, Z., and Pervukhin, A. (2009). Sirius: decomposing isotope patterns for metabolite identification. *Bioinformatics*, **25**(2), 218–224.

Brelstaff, G., Bicego, M., Culeddu, N., and Chessa, M. (2009). Bag of peaks: interpretation of nmr spectrometry. *Bioinformatics*, **25**(2), 258–264.

Chang, D., Banack, C. D., and Shah, S. L. (2007). Robust baseline correction algorithm for signal dense nmr spectra. *Journal of Magnetic Resonance*, **187**, 288–292.

Defernez, M. and Colquhoun, I. J. (2003). Factors affecting the robustness of metabolite fingerprinting using 1H-NMR spectra. *Phytochemistry*, **62**, 1009–1017.

Fletcher, R. (2000). *Practical Methods of Optimization*. Wiley VCH.

Forshed, J., Schuppe-Koistinen, I., and Jacobsson, S. P. (2003). Peak alignment of NMR signals by means of a genetic algorithm. *Analytica Chimica Acta*, **487**, 189–199.

Forshed, J., Torgrip, R. J. O., Aberg, K. M., Karlberg, B., Lindberg, J., and Jacobsson, S. P. (2005). A comparison of methods for alignment of NMR peaks in the context of cluster analysis. *Journal of Pharmaceutical and Biomedical Analysis*, **38**, 824–832.

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley, Reading, MA.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York.

Jarvis, R. M. and Goodacre, R. (2005). Genetic algorithm opimization for pre-processing and variable selection of spectroscopic data. *Bioinformatics*, **21**, 860–868.

Koh, H.-W., Lambert, J., Maddula, S., Hergenrder, R., and Hildbrand, L. (2008). Feature selection by lorentzian peak reconstruction for 1-h nmr post processing. In *Proc. of CBMS 2008*, pages 608–613. IEEE Press.

Lee, J. and Verleysen, M. (2005). Generalizations of the lp norm for time series and its application to self-organizing maps. In M. Cottrell, editor, *5th Workshop on Self-Organizing Maps*, volume 1, pages 733–740.

Levitt, M. H. (2008). *Spin Dynamics: Basics o Nuclear Magnetic Resonance (2nd Ed.)*. Wiley.

Mathworks Inc (2008). Matlab 2008b.

Mendes, P. (2006). Metabolomics and the challenges ahead. *Briefings in Bioinformatics*, **7**(2), 172.

Mitchell, M. (1995). *An Introduction to Genetic Algorithms*. MIT Press, Boston, MA.

Moco, S., Bino, R. J., Vos, R. C. D., and Vervoort, J. (2007). Metabolomics technologies and metabolite identification. *Trends in Analytical Chemistry*, **26**(9), 855–866.

Pierce, K. M., Wright, B. W., and Synovec, R. E. (2007). Unsupervised parameter optimization for automated retention time alignment of severely shifted gas chromatography data using the piecewise alignment algorithm. *Journal of Chromatography A*, **1141**, 106–116.

Savorani, F., Tomasi, G., and Engelsen, S. (2010). icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, **202**, 190–202.

Schleif, F.-M. (2006). *Prototype based Machine Learning for Clinical Proteomics*. Ph.D. thesis, Technical University Clausthal, Technical University Clausthal, Clausthal-Zellerfeld, Germany.

Schleif, F.-M. (2007). Preprocessing of nuclear magnetic resonance spectrometry data. *Machine Learning Reports*, **1**(MLR-01-2007). ISSN:1865-3960.

Schleif, F.-M., Riemer, T., Cross, M., and Villmann, T. (2008). Automatic identification and quantification of metabolites in h-nmr measurements. In *In Proceedings of the Workshop on Computational Systems Biology (WCSB) 2008*, pages 165–168.

Schlumm, T. and Riemer, T. (2001). Nmrj: A feasability study for a fully Java based platform independent MR processing and analysing program. In *Proceedings of the International Society for Magnetic Resonance in Medicine*, volume 9, page 798, Glasgow.

Smith, S., Levante, T., Meier, B., and Ernst, R. (1994). Computer simulations in magnetic resonance. an object oriented programming approach. *J. Magn. Reson.*, **106a**, 75–105.

Son, H., Hwang, G., Kim, K., Ahn, H., Park, W., Berg, F., Hong, Y., and Lee, C. (2009). Metabolomic studies on geographical grapes and their wines using 1H NMR analysis coupled with multivariate statistics. *Journal of Agric. Food Chem*, **57**(4), 1481–1490.

Swire, J., Fuchs, S., Bundy, J., and Leroi, A. (2009). The cellular geometry of growth drives the amino acid economy of caenorhabditis elegans. *Proceeding of the Royal Society B*, **276**(1668), 2747–2754.

Tiziani, S., Lodi, A., Khanim, F., Viant, M., Bunce, C., and Günther, U. (2009). Analysis of mixed lipid extracts using 1h nmr spectra. *PLoS ONE*, **4**(1).

Vogels, J. T. W. E., Tas, A., Venekamp, J., and v. d. Greef, J. (1996). Partial linear fit: a new NMR spectroscopy preporcessing tool for pattern recognition applications. *Journal of Chemometrics*, **10**, 425–438.

Weckwerth, W. (2003). Metabolomics in systems biology. *Ann. Rev. Plant Biol.*, **54**, 669–689.

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

Xi, Y. and Rocke, D. M. (2008). Baseline correction for nmr spectroscopic metabolomics data analysis. *BMC Bioinformatics*, **9**, 324–333.

Xia, J., abd P. Tang, T. C. B., and Wishart, D. S. (2008). Metabominer semi-automated identification of metabolites from 2d nmr spectra of complex biofluids. *BMC Bioinformatics*, **9**, 507–522.

Zhao, Q., Stoyanova, R., Du, S., Sajda, P., and Brown, T. R. (2006). Hiresa tool for comprehensive assessment and interpretation of metabolomic data. *Bioinformatics*, **22**(20), 2562–2564.