"An advanced clustering approach for assessing the repeatability and statistical relevance of 2D-COSY spectra"

Feraud, Baptiste ; Govaerts, Bernadette ; Verleysen, Michel ; de Tullio, Pascal

Abstract

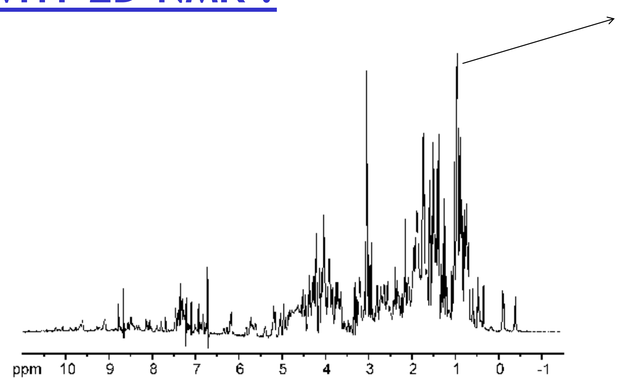
NMR techniques are widely used together with multivariate analysis approaches in order to characterize perturbations in metabolic pathways occurring during biological processes. A large amount of recent scientific and statistical works are available concerning 1D spectra (principally ^{1}H-NMR spectra). More recently, two-dimensional NMR spectroscopy techniques have been investigated: homonuclear (COSY,…) and heteronuclear ones (HSQC,…). It is commonly accepted by users (biologists, pharmacologists) that the recent introduction of 2D-NMR methods represents a huge qualitative gap for metabolomics investigations in terms of metabolites and biomarkers identifications. Indeed, it seems obvious that additional dimension means more predictive power. But, until now, no statistical study clearly proved this assumption. Therefore, a fundamental question is “Is supplementary information equivalent to relevant and crucial information ?”. In order to extend the statistical properties and t...

Document type : *Communication à un colloque (Conference Paper)*

An advanced clustering approach for assessing the repeatability and statistical relevance of 2D-COSY spectra

Baptiste Feraud (UCL, ISBA), Pascal de Tullio (CIRM, Chimie Pharmaceutique, Ulg), Bernadette Govaerts (UCL, ISBA), Michel Verleysen (UCL, MLG)

WHY 2D-NMR ?



biomarker? or biomarkers?

1D protein spectra are often far too complex for interpretation

- Signals overlap heavily
- Ambiguous or overlapping resonances
- …

Introduction of an additional spectral dimension = extra information (obvious)

- separate the contributions made by individual resonances
- analysis and quantization of **off-diagonal peaks !**

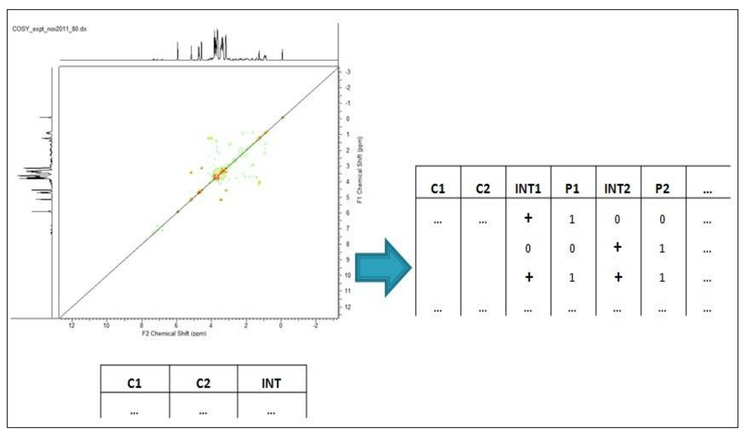
Statistical question: Is this supplementary information equivalent to relevant and crucial information for identification and prediction ?

EXPERIMENTAL DESIGNS AND PRE-PROCESSING

Two 1D (1H-NMR) and 2D COSY experimental designs

1) Four **cell culture systems** containing various levels of different metabolites (fetal bovine serum, glutamax, amino acids, vitamins, proteins, ...). Then, three samples per mixture have been collected. And, finally, three repeated measures have been collected on every sample. All these samples have been subject to freezing and defrosting steps, with real risks of degradation and bacterial contamination because of the duration of the 2D analysis process. In this design, signal is linked with the four initial mixtures and noise is linked with sampling, time repetitions, risks of degradation and other acquisition and condition parameters. 36 measures are finally stored, corresponding to 36 COSY spectra and 36 corresponding peak lists.
2) **Human serum** data. Four blood donors were engaged for the study. The design consists in eight days of measurements with repetitions within each days and multiple permutations (for instance, spectral techniques (1D or 2D COSY) have not been applied at the same moment of the day, thus creating different delays before the spectral measure). One donor is finally linked with eights measures/spectra/peak lists, with 32 measures all in all.

Some pre-processing steps :



- Concept of **Global Peak List**
- Symmetrization
- « Bucketing » by controlling the size of the final database via a chosen number of decimals for the coordinates (for instance, one decimal → (909 × 74), two decimals → (2348 × 74), three decimals → (3250 × 74)). It is also a way to propose different levels of spectral resolution.
- Normalization of each intensity vector such that sum = 1
- Detection of outliers (…)

CHECKING FOR THE REPEATABILITY OF HOMONUCLEAR 2D-NMR COSY SPECTRA : WHY CLUSTERING ?

An intuitive way to evaluate the repeatability / reproducibility of these 2D spectra consists in **non-supervised multivariate clustering** (blind, with **no a priori labeled information**).

Key idea : If we manage to separate and recover our 4 initial groups starting from the 36 spectral measures → Done !

CLUSTERING ON POSITION VECTORS

Working on the signals' positions or, in other words, on the simple existence of signals is motivated by a biological justification: a signal, or a particular metabolite, can be observed or not for a particular donor or in a particular media. If signals are present in detectable quantity, this presence or absence is supposed to be very stable, whereas intensities are supposed to be flexible and variable from a measure to another, according to different parameters or factors.Appropriate similarity measures such as Ochiai and Jaccard are needed to capture the binary specificity.
Ward and K-means algorithms are applied and different combinations are taken into account via different spectral resolutions.

Results: in the vast majority of cases, we can already isolate particular groups.

CLUSTERING ON NORMALIZED INTENSITY VECTORS

Again, Ward and K-Means algorithms are implemented along with different spectral resolutions. Classic euclidean distance is used.

Results: generally, all groups are well recovered by the algorithms in spite of the sampling procedure and time repetitions. Our best clustering result is obtained with the one-decimal matrix (this underlines the importance of the « bucketing ») with just ONE ERROR.

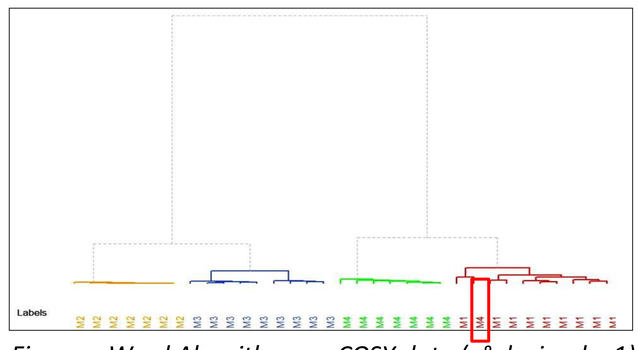


Figure – Ward Algorithm on COSY data (n° decimal = 1)

Conclusions: COSY appears to be a statistically repeatable tool. Initial groups and obtained clusters are mainly concordant.

COMPARISON WITH 1H-NMR RESULTS : NUMERICAL QUALITY INDEXES

The second objective is to compare these 2D results with corresponding 1D results (^{1}HNMR) obtained in the same conditions (elimination of negative intensities, resolution proportional or equal to the 2D horizontal axis, same number of decimals, without outliers…). So, besides the good repeatability of COSY, **the goal is to demonstrate that the additional information is relevant and crucial by improving the quality of the clustering results**.
Some internal validation indexes are needed to evaluate this quality, like Dunn index (DI), Davies-Bouldin index (DBI), Rand Index (RI) and Adjusted Rand Index (ARI).

Data	DI	DBI	RI	ARI
Positions, dec=1, Jaccard	0.7955	1.5688	**0.9373**	**0.8250**
Positions, dec=1, Ochiai	0.7215	1.5688	**0.9373**	**0.8250**
Positions, dec=2, Jaccard	0.9453	1.7998	0.7721	0.4005
Positions, dec=2, Ochiai	0.8988	1.7998	0.7721	0.4005
Positions, dec=3, Jaccard	**0.9813**	1.7920	0.6496	0.1705
Positions, dec=3, Ochiai	0.9633	1.7920	0.6496	0.1705
Intensities, dec=1, Ward	0.4186	**0.6432**	**0.9316**	**0.8040**
Intensities, dec=1, K-means	0.4186	**0.6432**	**0.9316**	**0.8040**
Intensities, dec=2, Ward	0.6886	1.5919	0.7892	0.4216
Intensities, dec=2, K-means	0.7061	1.7417	0.7350	0.2884
Intensities,dec=3, Ward	0.7780	1.6679	0.5784	0.0545
Intensities, dec=3, K-means	0.7644	1.8855	0.6467	0.1349

Table – Clustering results on COSY data (2nd design)

Data	DI	DBI	RI	ARI
Intensities, dec=1, Ward	**0.9810**	0.6878	0.7036	0.2324
Intensities, dec=1, K-means	0.3335	**0.6628**	0.7016	0.2294
Intensities, dec=2, Ward	0.9011	0.9862	0.7036	0.2324
Intensities, dec=2, K-means	0.6350	1.1195	**0.7399**	**0.2898**
Raw intensities, Ward	0.6698	1.0736	0.5880	0.0172
Raw intensities, K-means	0.4469	1.3207	0.6754	0.1446

Table – Clustering results on 1D data (2nd design)

Conclusions: On almost all results, although clusters are more compacts and well separated when using 1D spectra (see DI and DBI), COSY appears to be a better tool for prediction (see RI and ARI). And this, specially for « bucketed » data.

FURTHER WORK: Confirm these results with fastest COSY experiments, use of other medias and/or spectral datasets (heteronuclear HSQC), prediction and identification of discriminating zones and biomarkers.