Pergamon

0731-7085(94)00073-5

Journal of Pharmaceutical & Biomedical Analysis, Vol. 12, No. 10, pp. 1215–1225, 1994
Elsevier Science Ltd
Printed in Great Britain
0731-7085/94 $7.00 + 0.00

Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples

M. SPRAUL, P. NEIDIG, U. KLAUCK, P. KESSLER, E. HOLMES,† J.K. NICHOLSON,† B.C. SWEATMAN,‡ S.R. SALMAN,‡§ R.D. FARRANT,‡ E. RAHR,‡ C.R. BEDDELL‡ and J.C. LINDON*‡

Bruker Analytische Messtechnik GmbH, Silberstreifen, D76287-Rheinstetten 4, Germany
† Department of Chemistry, Birkbeck College, University of London, Gordon House, 29 Gordon Square, London WC1H 0PP, UK
‡ Department of Physical Sciences, Wellcome Research Laboratories, Langley Court, Beckenham, Kent BR3 3BS, UK

Abstract: A general method of automatically reducing NMR spectra to provide numerical descriptors of samples has been developed and investigated. These descriptors can be used as input to pattern recognition or multivariate algorithms for sample classification. The methods have been tested using 600 MHz one-dimensional ^{1}H NMR spectra of biofluids which are complex mixtures. The approach is, in principle, applicable to multidimensional and heteronuclear NMR spectra and to other types of liquid samples such as oils and foodstuffs as well as to situations such as ^{1}H or ^{31}P NMR *in vivo* and solid state NMR in drug formulation analysis. The method relies upon apportioning the information in the spectra to individual contiguous segments and allowing specified regions of the spectra to be omitted. Three approaches, based on the number of peaks, the summed peak heights and the summed peak areas respectively in each segment, have been tested. The effect of segment width and overlap and the effects of manipulation of the NMR spectra have been evaluated in terms of the classification of the samples using principal components analysis. A simple method of generating NMR based spectral descriptors for object classification is thus proposed.

Keywords: *NMR spectroscopy; data reduction; pattern recognition; sample classification; toxicity; biofluids.*

Introduction

In general there are two principal approaches in analytical chemistry for deciding whether a particular sample belongs to a given class [1]. It may be possible to make a determination of a single parameter with sufficient precision to be able to assign the sample to a particular class based on statistical significance. An example of this type of analysis could be the determination of the level of glucose in blood or urine as an indicator of diabetes. This approach requires an *a priori* selection of the analyte and relies upon a specific and proven hypothesis of the relationship between the analyte and the sample classification. Alternatively, it is possible to choose a series of analytes in the sample, and based upon multivariate statistics, the determined values of these analytes can then be used to assign the sample to a class. This method also relies on the preselection of analytes for study and the requirement that they are known to be the most significant for distinguishing the sample classes and there are many such examples of this type of analysis in clinical biochemistry and disease diagnosis [2].

Some modern analytical techniques allow the simultaneous determination of many descriptors of the sample without the need for preselection of the analytes. This allows the analyst to accept all the detectable data relating to the sample and provides the opportunity to determine which amongst the possible myriad of parameters are those which are significant in distinguishing sample classes. It may be that new and previously undiscovered markers of the sample classes will result from this approach. This mode of operation relies upon the presence of a 'training set' of samples with independently known classification so that significance testing of the novel descriptors can then be carried out.

* Author to whom correspondence should be addressed.
§ Present address: Institute of Industrial Chemistry, University of Science and Technology of Oran, B.P. 1505, El M'Naouar, Oran, Algeria. Permanent address: College of Science, University of Baghdad, Baghdad, Iraq.

Foremost among such analytical measurement methods must be high resolution NMR spectroscopy. This has the potential for simultaneous determination of many independent parameters in a complex mixture with the added advantage that relative signal intensities (if the experiment is suitably conducted) are directly proportional to concentrations of the species in the sample. One particular area where complex NMR spectra occur is the study of ^{1}H NMR spectra of biofluids such as urine where often thousands of resonances result from the presence of both endogenous biochemicals and drug metabolites [3–8]. Disease [9], physiological changes [10, 11] or toxic processes [12] can cause marked changes in the levels of endogenous metabolites and in severe cases, for example in the case of in-born errors of metabolism [13], completely new substances can be present. Under these circumstances, it is unnecessarily censorious to preselect metabolites for study, although such analysis of biofluids using NMR spectroscopy has been extremely successful in studying toxicity, drug metabolism and disease processes [3]. A study has, therefore, been made of methods to ensure that as comprehensive a set of descriptors as possible is collected in an automatic fashion such that multivariate analysis of the data can follow. This follows some preliminary work by us which has been reported previously [14].

Multivariate methods have been applied in a number of NMR-based studies. For example, in order to improve NMR spectral assignments, principal components (PC) analysis and cluster analysis have been used for recognition of resonances in 2-dimensional (2D) NMR spectra [15] and neural networks have helped the identification of antiphase cross-peaks in 2D NMR spectra [16]. A neural network approach has also been used to identify NMR spectra from sugars in oligosaccharides [17] and from sugar alditols [18]. Other computer-aided studies include the assignment of ^{13}C-^{1}H and ^{1}H-^{1}H correlation spectra [19]. The optimum features of 2D NMR spectra which can be used for classifying residues in peracetylated oligosaccharides using PC analysis and SIMCA [20] have been elucidated [21]. The use of a variety of multivariate methods based on PC analysis has been reported in order to predict NMR parameters, to relate chemical shifts to biological significance and to reduce noise and spectral artefacts [22]. Much work has been carried out on predicting ^{13}C chemical shifts using such multivariate methods [23] including more recently, the investigation of the usefulness of neural networks [24].

However, one potentially important area of the use of multivariate methods for classification is to use NMR data as descriptors of a biological condition, for example, disease or toxic stress. Some studies have reported monitoring the growth of tumours using NMR of blood serum [25] and distinguishing various types of tumour using the ^{1}H NMR spectra of extracts [26, 27]. In addition, attempts have been made to classify disease states from *in vivo* spectra using factor or PC analysis. In particular, a study has been made of ^{31}P NMR spectra from muscle myopathy patients [28] and of localized ^{1}H NMR spectra in neurological diseases [29].

Significant efforts have been made to classify ^{1}H NMR spectra of rat urine in terms of toxic insult [30, 31] where it was possible to distinguish the organ containing the toxic lesion, for example kidney, liver or testis, and within the kidney to identify toxins which affected different parts of the kidney, for example the cortex or medulla. Moreover, it was also possible to distinguish different biochemical mechanisms of toxicity within the kidney using PC methods [32]. A comparison has also been made of the relative usefulness of using ^{1}H NMR or conventional clinical chemistry tests for classifying the toxic nature of xenobiotics [33]. As well as distinguishing different types of toxicity, it has been shown that PC analysis is useful for demonstrating the time course of toxic effects and that trajectories in PC maps can be used to distinguish different types of kidney toxicity [34, 35].

The basic principle behind our earlier work on automatic data reduction was that in a complex mixture containing species that may be absent or undetectable in some abnormal situations, simply to carry out a 'peak-picking' exercise would be misleading in that the *n*th descriptor would not always correspond to the same biochemical marker. Hence in the earlier study, the ^{1}H NMR spectrum was reduced by carrying out a peak frequency and intensity listing of the spectrum, dividing the spectrum into regions of defined width (e.g. 0.05 ppm) and summing the peak heights in each region to obtain a series of numerical descriptors equally spaced along the NMR frequency axis. The validity of this approach was checked by

making a comparison with a more conventional manual measurement of the heights of peaks corresponding to 26 known metabolites [14].

This automatic approach has now been extended by considering alternative descriptors for such a segment analysis and by allowing regions of the spectra to be automatically excluded to avoid artefacts introduced by solvent suppression or drug metabolite resonances when seeking endogenous metabolite information. In addition, a study has been made of the effects of changing the segment chemical shift width, of allowing overlap of the segment regions and of changing spectral line-broadening parameters, on the ability to classify samples. Finally, the level of error introduced by manual phasing of the spectra has been investigated.

Experimental

Samples of rat urine were obtained from animals as part of on-going toxicological studies. Human urine samples were taken from healthy volunteers under different types of mild physiological stress who were being studied as part of an independent examination of the use of NMR spectroscopy to investigate normal human physiological variation. ^1H NMR spectra were measured at 600 MHz on a Bruker AMX600 instrument using water resonance presaturation. Chemical shifts were referenced relative to internal sodium trimethylsilyl[2,2,3,3,-$^2H_4$]propionate (TSP) at δ0.0. Detailed reports of the NMR spectroscopy and the pattern recognition results of these studies will be given separately. The NMR free induction decays were processed using the spectrometer operating and processing software UXNMR. To construct segment regions and hence to obtain spectrum descriptors, three approaches were used. In one case, the NMR spectra were reduced to a list of resonance frequencies and heights using the standard peak-picking routine and these lists were stored in ASCII files. Using routines written in the processing language of the table manipulation package RS/1 [36], these intensities were summed in discrete frequency bands (segments). This program also provided definition of the spectra in terms of the number of segment regions, allowed areas of the spectra to be excluded and provided for overlap of segment regions. Alternatively, similar software was written for the Bruker X32 computer for converting the NMR spectrum directly into descriptors and this allowed two further types of descriptor to be used (total integrated intensity in a defined region or number of peaks detected in a defined region). In each case the output of the analysis was subsequently manipulated using the same table generation and graph plotting software RS/1 [36] running on a DEC-VAX cluster. Pattern recognition and mapping methods were carried out using the program ARTHUR [37] on a VAX cluster. Data communication between computers was by TCP/IP. Significance testing and other statistical methods used Fisher weights and Student's *t*-tests.

Mapping of the samples was achieved using principal components (PC) analysis, a well-known multivariate exploratory data analysis technique. PC analysis [1] is a method which as well as devising the most informative descriptors in a data set irrespective of data classification, can also be used for dimension reduction. The PCs are eigenvectors from diagonalization of the covariance matrix ($p \times p$) of the $n \times p$ data matrix (based upon p descriptors for each of n samples). The first PC is a linear combination of the original p descriptors with appropriate weighting coefficients and contains the maximum variance in the data. The second PC is another linear combination of variables, orthogonal to the first and contains the next most complete description of the data. Successive PCs will, therefore, explain less and less of the data variance and at some point will consist of only the data noise. For a set of p descriptors there are p PCs but clearly plotting the sample coordinates for the first two PCs (a so-called scores plot) will provide the maximum information content of the data in two dimensions. Three-dimensional PC stereo-plots can also be useful in complicated data sets [31]. In this mode of exploratory analysis, no *a priori* assumptions are made about the samples and the sample classes are marked on the maps to investigate whether the descriptors are capable of distinguishing the classes. One advantage of PC analysis is that it is possible to determine how much each of the original descriptors contribute to the various PCs and hence which are important in explaining any sample groupings seen in the plot.

Results

An evaluation of the effectiveness of the

various methods of segmenting NMR spectra has been carried out, including a study of the effects of varying NMR processing and segment data reduction parameters. A set of 15 600 MHz ^{1}H NMR spectra of rat urine were used to test the software. These consisted of five control samples, five from animals which had received a single dose of an experimental nephrotoxic compound at dose 1 and five which had received an increased dose of the same compound, dose 2. Earlier work has shown that the controls can be separated from the drug treated animals using the levels of 26 predefined metabolites visible in the urine by NMR spectroscopy [14].

Figure 1 shows typical spectra for the control and drug treated classes of sample, indicating the level of difference in the spectra, together with a typical segment output based on integrated intensity in each segment. Differences between the spectra shown in Fig. 1(a) and (b) can be ascribed mainly to the effect of the dosed toxic substance and details of that toxicological study will be reported separately.

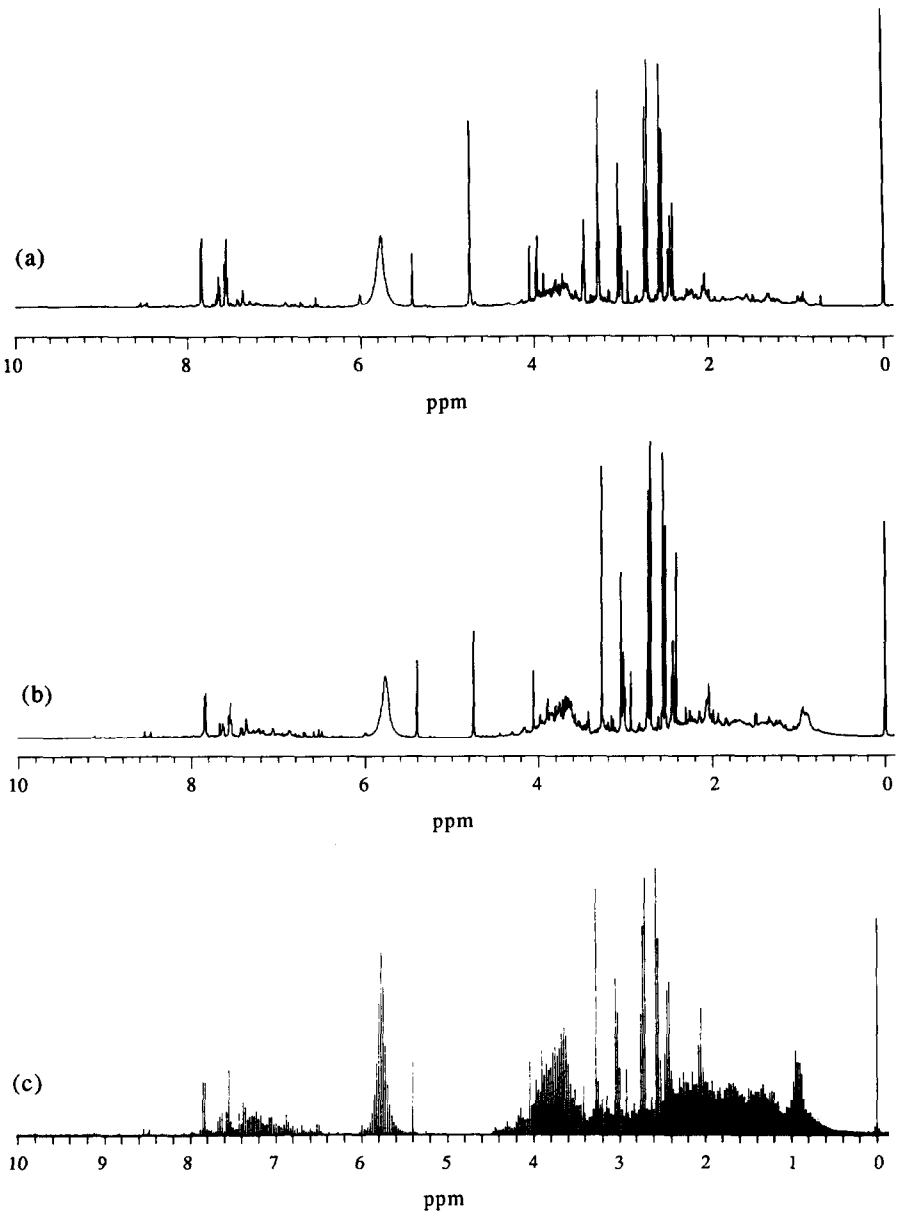

Figure 1
600 MHz ^{1}H NMR spectra of typical (a) control rat urine, (b) rat urine after dosing with a kidney toxin at dose 2, (c) summed peak integral distribution corresponding to (b) based on defined spectral segments.

NMR descriptors for this data set have now been generated using the procedures given above. The region of the spectrum between δ5.0 and 4.5 has been excluded from the analysis because it contains the distorted, suppressed water resonance. If a 0.05 ppm segment is chosen covering a range from δ10.025 to −0.125, this results in 203 descriptors for each spectrum with TSP in the middle of a segment. Three different types of NMR descriptor have been investigated. These are with the segment value defined as (a) the number of peaks in the segment, (b) the total integrated area in the segment and (c) the summed peak heights in the segment. In (b) and (c) in the present study, the values were scaled to the corresponding value for the reference compound TSP (δ0.0) as a fixed amount of this was added to each sample. Therefore, segment values in (b) will be directly related to metabolite concentrations. In other applications, it is possible to conceive situations where other types of data scaling would be more appropriate, for example, scaling to the total summed integral over all segments where effects of dilution need to be eliminated.

Preliminary investigations have shown that, in this data set, no sample classification is apparent using descriptors based on the number of peaks in each segment. This is not unexpected for this data set as there is no *a priori* reason why the number of peaks in a given segment should correlate with the toxin effects on biochemistry. In other applications such as in imaging or chromatography, the number of objects within a segment may be diagnostic. However, classification was possible using either of the other two approaches, namely integrated intensities and peak heights.

Figure 2 shows a plot of the first two principal components for the 15 samples using a 0.05 ppm segment based on integrated intensities and with various exponential weighting factors applied to the NMR data before Fourier transformation. It should be borne in mind that it is not necessarily the highest variance PCs which will show discrimination

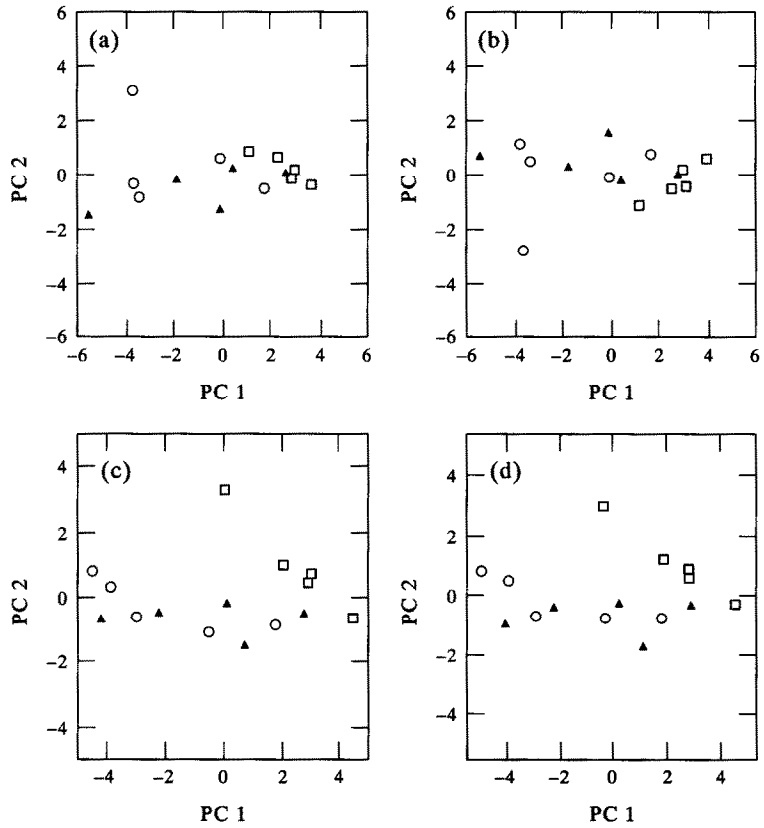

Figure 2
Plot of the first two principal components for the rat urine samples with automatic descriptor generation based on segment integrals. Segment regions of 0.05 ppm and with (a) no line broadening; (b) 5.0 Hz line broadening; (c) 25.0 Hz line broadening; and (d) 60.0 Hz line broadening. Key: □ — control urine, ○ — dose 1 urine, ▲ — dose 2 urine.

between classes, but given that 80–89% of the data variance is contained in the first two PCs in all of the test cases studied, the other PCs are largely modelling the noise in the data and will be less discriminating. In fact, for the test cases, three-dimensional maps based on the first three PCs did not improve the classification. The axis values of the PC plots depend on the number and range of the descriptors used and it is only the relative position of the points (each representing one NMR spectrum) which need to be considered. For zero line-broadening (Fig. 2a), the controls appear as a tight cluster, which also contains one sample from the dose 2 group. The urines from the drug treated animals, whether from dose 1 or dose 2, are intermixed on the map. Modest increases in the applied line-broadening of 5 Hz have little effect on the clustering apart from inverting PC2 (Fig. 2b). When the line-broadening factor approaches the segment width, the separation of the control group from the drug treated groups is improved (Fig. 2c, line-broadening 25 Hz, segment width 30 Hz) and increasing the line-broadening further has a negligible effect on the class separation (Fig. 2d).

A second manipulation of the data which could affect the segmented output is the need to phase an NMR spectrum to pure absorption mode. This has been investigated by manually rephasing a given spectrum five times during a working day, carrying out a segment analysis and measuring the standard error on the segment integrals. In regions where only noise exists (these regions are usually subsequently removed in the analysis) the error is of the order of 25%, but for segments containing resonances, the variation in the measurement due to manual phasing is about 1% of the total segment integral.

The effect of varying the segment width was

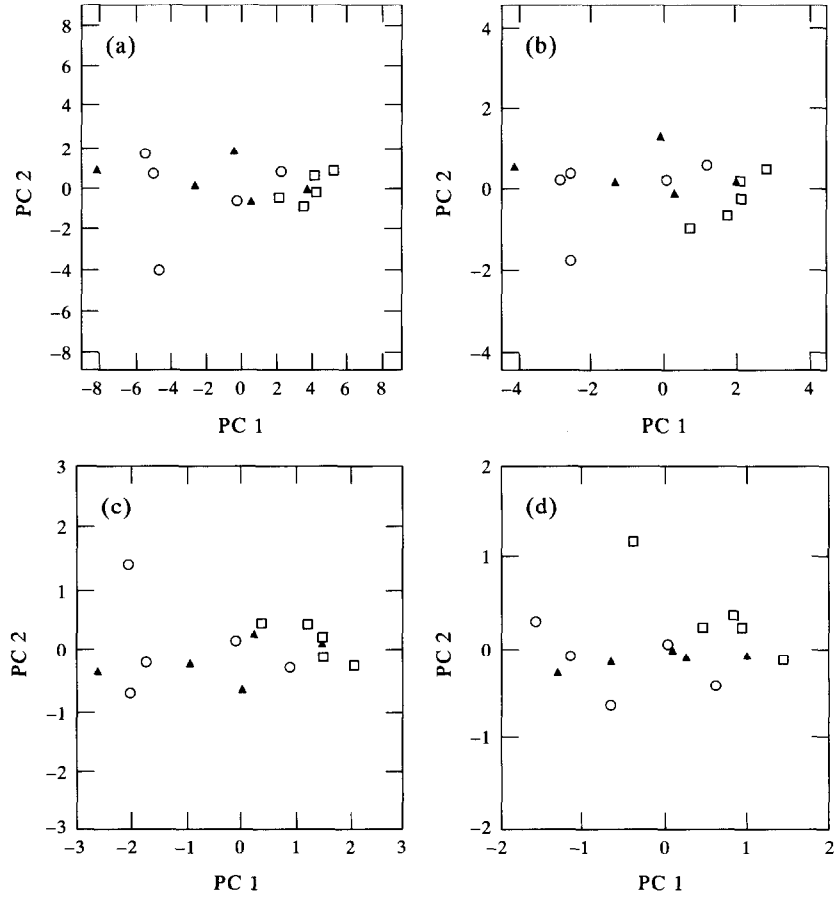

Figure 3
Plot of the first two principal components for the rat urine samples with automatic descriptor generation based on segment integrals. Line broadening of 1.0 Hz and with various segment region widths: (a) 0.025 ppm; (b) 0.1 ppm; (c) 0.2 ppm; and (d) 0.5 ppm. Key as for Fig. 2.

also investigated using the test set of spectra with a fixed line-broadening of 1 Hz. Figure 3 shows the result of this evaluation. The absolute PC axis values vary because the number of descriptors is now very different in each case. Figure 3(a) indicates the result of using a 0.025 ppm segment. The use of a 0.05 ppm segment (not shown), a 0.1 ppm (Fig. 3b) and a 0.2 ppm segment width (Fig. 3c, PC2 now inverted) also gave very similar results. Extension to a 0.5 ppm segment gave an improved separation of the control group as shown in Fig. 3(d), although one of the samples from the dose 2 group still appeared in the control region of the map.

A facility was built into the software to investigate the effect of allowing the segments to overlap. For a 0.05 ppm segment and a line-broadening of 1 Hz, allowing the segments to overlap by up to 20% had no effect on the maps. This overlap process can be thought of as increasing inter-parameter correlation and it is therefore not surprising that PC analysis, a decorrelation technique, is insensitive to this type of data reduction. Other multivariate analysis methods, particularly supervised techniques, may be highly sensitive to correlated data.

An estimate has been made of the class separation which can be effected by the use of a very coarse segment resolution. In this case, a 0.5 ppm segment with an exponential FID weighting equivalent to a 60 Hz line-broadening gave the PC plot shown in Fig. 4, with good separation of the control group of samples including separation of the sample from the dose 2 group not achieved in Fig. 3(d).

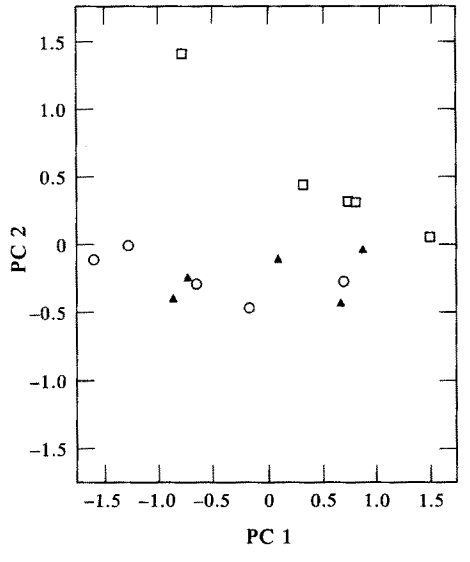

Figure 4
Plot of the first two principal components for the rat urine samples with automatic descriptor generation based on segment integrals. Line broadening of 60.0 Hz, segment region width of 0.5 ppm. Key as for Fig. 2.

The third approach of using summed peak heights is illustrated in Fig. 5. In this case, each segment region contains the summed heights of all peaks detected above a threshold equal to three times the peak–peak noise level. Figure 5 shows the effect of varying the segment width on the classification and the results can be compared with those given in Figs 3 and 4. As can be seen, the peak height method also yields similar classification results to using integrals within a segment. For example, comparison of the maps produced by the integral and peak heights approaches for a segment width of 0.2 ppm is given in Fig. 3(c) and Fig. 5(c), respectively. The information content of the first two PCs is very similar, in that one of the dose 2 samples is grouped within the control samples in both cases, and the relative disposition of the other dose 1 and dose 2 samples is also very similar, with the control group showing a tighter clustering in the peak height derived map. The results of using other segment widths are also comparable (Figs 3 and 5).

An illustration of the application of the automatic method of generating descriptors is shown in Fig. 6 using integrated areas. Each point on these plots involving the first three principal components corresponds to the 600 MHz ^{1}H NMR spectrum of a human urine sample. A cluster of samples from normal individuals is seen and a number of separate groups arising from samples taken from patients with a variety of inborn errors of metabolism can be seen. None of the inborn error of metabolism samples overlaps with the normal group using three dimensions and the various types of inborn error appear in different parts of the map. The use of the first three PCs was necessary in this case to achieve full class separation (the data variance explained by the PCs being 61% (first PC), 79% (first two PCs) and 85% (first three PCs)). The normal samples were taken from a study of normal physiological variance and the inborn error of metabolism samples constitute part of a larger study using NMR of urine on this subject. Both topics will be reported separately together with the details of the pattern recognition optmization approaches used.

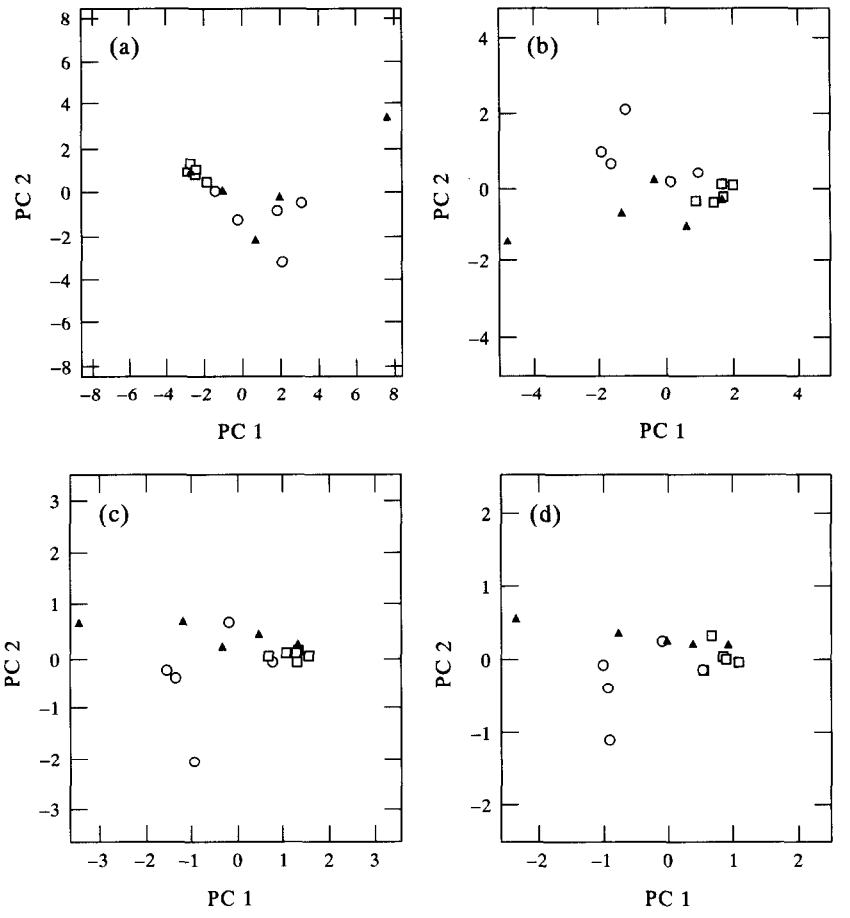

Figure 5
Plot of the first two principal components for the rat urine samples with automatic descriptor generation based on segment summed peak heights. Line broadening of 1.0 Hz and with various segment region widths: (a) 0.025 ppm; (b) 0.1 ppm; (c) 0.2 ppm; and (d) 0.5 ppm. Key as for Fig. 2.

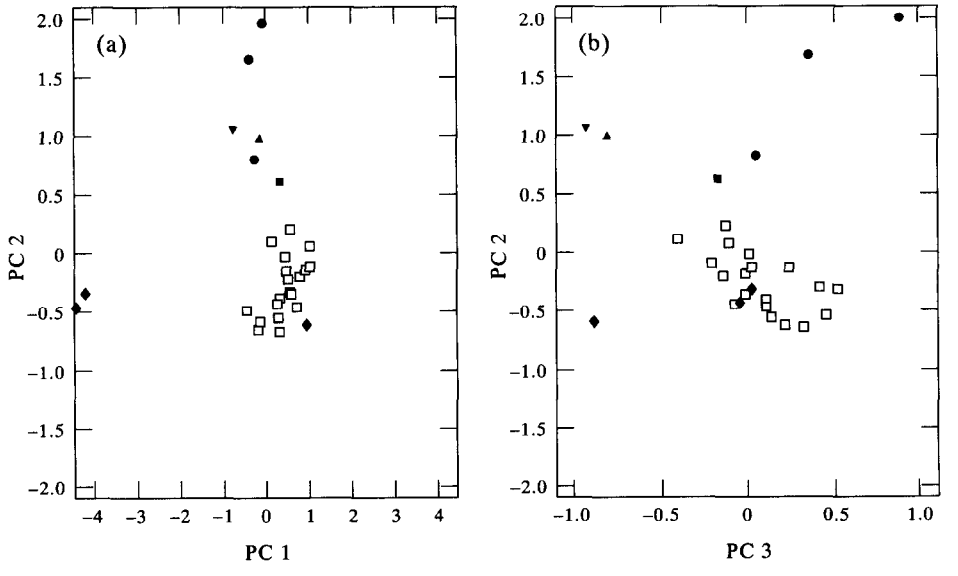

Figure 6
Illustration of the discriminating power of automatically generated NMR descriptors on human urine samples. Plot of the first three principal components for the samples with automatic descriptor generation based on segment integrals, (a) PC2 vs PC1 and (b) PC2 vs PC3. Key: □ — normal, ● — cystinuria, ▲ — porphyria, ■ — 5-oxoprolinuria, ▼ — Fanconi syndrome, ◆ — oxalic aciduria.

Discussion and Conclusions

Individual resolved peaks in an NMR spectrum reflect resonances of chemically distinct nuclei and the peak areas can be related to the concentration of the respective atom types and hence compound if the resonances can be assigned. However, in a typical complex biofluid NMR spectrum, even if measured at the highest possible field strength, there are many signals which are only partially resolved and others which only appear or disappear in certain pathological conditions. In such circumstances, the extraction of parameters which reproducibly represent the spectra is far from ideal.

One approach which retains the characteristics of a peak area is to integrate signals over adjacent chemical shift windows of equal width. No signals are lost but the resulting integrals will represent the sum of all signals and parts of signals within the window boundaries. If the window width is narrower than a resolved peak, then all peaks will occupy more than one window, and if the window width is comparable to the peak width, then any window may contain one peak or part of a peak plus parts of additional peaks where there is peak overlap.

Another approach is to record peak heights. In a group of partially resolved signals, the peak height for a dominant signal will be fairly well preserved, but smaller signals in the wings of the major peak may fail to be picked out as peaks, and will not be represented. In order to enable corresponding peaks in different spectra to be perceived as related (bearing in mind small possible differences in the associated chemical shift values between samples, the absence of peaks in some cases or the appearance of new species in other cases), an equivalencing procedure is needed, such as the assignment of each peak to one of a series of adjacent chemical shift windows together with the summing of the peak heights within each window, thus providing a set of windowed peak height sums across all spectra which may provide a data matrix for pattern analysis.

These two approaches have been investigated here. More sophisticated methods can be conceived, such as the use of maximum entropy deconvolution or artificial intelligence peak equivalencing methods, but they would seem less readily suited to automation, or they would need a more fundamental degree of development and evaluation.

The approaches described in this work have been demonstrated to yield a robust and rapid method for generating descriptors from ^{1}H NMR spectra which can be used as input to pattern recognition routines for sample classification. The method relies on altered concentrations of individual metabolites and it might be argued that concentration-dependent chemical shifts would mean that a given metabolite could appear in different segments in different samples. Whilst this is possible, the observed concentration dependence of chemical shifts in such complex mixtures is negligible. On the other hand, many metabolite chemical shifts are pH dependent and a spread in pH over a sample set could cause ambiguity in that the exact segment location of a given metabolite could be variable and might indicate the need for controlled sample pH. However, one benefit of the use of the current approach is that the segment width can be chosen to ensure generally that any expected pH-dependent chemical shifts are encompassed within the segment width. The application of the automatic data reduction approach may have widespread application for the automatic analysis of complex mixtures in diverse fields such as biofluids, wines, juices and any complex liquid mixture. The methodology could also be adapted for solid state NMR or *in vivo* NMR or to other forms of spectroscopy. In addition, in principle it can be applied to heteronuclear NMR spectroscopy and can be adapted to two-dimensional and multi-dimensional experiments, and in particular should be useful for sample class separation based upon heteronuclear inverse ^{1}H-^{13}C NMR correlation spectra.

Three types of automatic descriptor have been tested, comprising segment values based on the number of peaks within a given region, total summed peak heights within a region and total summed integrals within a region. For the test data set, no classification was observed when peak numbers were used as descriptors. Good classification was possible using either summed peak heights derived from a peak-pick file or by using integrals in a spectral region. The use of integrals provides a general approach which will be applicable in many applications but which is likely to be susceptible to baseline distortions unless effective baseline corrections are carried out. On the

other hand the use of peak height data derived from a prior peak-picking exercise does not suffer so much from baseline problems and allows easy referral back to the spectrum to identify those NMR peaks and hence, in the present case, endogenous biochemicals which are responsible for the class separation. This could be achieved through the application of supervised learning methods where the sample class is input to the algorithm in such approaches as SIMCA [20].

The high-field high-resolution NMR spectra of complex mixtures are in general too rich in information content for a full interpretation of the features by eye, thus providing for the chance that important classification information may be missed in a trivial selective approach. In addition, manual data reduction is very time consuming, is open to the possibility of bias and peak overlap may lead to errors in quantitation and biochemical over-interpretation, which even so may only be drawn from a small fraction of the latent information in the spectrum.

The automatic segment generation approach appears to offer a rapid and robust method of primary data reduction giving at least comparable results to manual approaches and at the same time encapsulating more of the latent information in the spectrum. An interesting speculation is possible on the relative information loss caused by the reduction of an NMR spectrum of for example 64K data points to 500 segments or the reduction from 500 segments down to one or a few principal components. If the latter is a bigger compression, then it may be expected that the use of the present automatic methods will be very robust. Nevertheless, it is possible that the data compression methods described here could lead to some information loss, but as long as the information which is important for sample classification is retained, then the methods are adequate. It is worth pointing out that NMR spectra of biofluids contain, as well as regions with no information content, much redundant information in that any one molecule will often give rise to several resonances, the areas of which are ideally related to the concentration of the molecule and the number of nuclei giving rise to each peak. In addition, many resonances will be further split by spin–spin coupling. Thus many of the peak intensities in a biofluid NMR spectrum will show high levels of correlation. There is a likelihood of retaining this information and hence correlation after the data compression procedure since the signal intensities are preserved. Furthermore, additional redundancy may arise in certain specific situations where changes in metabolite levels are correlated, for example through linked biochemical processes. However, one of the benefits of using the pattern recognition approach is that it is possible to explore, take advantage of or remove, if necessary, such correlated effects within a spectrum and to discover the underlying classifying descriptors.

References

[1] M.A. Sharaf, D.L. Illman and B.R. Kowalski, *Chemometrics*. John Wiley & Sons, Chichester, UK (1986).

[2] J.F. Zilva and P.R. Pannal, *Clinical Chemistry in Diagnosis and Treatment* 4th Ed. Lloyd–Luke, London, UK (1984).

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

[4] J.K. Nicholson and I.D. Wilson, *Prog. Drug Res.* **31**, 427–479 (1987).

[5] O.A.C. Petroff, R.K. Yu and T. Ogino, *J. Neurochem.* **47**, 1270–1276 (1986).

[6] J.K. Nicholson, M.J. Buckingham and P.J. Sadler, *Biochem. J.* **211**, 605–615 (1983).

[7] J.R. Bales, D.P. Higham, I. Howe, J.K. Nicholson and P.J. Sadler, *Clin. Chem.* **30**, 426–432 (1984).

[8] B.C. Sweatman, R.D. Farrant, E. Holmes, F.Y. Ghauri, J.K. Nicholson and J.C. Lindon, *J. Pharmaceut. Biomed. Anal.* **11**, 651–664 (1993).

[9] P.J.D. Foxall, G.J. Mellotte, M.R. Bending, J.C. Lindon and J.K. Nicholson, *Kidney Int.* **43**, 234–245 (1993).

[10] J.R. Bales, J.D. Bell, J.K. Nicholson and P.J. Sadler, *Magn. Reson. Med.* **3**, 849–856 (1986).

[11] J.D. Bell, J.C.C. Brown, R.E. Norman, P.J. Sadler and D.R. Newell, *NMR in Biomed.* **1**, 90–94 (1988).

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

[13] R.A. Iles, A.J. Hind and R.A. Chalmers, *Clin. Chem.* **31**, 1795–1801 (1985).

[14] R.D. Farrant, J.C. Lindon, E. Rahr and B.C. Sweatman, *J. Pharmaceut. Biomed. Anal.* **10**, 141–144 (1992).

[15] K.-P. Neidig, R. Saffrich, M. Lorenz and H.R. Kalbitzer, *J. Magn. Reson.* **89**, 543–552 (1990).

[16] M. Kjaer and F.M. Poulsen, *J. Magn. Reson.* **94**, 659–663 (1991).

[17] B. Meyer, T. Hansen, D. Nute, P. Albersheim, A. Darvill, W. York and J. Sellers, *Science* **252**, 542–544 (1991).

[18] J.U. Thomsen and B. Meyer, *J. Magn. Reson.* **84**, 212–217 (1989).

[19] J.-M. Nuzillard and G. Massiot, *Anal. Chim. Acta* **242**, 37–41 (1991).

[20] S. Wold, *Pattern Recog.* **8**, 127–139 (1978).

[21] G. Okide, D.S. Weber and W.J. Goux, *J. Magn. Reson.* **96**, 526–540 (1992).

[22] U. Edlund and H. Grahn, *J. Pharmaceut. Biomed. Anal.* **9**, 655–658 (1991).

[23] D. Johnels, U. Edlund, E. Johansson and M. Cocchi, *J. Chem. Soc. Perkin Trans. II* 1773–1777 (1989).

[24] L.S. Anker and P.C. Jurs, *Anal. Chem.* **64**, 1157–1164 (1992).

[25] S. Kruse, O.M. Kvalheim, G. Gadeholt, L. Halsteinslid and E. Sletten, *Chemometrics and Intell. Lab. Systems* **11**, 191–196 (1991).
[26] S.L. Howells, R.J. Maxwell and J.R. Griffiths, *NMR in Biomed.* **5**, 59–64 (1992).
[27] S.L. Howells, R.J. Maxwell, A.C. Peet and J.R. Griffiths, *Magn. Reson. Med.* **28**, 214–236 (1992).
[28] D. Morvan, P. Jehenson, D. Duboc and A. Syrota, *Magn. Reson. Med.* **13**, 216–227 (1990).
[29] S. Confort-Gouny, J. Vion-Dury, F. Nicoli, P. Dano, J.-L. Gastaut and P.J. Cozzone, *C.R. Acad. Sci. Paris* **315**, 287–293 (1992).
[30] K.P.R. Gartland, S.M. Sanins, J.K. Nicholson, B.C. Sweatman, C.R. Beddell and J.C. Lindon, *NMR in Biomed* **3**, 166–172 (1990).
[31] K.P.R. Gartland, C.R. Beddell, J.C. Lindon and J.K. Nicholson, *Mol. Pharmacol.* **39**, 629–642 (1991).
[32] M.L. Anthony, PhD Thesis, High resolution NMR spectroscopic and pattern recognition approaches to the biochemical characterisation of experimental nephrotoxicity states. London University (1993).
[33] K.P.R. Gartland, C.R. Beddell, J.C. Lindon and J.K. Nicholson, *J. Pharmaceut. Biomed. Anal.* **8**, 963–968 (1990).
[34] E. Holmes, J.K. Nicholson, F.W. Bonner, B.C. Sweatman, C.R. Beddell, J.C. Lindon and E. Rahr, *NMR in Biomed.* **5**, 368–372 (1992).
[35] E. Holmes, F.W. Bonner, B.C. Sweatman, J.C. Lindon, C.R. Beddell, E. Rahr and J.K. Nicholson, *Mol. Pharmacol.* **42**, 922–930 (1992).
[36] RS/1. BBN Software Products, UK Ltd, Staines, Middlesex, UK.
[37] ARTHUR 81. Version 4.1. B & B Associates, Seattle, WA, USA.

[Received for review 21 March 1994; revised manuscript received 3 May 1994]