CHENOMX
NMR Mixture Analysis

COMPLETE AUTOFIT: Flexible and automated NMR spectral fitting in Chenomx NMR Suite V10

Pascal Mercier (pmercier@chenomx.com),
Chenomx Inc, Canada

www.chenomx.com

INTRODUCTION

The mathematical challenges associated with automatic reconstruction and fitting of high-resolution proton NMR spectra has been the bottleneck of high-throughput analysis of NMR metabolomics data since its infancy.

The superiority of targeted profiling over fast binning methods has long been established, but the identification and the quantification of compounds via targeted profiling has been hindered by the time required to perform manual or computer-assisted spectral fitting.

Quantitative or targeted profiling offers an alternative route to spectral reduction techniques such as binning, where experimental spectra are reconstituted at their full resolution from a sum of their underlying components using a reference compound library. In this approach, compounds are identified and quantified prior to performing any kind of multivariate statistical analyses. The challenge in quantitative metabolomics lies in the time and effort needed to identify and quantify compounds in biofluid mixtures.

Different approaches have been formulated over the years, with more or less success, or limited applicability, due to the sensitive nature of the NMR signal with experimental conditions (solvent, magnet strength, pulse sequence and parameters) and the mathematical complexity of spectral deconvolution, whereby the parameter space to explore is enormous and hundreds if not thousands of variables need to be optimized simultaneously.

Here we present a new, fully automated fitting algorithm based on mathematical methods derived from artificial intelligence that improve peak positioning and concentration fitting.

METHODS

The query spectrum $\bar{S}$ is reconstructed in the frequency domain using a linear combination of each of the reference compound spectra following a Lorentzian model:

$$\bar{S}(x) = \sum_{i=1}^{metabolites} c_i \sum_{j=1}^{nclusters} \sum_{k=1}^{npeaks} \frac{h_{i,j,k} \bullet w_{i,j,k}^2}{4\left(x - \left(\delta_{i,j,k} + \Delta\delta_{i,j,k}\right)\right)^2 + w_{i,j,k}^2}$$

where $\bar{S}$ is the predicted spectrum, (x) is the predicted spectral intensity at point index x, c_i the concentration of metabolite i (linear variables), $\delta_{i,j,k}$ the resonance frequency of peak k of cluster j of metabolite i, $\Delta\delta_{i,j}$ the cluster center offset (called "transform") of peak cluster j of metabolite i (non linear variables), and $h_{i,j,k}$ and $w_{i,j,k}$ the intensity and line width of peak k in cluster j of metabolite i, respectively.

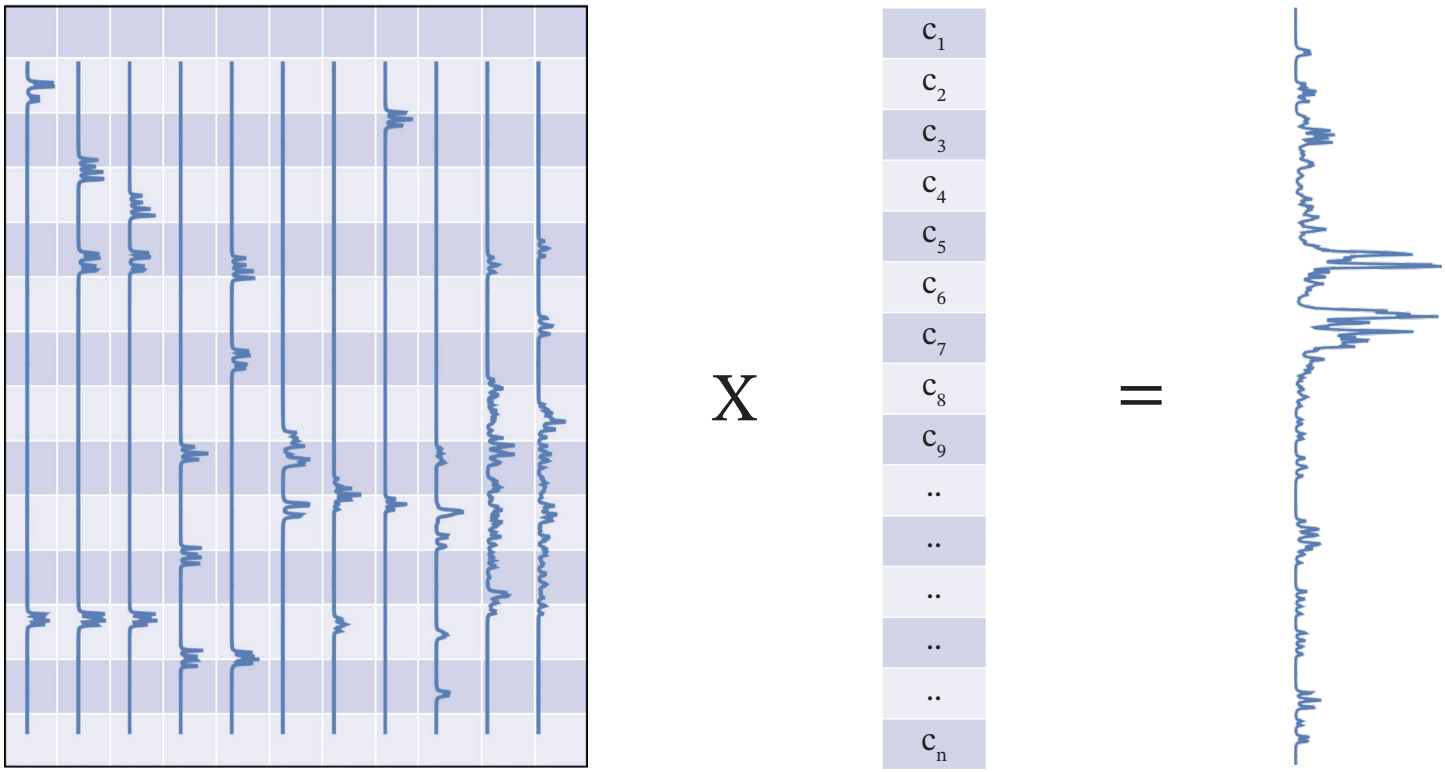
The parameters $h_{i,j,k}$ and $w_{i,j,k}$ are constants from the compound library once the library has been calibrated to the provided spectrum. A non-negativity condition is imposed on all c_i such that $c_i \geq 0$. Upper bounds also be applied on c_i based on maximal potential concentrations.

The Chenomx advanced pH-sensitive reference compound libraries are used to set starting peak cluster centers ($\Delta\delta_{i,j}$) and to provide lower and upper bound values $\Delta\delta_{i,j}$.

The automated fitting procedure consists of finding the c and $\Delta\delta$ values that minimizes the 2-norm target function $\chi^2 = \|S - \bar{S}\|^2$.

When expressed in matrix-form, the problem simplifies to $\bar{S} = AC$, where A is the model matrix, and C is the concentration vector. For each possible resulting model matrix A, there exists a corresponding concentration vector C that minimize the 2-norm target function. The vector C is solved by iterative reweighted bounded linear least-square. A weight penalty can be imposed on spectral points at locations where the predicted/reconstructed intensities exceed the experimental spectrum.

METHODS *...continued*



Model matrix A | Concentration vector C | Spectrum $\vec{f}$

Both the construction of A by automatic positioning of peak clusters and the optimization of vector C derive from mathematical methods developed for artificial intelligence.

PROCEDURE

- The model matrix A is filled one column (compound) at a time by snapping peak cluster shapes at their most likely positions on the spectral subtraction line, i.e. the experimental spectrum minus the current contribution of all compounds except for the one being processed. The order in which compounds are processed is randomized at each cycle.

- Optionally, the matrix can be downsized (binned) to a slightly lower resolution . This can help to account for different experimental peak shapes between the experimental spectrum and what is contained in the reference compound library being used.

- For each generated model matrix the compound concentrations are obtained by iterative weighted constrained linear least squares.

- After n internal cycles, the best reached solution (in terms of residual χ^2) is output.

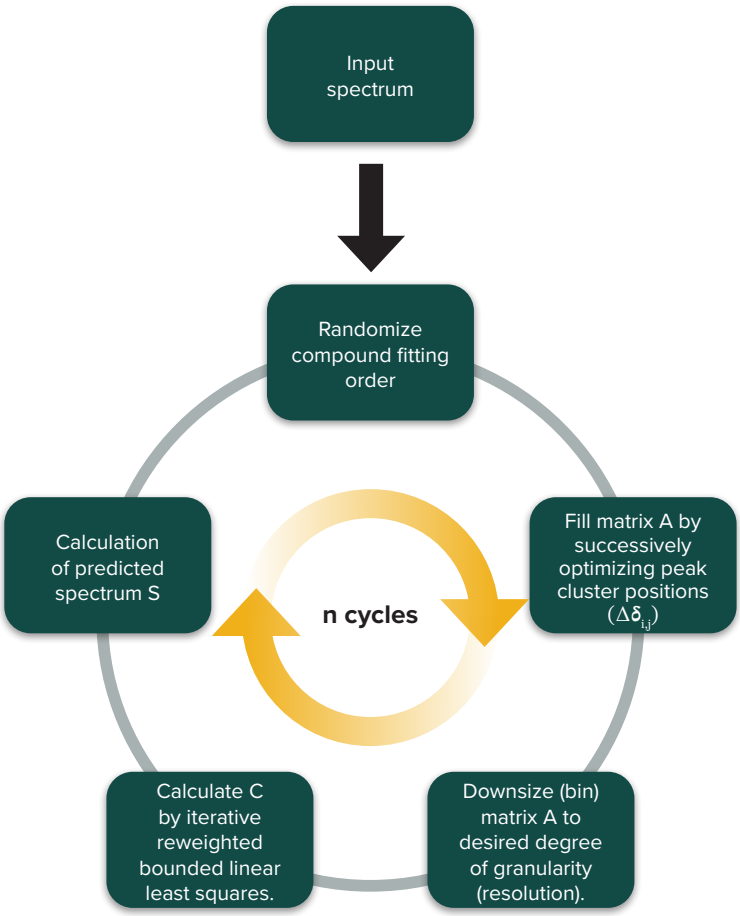
- For each input spectrum, several solutions are produced in parallel using several CPUs. Each resulting fit (cnx file) can be inspected using Chenomx's regular spectral Profiler module.

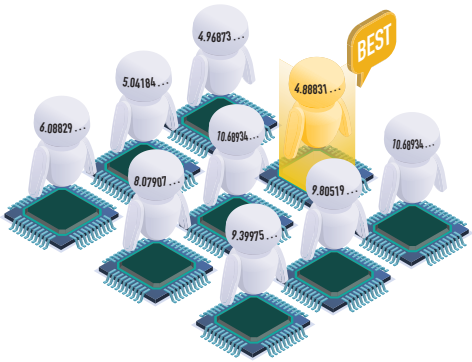
- The best solution(s) can be selected from the reported χ^2 in the generated fit report.

- Inspection of concentration distributions can provide insights about the range of values and errors (std. deviation) obtained for each compound concentration across the produced fits. Statistical methods can be used to further keep or reject compounds for further analysis (PCA, etc).

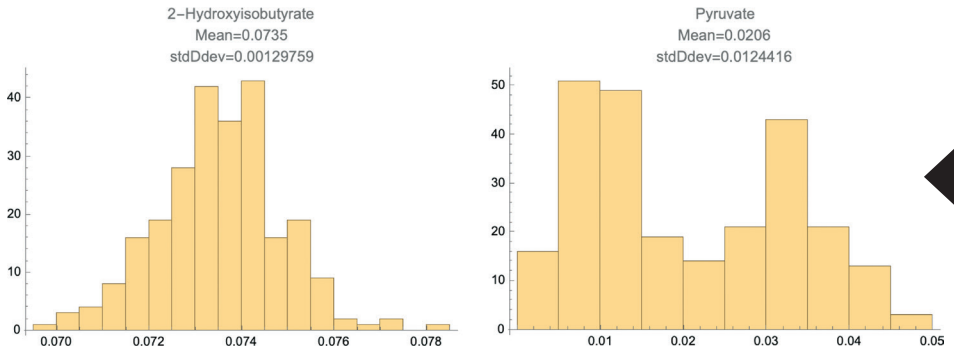
- The algorithm mimics a lab setting where an ensemble of fit solutions are expected from human estimates.

PROCEDURE *...continued*





Output	Concentrations
Spectrum fit #1	$c_{1,1,}\ c_{1,2,}\ c_{1,3,}\ \ldots$
Spectrum fit #2	$c_{2,1,}\ c_{2,2,}\ c_{2,3,}\ \ldots$
Spectrum fit #3	$c_{3,1,}\ c_{3,2,}\ c_{3,3,}\ \ldots$
Spectrum fit #4	$c_{4,1,}\ c_{4,2,}\ c_{4,3,}\ \ldots$
....	....



Output of several solutions (multi-CPUs) of similar fit quality (χ^2 residuals)

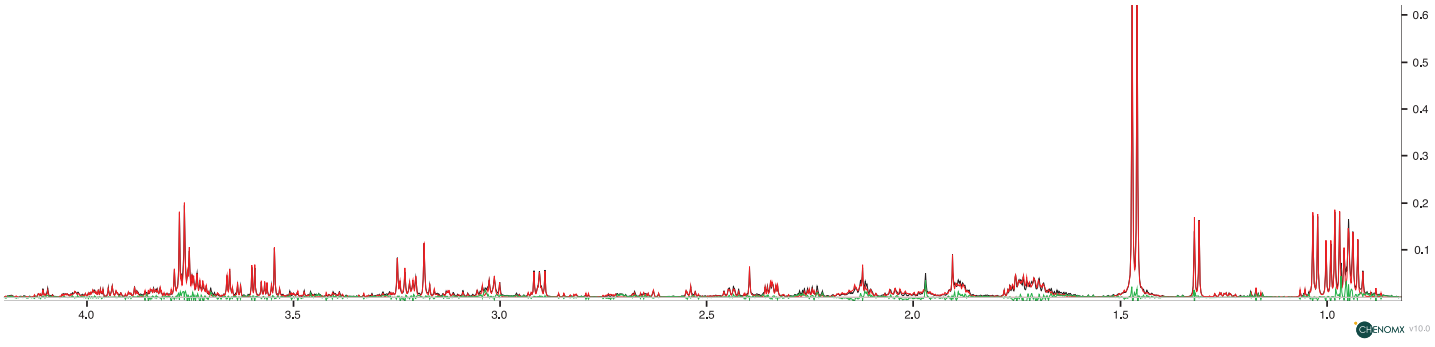
Analysis of concentration distributions prior to post-statistical treatment (PCA, OPLS-DA, ...)

CHENOMX NMR Mixture Analysis

IMPORTANT CONSIDERATIONS

Contrary to mass-spectroscopy where masses are constant, NMR signals are not. Peak intensities and frequencies change with spectrometer frequency, temperature, solvent conditions (pH, solvent type), pulse sequence and pulse sequence parameters.

- For maximum compatibility and accuracy, the reference compound library and the experimental spectra should be acquired under the same conditions.
- Both underfitting and overfitting must be avoided by choosing a compound set that is representative of the type of sample being analyzed.
- Proper phasing, baseline correction and shim correction (reference deconvolution) must have been applied properly à priori. The algorithm does not evaluate the baseline on the fly in its current form. Background signal/large humps associated with large molecules (proteins, lipids) must be eliminated prior to automatic fitting.



Selected region of autofit 1H NMR spectrum of cell media sample at 600 MHz with 55 compounds.