GSD (Global Spectrum Deconvolution) in Metabolomics: Absolute Quantification and GSD Binning

Silvia Mari[1], Valeria Mannella[1], Giacomo Quilici[1], Giovanna Musco[1], Carlos Cobas[2], Felipe Seoane[2], Stan Sykora[3]

[1]San Raffaele Scientific Institute, Center of Genomics, BioInformatics and BioStatistics, Milan, Italy, www.sanraffaele.org

[2]Mestrelab Research, Santiago de Compostela, Spain, www.mestrelab.com

[3]Extra Byte, Castano Primo, Italy, www.ebyte.it

Introduction

NMR spectroscopy is an important tool in many metabolomic applications. Its potential capability to handle complex mixtures of metabolites makes it a prime choice in both identification and quantification of the multitude of different species constituting unprocessed biological mixtures. In quantitative metabolomics (targeted metabolomic profiling) all detectable compounds in a biosample are identified and quantified by comparing the biosample spectrum to a library of reference spectra of pure compounds [1]. The underlying assumption is that a spectrum is the weighed sum of the spectra of all the individual metabolites which constitute the mixture (spectral additivity). In practice, the inverse problem of decomposing an experimental spectrum into its component parts corresponding to individual metabolites is arduous due to their large number and the complexity of their spectra, massive overlap of spectral peaks, non-trivial deviations of peak shapes from the ideal Lorentzian profile [2], lack of a suitable orthonormal base in the vector space spanned by sets of Lorentzian peaks, and presence of artifacts such as receiver noise, irregular baseline drifts, and magnetic field inhomogeneity effects. Due to all these factors, attempts to overcome peaks overlap problems by means of conventional deconvolution (fitting) of selected spectral areas have in general limited success. For this reasons, and because the actual nature of all the metabolites is rarely known in advance, metabolomics often uses alternative statistical evaluation methods, such as multivariate factor analysis [3], which sidestep the need for a complete interpretation of the spectra and a full solution of the inverse problem, while still permitting the correlation of the spectra with specific biological aspects. However, such approaches require integration over predefined intervals (bins) and a meaningful integration of such intricate and artifact-burdened spectra may often be just as arduous as peaks fitting. Recently, a new algorithm called GSD (Global Spectrum Deconvolution) has been developed [4] and made available in the Mnova software package of Mestrelab. GSD is capable of identifying even poorly resolved spectral peaks and of fitting all recognizable peaks in even a very complex 1D spectrum in a surprisingly short time (typically a dozen seconds for up to 1000 peaks). Moreover, it is fully automatic and objective (no human intervention is required) and produces a table of all detectable spectral peaks and their parameters. Such a table can be then used for various purposes such as generation of artifact-free synthetic spectra (with or without resolution enhancement), stick spectra, artifact-free integrals, as well as accurate binning void of any bin-crossover problems due to the overlapping wings of spectral peaks. Because of these attractive features, GSD is likely to become a very important pre-processing tool for all metabolomic approaches to the evaluation of NMR spectra of whole biosamples.

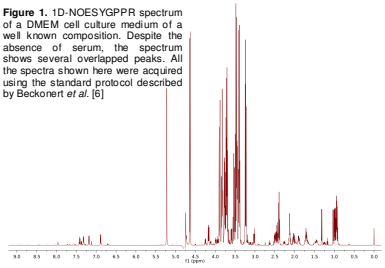

Figure 1. 1D-NOESYGPPR spectrum of a DMEM cell culture medium of a well known composition. Despite the absence of serum, the spectrum shows several overlapped peaks. All the spectra shown here were acquired using the standard protocol described by Beckonert *et al.* [6]

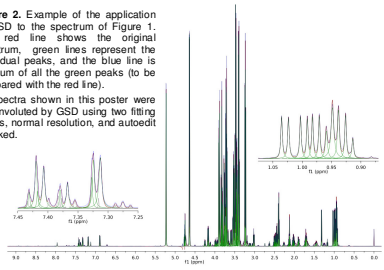

Figure 2. Example of the application of GSD to the spectrum of Figure 1. The red line shows the original spectrum, green lines represent the individual peaks, and the blue line is the sum of all the green peaks (to be compared with the red line).

All spectra shown in this poster were deconvoluted by GSD using two fitting cycles, normal resolution, and autoedit checked.

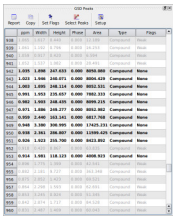

Table 1. The GSD table generated after the deconvolution illustrated in Figure 2. All deconvoluted peaks (green lines in Figure 2) are listed in the table and can be exported.

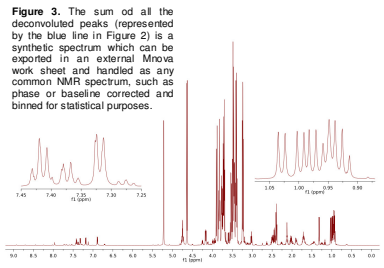

Figure 3. The sum od all the deconvoluted peaks (represented by the blue line in Figure 2) is a synthetic spectrum which can be exported in an external Mnova work sheet and handled as any common NMR spectrum, such as phase or baseline corrected and binned for statistical purposes.

Metabolite Quantification and Multivariate Statistic

In addition, we have taken advantage of GSD to implement under Mnova the quantitative referencing strategy known as PULCON [5]. It is well known that when dealing with biological samples containing lipids and proteins, one can not use an internal standard to perform absolute quantification of metabolites. This is due to the fact that all possible reference compounds (such as TSP or DSS) interact with large biological molecules, making the quantification error prone. By combining the GSD algorithm with a PULCON script we are able to deconvolute overlapped regions and perform absolute quantification even of metabolites whose peaks are buried under these areas. In order to verify the quality of the quantification, we have acquired spectra of a commercial cell culture medium which is a complex but well known mixture of more than 20 compounds. We then compare the metabolite concentrations obtained from experimental spectra by means of the GSD - PULCON algorithm with

Equation 1.

$$\frac{A_{ref}\cdot cal_{ref}}{conc_{ref}} = \frac{A_{sample}\cdot cal_{sample}}{conc_{sample}}$$

$$where\ \ cal_i = \frac{T_i\cdot PI_i}{RG_i\cdot NS_i\cdot SI_i}\ ,\ \ A_i\ \ are\ absolute\ integrals$$

Here we present the first controlled tests of the applicability of GSD to the multivariate analysis of complex metabolite mixtures such as cell culture media containing serum and mouse urine. GSD allows the user to generate synthetic spectra that can be used in the subsequent steps of multivariate statistical analysis.

In particular, we compare the quality of PCA plots obtained by means of conventional binning of experimental spectra (NOESY and CPMG) with the PCA plots obtained by binning the synthetic spectra generated by the GSD algorithm.

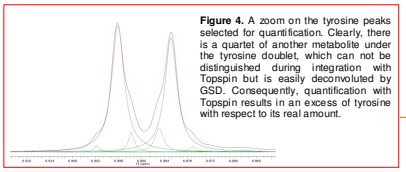

Figure 4. A zoom on the tyrosine peaks selected for quantification. Clearly, there is a quartet of another metabolite under the tyrosine doublet, which can not be distinguished during integration with Topspin but is easily deconvoluted by GSD. Consequently, quantification with Topspin results in an excess of tyrosine with respect to its real amount.

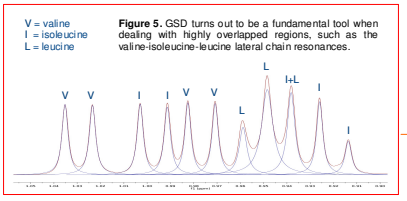

Figure 5. GSD turns out to be a fundamental tool when dealing with highly overlapped regions, such as the valine-isoleucine-leucine lateral chain resonances.

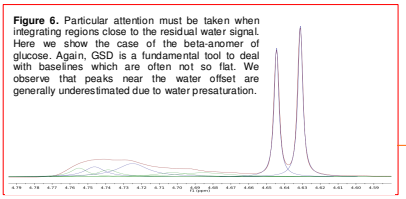

Figure 6. Particular attention must be taken when integrating regions close to the residual water signal. Here we show the case of the beta-anomer of glucose. Again, GSD is a fundamental tool to deal with baselines which are often not so flat. We observe that peaks near the water offset are generally underestimated due to water presaturation.

More than one region per metabolite was chosen for the quantification. In general an average peak area was calculated and then divided by the number of protons contributing to the integrated peaks. According to Equation 1, each metabolite's absolute area was then compared with the one obtained for the sucrose+DSS standard sample in H_20/D_2O, resulting in the following **Table 2.**

		NOESY		CPMG	
	composition as declared by vendor (mM)	Topspin quantcal/ quantint AU programs (mM)	Mnova pulcon (mM)	Topspin quantcal/ quantint AU programs (mM)	Mnova pulcon (mM)
niacinamide	0.0328	0.0234 (±0.0158)	0.0296 (±0.0015)	0.0343 (±0.0036)	0.0347 (±0.0026)
histidine[a]	0.2	0.19 (±0.01)	0.18 (±0.02)	0.23 (±0.01)	0.23 (±0.01)
tyrosine	0.398	0.429 (±0.049)	0.312 (±0.011)	0.534 (±0.036)	0.411 (±0.016)
leucine	0.802	0.672 (±0.027)	0.631 (±0.016)	0.863 (±0.036)	0.838 (±0.025)
valine	0.803	impossible to integrate	0.662 (±0.009)	impossible to integrate	0.870 (±0.021)
isoleucine	0.802	impossible to integrate	0.645 (±0.029)	impossible to integrate	0.882 (±0.022)
glucose	25	19.4 (±0.5)[b]; 25.5 (±0.7)[c]	7.8 (±0.1)[b]; 22.7 (±0.1)[c]	30.4 (±1.4)[b]; 30.8 (±0.8)[c]	13.26 (±0.03)b; 29.52 (±0.09)c

[a] vendor declares histidine in its hydrochloride-H2O form. [b] obtained by the sum of alpha-anomer (5.225ppm) doublet and beta-anomer doublet (4.635ppm) [c] obtained by the sum of alpha-anomer (5.225ppm) and H5 of the beta isomer (3.235ppm)

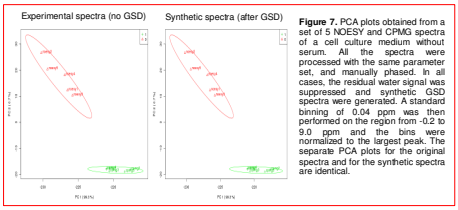

Figure 7. PCA plots obtained from a set of 5 NOESY and CPMG spectra of a cell culture medium without serum. All the spectra were processed with the same parameter set, and manually phased. In all cases, the residual water signal was suppressed and synthetic GSD spectra were generated. A standard binning of 0.04 ppm was then performed on the region from -0.2 to 9.0 ppm and the bins were normalized to the largest peak. The separate PCA plots for the original spectra and for the synthetic spectra are identical.

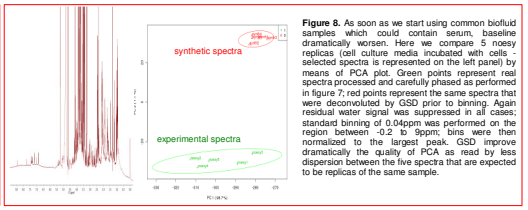

Figure 8. As soon as we start using common biofluid samples which could contain serum, baseline dramatically worsen. Here we compare 5 noesy replicas (cell culture media incubated with cells selected spectra is represented on the left panell by means of PCA plot. Green points represent real spectra processed and carefully phased as performed in figure 7, red points represent the same spectra that were deconvoluted by GSD prior to binning. Again residual water signal was suppressed in all cases; standard binning of 0.04ppm was performed on the region between -0.2 to 9ppm; bins were then normalized to the largest peak. GSD improve dramatically the quality of PCA as read by less dispersion between the five spectra that are expected to be replicas of the same sample.

References

1. Wishart D.S., Quantitative metabolomics using NMR, TrAC, 27, 228-237 (2008). DOI 10.1016/j.trac.2007.12.001
2. S.Sykora, C.Cobas, Peak Shapes in NMR Spectroscopy, www.ebyte.it/stan/Talk_MMCE_2009.html.
3. HG.J.Issaq, Q.N.Van,T.J.Waybright,G.M.Muschik,T.D.Veenstra, Analytical and statistical approaches to metabolomics research, J.Sep.Sci., 32, 2183–2199 (2009). DOI 10.1002/jssc.200900152.
4. C.Cobas, S.Sykora, The Bumpy Road Towards Automatic GSD, Poster at 50th ENC, DOI 10.3247/sl3nmr09.003.
5. G.Wider, L.Dreier, Measuring Protein Concentrations by NMR Spectroscopy, DOI 10.1021/ja055336t, JACS 128, 2571-2576 (2006). See also www.earcr-meeting.ethz.ch/IMB/groups/wider_group/publications/.
6. Beckonert O. et al, Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. Nature Protocol, 2, 2692-2703 (2007)

This poster was presented at the **51st ENC**, April 18-23, 2010 Daytona Beach, FL, USA.

DOI: 10.3247/sl3nmr10.002