Description of the citrate-hippurate database

From Rousseau (2011) and Guisset* et al. *(2019)

This article presents the Semi-artificial Urine database originally used in Rousseau (2011) and Guisset *et al.* (2019). This database was designed with spectroscopists from Eli Lilly and the University of Liege (ULg).

Introduction

The database has been experimentally created in order to control the spectral locations of the biomarkers to find. This property allows us to evaluate the performances of the data analysis of various statistical methods.

This urine experimental database is also designed in order to explore the influence on spectra of diuretic fluctuations, intra-sample ^{1}H NMR replications, and inter-day ^{1}H NMR measurements.

Motivations for creating this database

This database was collected according to an experimental design with three objectives:

① The first one was to study the ability of statistical methods to find as biomarkers the descriptors of the spectra for which a variability was experimentally controlled. In this experiment homogenised medium urine samples were spiked with two products at different levels of concentration and analysed by spectroscopy. The "concentrations" factor of each added product has to mimic the variability focused in a biomarker search study.

② The second objective was to study the quality of normalisation techniques. Two dilutions in water of a same pool of urine were considered in these experiments.

③ The third objective was to study the natural variability of spectra related to several factors: the day of analysis, and the repetition of a spectral acquisition (replication).

Statistical experimental design

Four factors of variability were considered in the experimental design:

1. The medium: two media are considered: the urine from a pool ("B04-331, fisher 344 rats, female") and the urine coming from the same pool but diluted with a dilution rate of 50 %.
2. The concentration of each of the two products into the samples: hippuric acid and citric acid were added to the samples in different concentrations described in Figure 1. The maximal concentrations are 8mM for the citric acid (QC) and 4mM for the hippuric acid (Qh). Four levels of each acid were considered, determining fourteen experimental conditions.

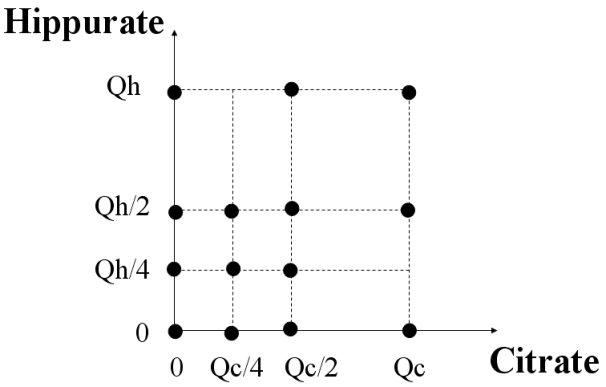

Figure 1: *The urine experimental design.*

As shown in Figure 2, the peaks corresponding to each product are located in distant ppm domains. The hippurate has three peaks with two in the region of high ppms containing a low noise level. On the contrary, citrate peaks are located in the noisy region of low ppms. Note that in the spectral pre-treatments the citrate peaks are aggregated in one peak to avoid alignment problems.

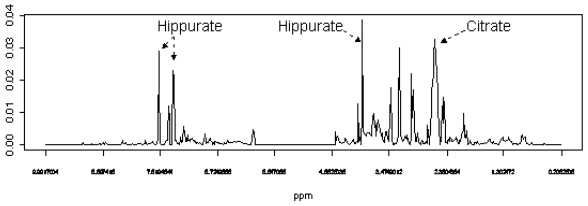

Figure 2: *A typical urine spectrum with spiked citrate and hippurate.*

3. The replicates: each mixture or experimental condition described in Figure 1. was repeated at least twice. Consequently, 28 samples were prepared for each medium.
4. The day of ^{1}H NMR measurement: for each medium, the preparation of the 28 mixtures (14 experimental conditions $\times$ 2 replicates) were performed in three series. The two media samples of each series correspond to a plate. Each day, a plate was defrosted in order to be analysed by the ^{1}H NMR spectrometer. Additionally, plate 1 was analysed twice: the plate was analysed on day 1, refrost, then defrost and analysed again on day 5. Values for the day of measurement factor are: day 1, day 2, day 3, day 5.

In each plate, according to the previous description, we have 56 samples (2 media $\times$ 14 experimental conditions $\times$ 2 replicates). In addition, other replicates were included for experimental conditions with no spiked compounds. In this way, over the four days, we obtained 269 samples to analyse.

Sample preparation and acquisition of the ^{1}H NMR data

Each sample had a final volume of 1200 μl. For each prepared sample, a mixture of 600 μl was added to 600 μl of media. The media was either urine coming from a pool of 344 female Fischer rats or the same urine diluted at 50%. The mixture included the two products (citrate and hippurate) in the chosen concentrations and phosphate buffer containing TMSP. The volume of the phosphate buffer was adapted in order to obtain a volume of 600 μl to add to the 600 μl of urine. For example, a final urine media sample contains 600 μl of urine and 600 μl of mixture composed of 150 μl of citrate, 150 μl of hippurate and 300 μl of buffer. The corresponding diluted media sample was obtained by taking half of the citrate volume, half of the hippurate and a larger volume of buffer to keep the mixture volume at 600 μl. In the example, the corresponding final diluted urine media sample contains 600 μl of the diluted urine media and 600 μl of the mixture, here composed by 75 μl of citrate, 75 μl of hippurate and 450 μl of buffer. In each preparation, the mixture was added to the urine, centrifuged, frozen at $-$ 80°C and unfrozen at 4°C the day before the ^{1}H NMR analysis.

Sample tubes were analysed within each day of experiment with a NOESY presaturation sequence.

The pre-treatments

Each spectrum was processed using the pre-treatment procedure advised in Rousseau (2011). The part of the spectrum between 0.2 and 10 ppm has been reduced to 600 descriptors. Attention must be paid to the facts that:

- The ppm values corresponding to the large non-informative urea and to the water peak (4.5 - 6.0 ppm) were set to zero.
- The region around the citrate resonances (2.56 - 2.72 ppm) was integrated into one peak to suppress the high shifts of the citrate peaks.
- the data were normalised by a constant sum normalisation.

The full final urine database

The spectra are labelled as: $Mm - Cxy - Dd - Rr$ where m is the code for the medium (1 or 2 for urine and diluted urine respectively), xy are the product proportions of the maximal concentrations, d is the day, and r is the time index (replicate). Among all the the samples analysed, three presented problems during the spectral acquisition (M1C04D1R2, M2C00D1R7, M2C48D1R1).

The spectral data matrix X of dimensions (269 $\times$ 600) is contained in the data file *HC_PtotP.Rdata*. Indeed, in addition to the 224 samples were 2 replicates are present, blank conditions were replicated 8 times.

The data file *HC_Destot.Rdata* corresponds to the experimental design of dimensions with 14 columns describing the spectra. The variables contained in it are: the name of the spectrum, the medium, the combined level of hippurate and citrate (qualitative variable), the level of hippurate separately (qualitative variable), the level of only citrate (qualitative variable), the volume of added hippurate (quantitative variable), the volume of added citrate (quantitative variable), the index of the replicate and the day.

The UCH sub-dataset used for Chapter 4

In the subset of this database used in Chapter 4, and further referred to as the UCH data set, citrate and hippurate factors have three levels: no added chemical (a concentration of 0), a medium concentration (2 mM for the citrate factor (Qc/4) and 1 mM for the hippurate factor (Qh/4)) and a high concentration (4 mM for the citrate factor (Qc/2) and 2 mM for the hippurate factor (Qh/2)). This leads to an experimental design with 9 possible combinations of concentrations as illustrated in Figure 3(a). Two different days were retained (days 2 and 3) in the subset. The UCH data originally contained 36 spectra *i.e.* 9 combinations of concentrations $\times$ 2 time points $\times$ 2 replicates. However, two outliers were removed, resulting in 34 observations and a slightly unbalanced experimental design (*cf.* Figure 3(b)). The resulting spectral matrix is of size (34 x 600).

Finally, Figure 3(d) presents the scores plot of the first two PCs for the spectral matrix. It shows that the information on the citrate and hippurate factor levels from the experimental design can be recovered, but only to some extent.

The data file *UCH.Rdata* contains this subdataset where *design* represents the experimental design with variables *Hippurate*, *Citrate*, *Dilution*, *Day* and *Time*; *outcomes* represents the metabolomic spectral profiles; and *formula* is the model formula that was used in Chapter 4.

References

Guisset, S., Martin, M., & Govaerts, B. 2019. Comparison of PARAFASCA, AComDim, and AMOPLS approaches in the multivariate GLM modelling of multi-factorial designs. *Chemometrics and Intelligent Laboratory Systems*, **184**, 44 – 63.

Rousseau, R. 2011. *Statistical contribution to the analysis of metabonomics data in ^1H NMR spectroscopy.* Ph.D. thesis, Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain, Belgium.

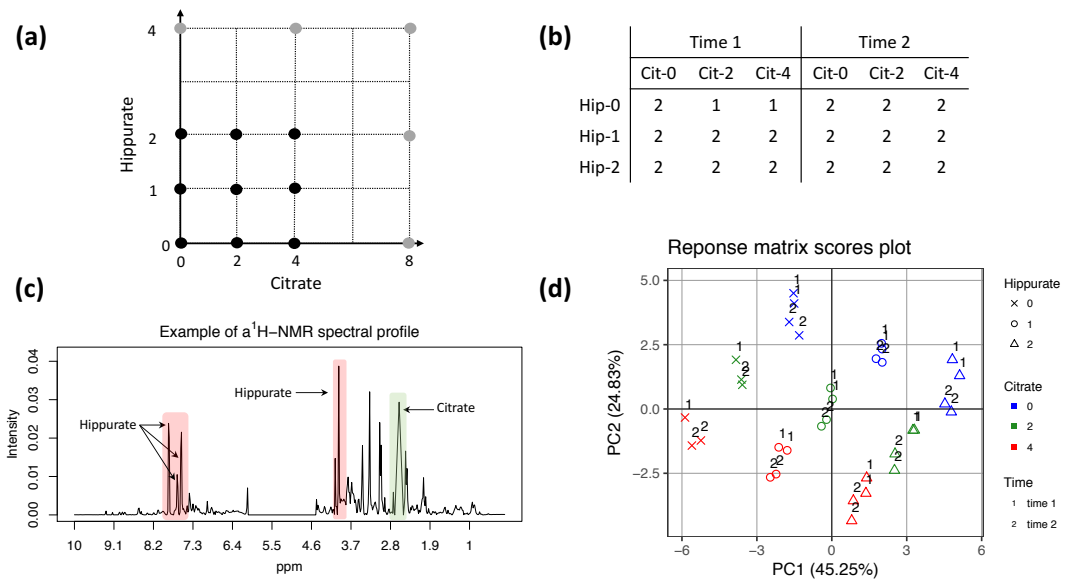

Figure 3: *Presentation of the UCH sub-dataset (a) Experimental design (b) Sample sizes for each of the 9 combinations of citrate, hippurate and Time levels (c) Example of a 1H NMR spectral profile showing the hippurate and citrate peaks (d) Response matrix scores plots for the first two PCs of the UCH data set.*