Journal of Pharmaceutical & Biomedical Analysis
Vol. 10, No. 2/3, pp. 141–144, 1992
Printed in Great Britain

0731–7085/92 $5.00 + 0.00
Pergamon Press Ltd

An automatic data reduction and transfer method to aid pattern recognition analysis and classification of NMR spectra

R.D. FARRANT, J.C. LINDON,* E. RAHR and B.C. SWEATMAN

Department of Physical Sciences, Wellcome Research Laboratories, Langley Court, Beckenham, Kent BR3 3BS, UK

Abstract: A method of automatically generating reduced NMR data and transferring it between computers is proposed. These data can then be used as descriptors for input to non-parametric statistical routines for classification of the samples.

Keywords: *NMR; data transfer; pattern recognition; principal components.*

Introduction

A common problem in analytical science is the need to categorize a sample into one of a number of classes based upon a series of measurements. These measurements could be a series of spectroscopic and/or chromatographic determinations and the results would be compared against a series of standard samples where the outcome was known. The simplest case is to pass or fail a sample from, for example, a clinical chemistry test by comparison of a single measured marker such as occurs in the urinary glucose test for diabetes. However, the situation is often less clearly defined, there may not be a single descriptor, and more complex methods are needed. One approach is to generate a 'training set' of examples where the outcome is known and to compare, by some statistical test, the similarity of the test sample to various categories of sample in the training set.

This type of situation is found in studies of ^{1}H NMR of body fluids [1] such as urine [2] or bile [3], where it is possible to generate several hundred NMR spectra each containing many hundreds of resonances arising from the natural endogenous metabolites and often very many spectra are measured either from a wide range of patients in clinical studies or in a statistically valid number of animals when studying drug effects. In this case, one possible need is to categorize the spectra and hence the animals from which they arise on the basis of the observed biochemical changes which can be related to the toxicity of administered substances [1, 2, 4–6]. At present, the spectra are quantified by measuring a series of peak heights or areas for specific known endogenous metabolites (up to about 30) and these are input manually into statistics routines on a separate computer to that controlling the NMR spectrometer. We have used these data as input to pattern recognition software in order to obtain the maximum classification information, rather than rely on individual biochemical markers [7–9]. Clearly it would be highly desirable to have some form of automatic data reduction of the spectra, automatic transfer to the statistics computer, and routines to convert the data into a format suitable for analysis by pattern recognition software.

We describe herein a simple yet effective prototype approach which achieves all of these goals and we apply it to data sets acquired in a recent toxicological study and compare the method to manually inputting peak area data for selected metabolites.

Method

Proton NMR spectra are acquired in the normal fashion either with or without water suppression [1] using TSP (3-trimethylsilyl-[2,2,3,3,-$^2H_4$] 1-propionate) as an internal reference at a fixed concentration. Each

* Author to whom correspondence should be addressed.

acquired spectrum is processed automatically using a microprogram on the spectrometer or associated data-station, with exponential weighting, Fourier transformation, phase correction and baseline correction. Using modern spectrometers, in our case a Bruker AM-360 running DISNMR, it is possible to use the NMR software to identify a peak close to 0δ and assign a chemical shift of exactly 0.0δ. Regions of the spectra are assigned for peak picking with a predefined threshold fixed by examination of the first spectrum; we have chosen these regions for body fluids to be nominally 10.0δ to 5.2δ and 4.4δ to −0.1δ, thereby omitting artefacts from any water suppression. Using the NMR computer each spectrum is peak-picked into a file containing resonance frequencies in ppm and peak heights in ASCII format. If all the spectra are acquired under the same conditions then this fully automatic procedure works well but if for historical reasons the spectra have been measured under different conditions it may be necessary to set, for example, the peak-picking threshold manually, this being a minor task. This procedure can be used on any FT spectrometer from any of the major manufacturers. In our laboratory the ASCII files are transmitted over a serial line using KERMIT [10] to a DEC VAX-8550 which is used for the statistical analysis and imported into the data handling software suite RS/1 [11]. Using software written in RS/1 procedure language, the TSP peak is identified by its chemical shift and all other resonances are scaled by this height. This routine could equally well be written in any high level language such as FORTRAN or C or even in the specialist commands as widely used by spreadsheet software. Because we are usually interested only in relative changes to the biochemical profiles no attempt is made to convert the peak heights into concentrations (this would require detailed assignment of the spectra, something the method is attempting to overcome, along with knowledge of urinary flow rates, urine collection times, number of protons contributing to the resonance and identification of coupling patterns). Each spectral file is split into integration regions, the width of which can be varied but is typically of the order of 0.025–0.1 ppm. All peak heights within each region are summed thus producing for the case of a 0.05 ppm region over a 10.0δ to 5.2δ and 4.45δ to −0.15δ spectral range, a total of 188 descriptors defining that spectrum. Since TSP is taken to resonate at 0.0δ, this peak could lie on the boundary of two integration regions if the regions are chosen injudiciously and hence the software offsets the regions by half of the integration width, i.e. for an integration region of 0.05 ppm, the process would begin at −0.075δ, ensuring that TSP appears in the middle of a region. From these data, an RS/1 table is constructed where each row relates to a sample with the row entries consisting of the 188 descriptors in decreasing chemical shift order. This method avoids simply using the observed peak intensities because in many cases peaks for defined substances may or may not be present, for example creatine is not normally observed in rat urine but is a major metabolite when the testicular toxin cadmium is adminstered [12]. The above histogram method ensures that each descriptor always relates to the same point in the spectrum.

The RS/1 data table is then used as input to the pattern recognition software ARTHUR [13] which produces, amongst other output, two-dimensional representations of the multidimensional data (each of the 188 descriptors can be thought of as a coordinate in a 188 dimensional space). Points (i.e. individual spectra corresponding to single animals or patients) which are close together in a map have by definition a similarity of input variables. We have earlier demonstrated the classification of toxins according to site of action based upon NMR of urine using two such dimension reduction techniques, namely nonlinear mapping [14] and principal component analysis [15] and details of the techniques as applied to bio-fluids have been published [7–9].

Results

Figure 1 shows a representative ^{1}H NMR spectrum of urine from a control rat measured at 360 MHz, indicating the spectral complexity and the artefacts which can be introduced by water suppression. The data used here to test the data processing method arose from a toxicological study comprising five male and five female control Wistar rats and equal numbers of males and females dosed with a potentially toxic substance (B.C. Sweatman, C.R. Beddell, J. Wood, J.C. Lindon and G.O. Evans, unpublished results). Figure 2(a) shows a principal component map obtained by

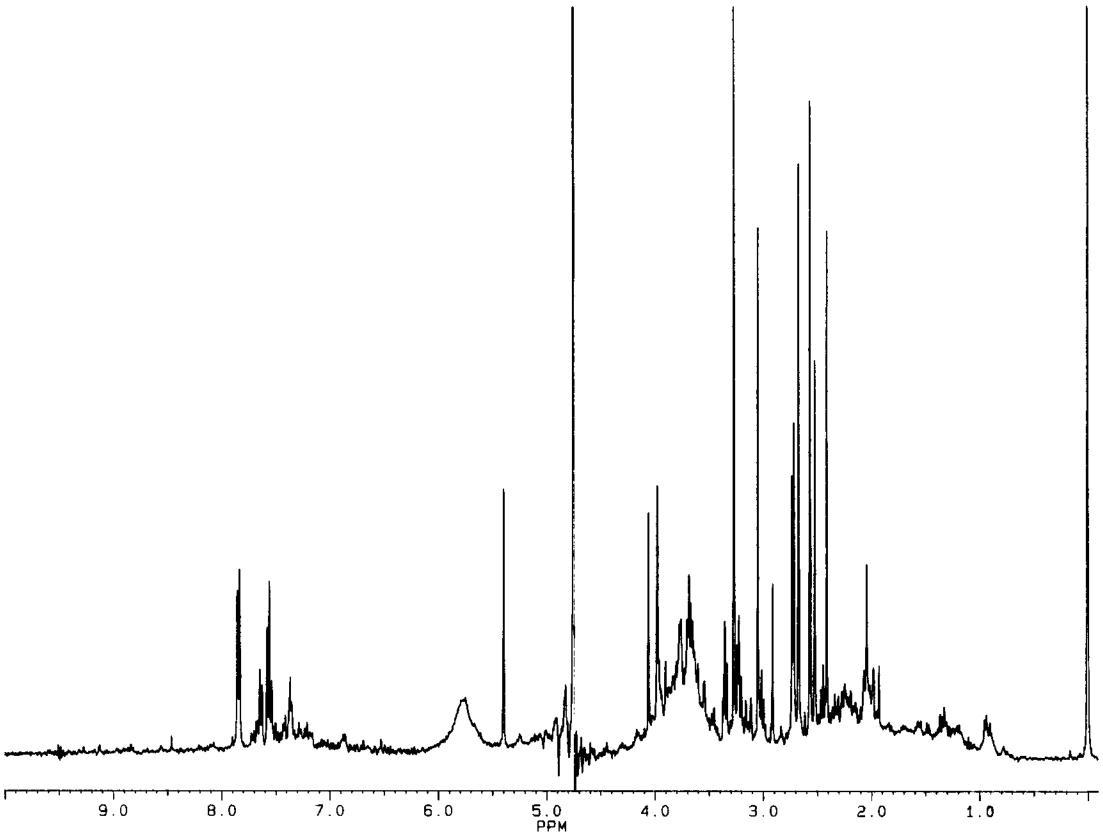

Figure 1
Proton NMR spectrum (360 MHz) of urine from a control animal.

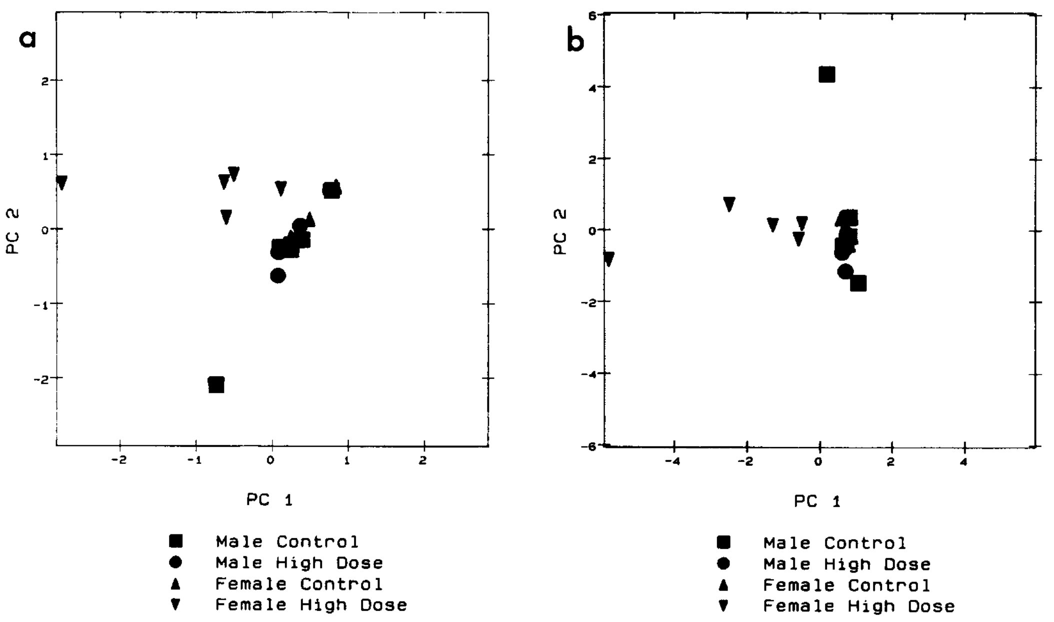

Figure 2
(a) Plot of the first two principal components for 26 NMR derived specific metabolite peak areas. (b) Plot of the first two principal components for 188 NMR descriptors generated automatically as defined in the text. Each point represents one animal, the key to the groups is as shown.

measuring peak areas relative to TSP for 26 selected endogenous metabolites. The data were autoscaled to give each metabolite a mean of zero and a variance of unity, thus ensuring equal weighting of each metabolite. Each point on the map represents one animal and the distance between points is related to their similarity in terms of the descriptors used. The female dosed group is clearly well separated from all the other animals as a result of altered biochemistry. In addition, one of the male control animals shows an abnormal NMR spectrum and appears on the map away from the other controls.

Figure 2(b) shows a principal components map processed with autoscaling of the data as for Fig. 2(a) but using the automatic data generation method described above. A histogram region width of 0.05 ppm was used resulting in 188 descriptors for each spectrum. Using a histogram region of 0.1 ppm (94 descriptors) gave a similar result. The similarity of the maps generated from the two methods is evident and both show the same separation of the different classes of animal, namely high dose females, one male control and the remainder of the control animals.

A detailed investigation is being undertaken to optimize such parameters as FID line-broadening, integration region and peak-picking threshold, but it is expected that these will vary according to the samples and it will be safer to determine a sufficient, if not optimal set of parameters, from the examination of the acquired data. Many NMR observable metabolites give rise to more than one chemically shifted resonance and hence there exists the probability that some of the descriptor values are intercorrelated. We have investigated this possibility and have within the software package the facility to remove descriptors from pairs which have a predefined, high correlation coefficient so that we can choose to include or exclude highly correlated data.

The method demonstrated here should enable the incorporation of data from spectrometers from any manufacturer and we have also demonstrated its utility on cerebrospinal fluid NMR spectral data from a Jeol GSX-500 instrument (F.Y.K. Ghauri, J.K. Nicholson, C.R. Beddell, R.D. Farrant and J.C. Lindon, unpublished results). In addition, the method could possibly allow the combination of data from spectrometers at different field strengths.

Conclusions

We have demonstrated a robust and efficient method of automatically reducing and transferring NMR data for input to statistical analysis routines which is potentially capable of coping with spectra from instruments from different manufacturers and at different field strengths. We have exemplified the method with ^{1}H NMR spectra from urine for classifying samples according to drug toxicity, but the method is equally applicable to other body fluids, to NMR spectra of small organic molecules to classify according to chemical structures, to NMR spectra of macromolecules for comparing carbohydrate or amino-acid sequences or indeed to other types of spectroscopy including ^{13}C NMR or to chromatography for a wide variety of structural or analytical purposes.

References

[1] J.K. Nicholson and I.D. Wilson, *Prog. NMR Spectrosc.* **21**, 449–501 (1989).

[2] J.K. Nicholson, J.A. Timbrell and P.J. Sadler, *Mol. Pharmacol.* **27**, 644–651 (1985).

[3] K.P.R. Gartland, K.E. Wade, C.T. Eason, F.W. Bonner and J.K. Nicholson, *J. Pharm. Biomed. Anal.* **7**, 699–707 (1989).

[4] K.P.R. Gartland, F.W. Bonner and J.K. Nicholson, *Mol. Pharmacol.* **35**, 242–250 (1989).

[5] K.P.R. Gartland, F.W. Bonner, J.A. Timbrell and J.K. Nicholson, *Arch. Toxicol.* **63**, 97–106 (1989).

[6] P. Foxall, M. Bending, K.P.R. Gartland and J.K. Nicholson, *Hum. Toxicol.* **9**, 491–496 (1989).

[7] K.P.R. Gartland, C.R. Beddell, J.C. Lindon and J.K. Nicholson, *NMR Biomed.* **3**, 166–172 (1990).

[8] K.P.R. Gartland, C.R. Beddell, J.C. Lindon and J.K. Nicholson, *J. Pharm. Biomed. Anal.* **8**, 963–968 (1990).

[9] K.P.R. Gartland, C.R. Beddell, J.C. Lindon and J.K. Nicholson, *Mol. Pharmacol.* **39**, 629–642 (1991).

[10] F. Da Cruz and W. Catchings, *Byte* **9**, 255–278 (1984).

[11] RS/1. BBN Software Products UK Ltd, Staines, Middlesex (1988).

[12] J.K. Nicholson, D.P. Higham, J.A. Timbrell and P.J. Sadler, *Mol. Pharmacol.* **36**, 398–404 (1989).

[13] ARTHUR 81. Version 4.1, B&B Associates, Seattle, WA 98105 (1981).

[14] B.R. Kowalski and C.F. Bender, *J. Am. Chem. Soc.* **94**, 5632–5639 (1972).

[15] H. Seal, *Multivariate Statistical Analysis for Biologists*, pp. 101–102. Methuen, London (1968).

[*Received for review 11 September 1991*]