Journal of Pharmaceutical & Biomedical Analysis
Vol. 9, No. 8, pp. 655–658, 1991
Printed in Great Britain

0731–7085/91 $3.00 + 0.00

Multivariate data analysis of NMR data*

ULF EDLUND†‡ and HANS GRAHN§

‡*NMR Research Group, Department of Organic Chemistry, Umeå University, S-901 87 Umeå, Sweden*
§*Astra Research Centre, S-151 85 Södertälje, Sweden*

Abstract: Multivariate methods based on principal components (PCA and PLS) have been used to reduce NMR spectral information, to predict NMR parameters of complicated structures, and to relate shift data sets to dependent descriptors of biological significance. Noise reduction and elimination of instrumental artifacts are easily performed on 2D NMR data. Configurational classification of triterpenes and shift predictions in disubstituted benzenes can be obtained using PCA and PLS analysis. Finally, the shift predictions of tripeptides from descriptors of amino acids open the possibility of automatic analysis of multidimensional data of complex structures.

Keywords: *Multivariate data analysis; PCA/PLS; NMR data reduction; shift predictions; 2D NMR.*

Introduction

The impressive development of NMR spectroscopy during the last decade has positioned NMR to be the leading spectroscopic technique. Due to high magnetic fields and improved probe designs, NMR can be classified as an analytical technique for many nuclei and the introduction of 3D/4D techniques combined with isotopic labelling (^{15}N, ^{13}C) has steadily increased the upper limit for molecular complexity that could be studied by NMR [1].

This condition coupled to the fact that structural and dynamic information is to be found in many NMR parameters (chemical shifts, coupling constants, relaxation times, NOE, etc.) has resulted in a need for models that could extract the information content in relation to specific questions. Numerous "hard" models based on kinetic or thermodynamic conditions, connectivities, library data banks, etc. have been suggested [2]. The most recent need for automatic analysis of 2D/3D data of large proteins has clearly pointed out the weakness of such "hard" approaches.

In our opinion "soft" models such as those based on principal components (PCA, PLS) are to be preferred in handling NMR data. The complexity of the spectral responses and the fact that NMR is far from robust compared with other spectroscopic techniques are both strong arguments in favour of soft modelling. In practice it is impossible, for instance, to obtain a 2D NOESY spectrum of a reasonably large protein without getting spurious peaks in the spectrum.

In this report we will give several examples how PCA/PLS methods can be used in NMR spectroscopy, from processing of NMR spectra, identification or assignment of signals to structural classification problems.

Results and Discussion

Data processing

Besides the problem with overlapping peaks, NMR spectroscopy suffers sometimes from signal-to-noise limitations. Characteristic noise includes random thermal noise and systematic noise ridges. The t_1 noise, usually the most troublesome artifact in 2D NMR spectroscopy, appears as bands of spurious peaks running parallel to the t_1 spectral dimension through large signals in the data such as peaks from methyl groups or solvent lines. Causes of t_1 noise include instrumental instabilities (field/frequency instability in particular), signal truncation and numerical errors in fast Fourier transform. These noise ridges pose serious difficulties for accurate signal assignments because these ridges introduce "false" peaks while obscuring "true" ones. By the combination of experimental methods for solvent peak suppression and various data-processing

*Presented at the Symposium on "Chemometrics in Pharmaceutical and Biomedical Analysis", November 1990, Stockholm, Sweden.

†Author to whom correspondence should be addressed.

procedures such as symmetrization and noise-profile subtraction it is possible to extract information that would otherwise be hidden by noise bands. However, quite a few of these techniques can pose large difficulties in interpreting the data, because some processing schemes introduce artifacts.

In this study we have chosen a multivariate representation of the entire 2D spectra in conjunction with principal component analysis. PCA is used to approximate the spectrum to a degree which retains the systematic spectral features while excluding random noise. PCA can also be used to form a mathematical model of the spectral noise bands. This noise model is later subtracted from the original spectrum, leading to substantial suppression of the noise (Fig. 1).

Shift assignments

In an early study of ^{13}C NMR substituent chemical shifts (SCS) of monosubstituted benzenes it was found that the majority of substituents belonged to one out of four

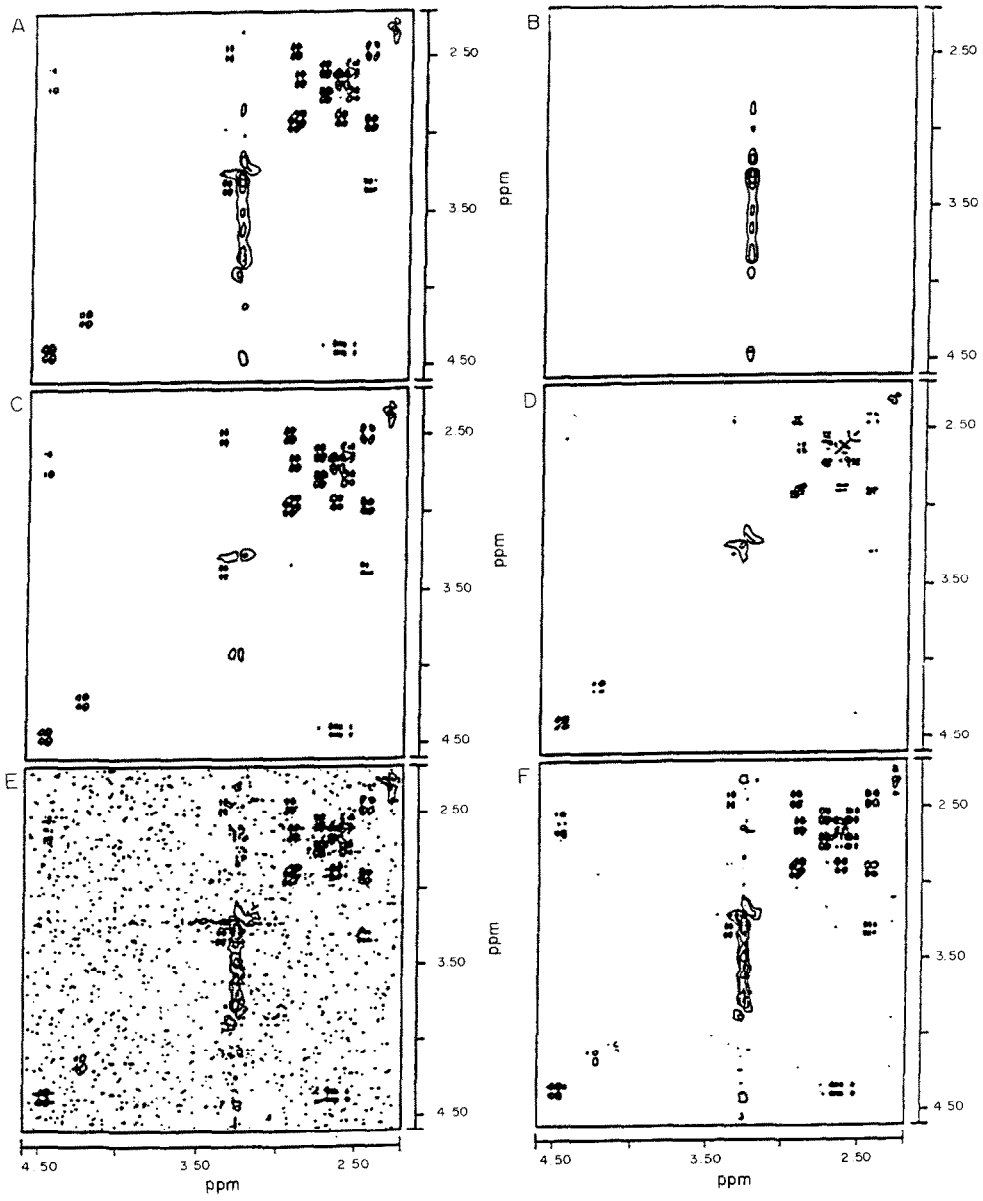

Figure 1
(A) The aliphatic region of Trp-Phe-Trp-Ala is shown (DQF-COSY, DMSO-d_6). Note the strong t_1 ridge, which is originating from the solvent signal at 3.3 ppm. (B) Second principal component, reconstruction of region in (A). (C) Original data with noise ridge components subtracted. (D) Symmetrization of original data. (E) Original data plus simulated noise, S/N = 5:1. (F) Reconstruction of spectrum in (E) with 10 principal components, S/N = 12:1.

groups, alkyls, acceptors, donors or halogens [3]. Only a limited number of "unusual" substituents were found in between these clusters. Hence, by using local models for each class of substituents, a significantly improved description of SCS was found compared with global models such as dual substituent parameter models. This view has been frequently criticized, but a positive outcome of our investigation is that the many reports published later actually test for clustering before using a unifying model [4]. Unfortunately, the support for a unifying view of SCS has been based on data sets where the "well-behaved" objects are reduced in number and "odd" substituents in between the claimed clusters are more frequent.

To illustrate the strength of partial least-squares (PLS) data analysis as a tool to quantify analogy reasoning for shift assignments, the non-additivity of SCS in 1,3- and 1,4-disubstituted benzenes was specifically studied [5]. By using the ^{13}C SCS of monosubstituted benzenes, it was shown that the dependent "non-additivity SCS" matrix could be fully described by the monosubstituted SCS pattern. This explicitly means that no additional effects are necessary to explain the non-additivity in these disubstituted benzenes, an important simplification to remember in molecular design schemes.

Two dimensional NMR of large molecular systems is an area where there is a special need for efficient strategies for peak assignments. One approach is to transform the normal 2D output into a peak table, where in principle each data point in one dimension is an object defined by the intensity value along the other dimension (variables). Such intensity matrix can be treated in a PCA manner and weakly coupled spin systems can be identified [6]. To illustrate this approach, we show how the unambiguous spin topology information in a double-quantum-filtered COSY spectra of leucine enkephalin (Tyr-Gly-Gly-Phe-Leu) can be extracted. By interpreting PCA score plots it was possible to successfully separate the seven detectable spin systems. These score plots show the spin systems as a series of mutually orthogonal hyperplanes. In the PCA score plots, the spin systems in tyrosine, phenylalanine and leucine are separated from each other; the two glycine spin systems occur together in a single hyperplane. It is important to note that, based on topology of this DQF-COSY spectrum, it is possible to separate the overlapping glycine systems without resorting to information from other experiments.

It must be stressed that the steps to generate these results were performed in a totally automated way, including the steps of spectral processing, peak picking, construction of the multivariate matrix representation and PCA.

There are several problems any automatic or semi-automatic approach must surmount in order to provide a practical method for 2D spectral interpretation of biomolecules. Overlapping peaks, spectral changes due to temperature or pH variation, missing peaks, etc. put strong demands on the used methodology. This should be added to the fact that changing an amino acid in the vicinity of given amino acid in a protein will induce significant shift changes. We have approached this problem by tabulating a large number of shift data for all protons in combinations of three amino acids. By considering that there are 23*22*23 three unit sequences, or residue combinations, we have currently tabulated about 20% of all possible combinations. This is already a sizeable share of the total "search space". In addition to this we have included a large number of physical descriptors for each amino acid and also for the neighbouring amino acids. These descriptors contain various properties ranging from molecular size to electronic effects for the individual amino acids.

By using soft modelling methods of the data set(s) it was possible to classify each proton class, and the technique provides a good tool for evaluating the significance of the proposed assignments.

Classification

A very common situation in NMR is to find suitable probes, for instance specific signals, whose parameters (shifts, coupling relaxation) you hope monitor a certain molecular property, let us say configuration. For a series of compounds, the NMR information related to your question (configuration) could be hidden in effects due to a variable substitution. The required information might also to a varying degree be distributed to many positions; a condition, which hardly could be detected by traditional eye-balling. Such a situation is prevalent in a series of pentacyclic, pharmacologically interesting triterpenes [7]. Although the main skeleton was substituted differently in five positions we were able to (1) identify

which positions discriminated between α/β configuration, and (2) to classify unknowns in a statistical correct manner. It is evident that a soft modelling procedure of this kind has clear advantages over other "empirical" one-variable rules.

To summarize, the high information content and the spectral sensitivity to effects such as temperature, concentrations, etc. are all factors that strongly favour soft modelling approaches in handling NMR data.

Acknowledgement — We thank Mr Frank Delaglio, NMRi Inc., Syracuse, for valuable suggestions.

References

[1] G. Wagner, in *Progress in NMR Spectroscopy* (J.W. Emsley, J. Feeney and L.H. Sutcliffe, Eds), Vol. 22, pp. 101–139. Pergamon Press, Oxford (1990).
[2] J. Kalchhauser and W. Robien, *J. Chem. Inf. Comput. Sci.* **25**, 103 (1985).
[3] D. Johnels, U. Edlund, H. Grahn, S. Hellberg, M. Sjöström, S. Wold, S. Clementi and W.J. Dunn III, *J. Chem. Soc., Perkin Trans.* **2**, 863 (1983).
[4] O. Exner and M. Budesinsky, *Magn. Reson. Chem.* **27**, 27 (1989).
[5] D. Johnels, U. Edlund, E. Johansson and M. Cocchi, *J. Chem. Soc., Perkin. Trans.* **2**, 1773 (1989).
[6] H. Grahn, F. Delaglio, G.C. Levy and M.A. Delsuc, *J. Magn. Res.* **77**, 294 (1988).
[7] H. Grahn, C.J. Masino, B. Nordén and U. Edlund, *Magn. Reson. Chem.* **26**, 1097 (1988).

[*Received for review 26 November 1990*]