PEAK DETECTION AND QUANTIFICATION IN NMR SPECTRA USING THE PIQABLE ALGORITHM

Sarah J. Nelson and Truman R. Brown

Department of NMR
Fox Chase Cancer Center,
7701 Burholme Avenue
Philadelphia, PA 19111.

INTRODUCTION

The identification of peaks and quantification of peak positions, heights and areas are critical steps in the interpretation of all NMR spectra. Conventional methods of performing these analyses have relied heavily on manual input for phasing, baseline removal and definition of peak margins. For studies involving large datasets, such as the determination of protein structure from high resolution 2-D or 3-D ^{1}H spectra, the development of reliable, automatic methods for "peak-picking" would save considerable time and provide a convenient input for newly proposed pattern recognition techniques (1-4). In the case of *in vivo* localized spectroscopy using, for example, chemical shift imagings, the problems associated with multi-dimensional datasets are compounded by poor spectral resolution, low signal to noise ratio and significant baseline distortions. PIQABLE is an automatic algorithm (6,7) which can be used to detect and quantify peaks in both of these types of data. In this presentation we will first outline the major features of PIQABLE and then illustrate its applications.

OVERVIEW OF PIQABLE

The two steps in the PIQABLE algorithm are firstly separation of statistically significant peaks from baseline and noise components and, secondly, the estimation of peak parameters. Identification of peaks is based upon the minimal assumption that a spectrum comprises three components: baseline, peaks and random noise.

$$D_i = B_i + P_i + E_i.$$

The baseline, B_i, is assumed slowly varying in at least one dimension. Peaks are univariate (1-D spectra) or bivariate (2-D spectra) regions which are statistically significantly biased above or below the baseline component. For unfiltered data, the noise is random and uncorrelated with a gaussian distribution of constant or slowly changing variance. If the data are prefiltered (e.g. by apodization), the power spectrum of the filtered noise must be taken into account in identifying peaks. PIQABLE uses a simple iterative technique to estimate the baseline and noise components, while determining the regions of the spectrum containing peaks.

Non-parametric estimates of peak parameters are obtained by searching the baseline-subtracted spectrum within individual peak regions and identifying statistically significant maxima. An adaptation of Marshall's algorithm (8) is used to test for the existence of partially overlapping peaks. Fully overlapping peaks cannot be separated at this stage. To quantify the contributions from overlapping peaks or test for the existence of "shoulders," parametric assumptions about peak shape are required. For analysis of *in vivo* spectra in particular, great caution must be exercised in modelling peak shapes. The two common assumptions of either Lorentzian or Gaussian shapes may be used, but it is essential that statistical tests are applied to determine whether the estimated model parameters provide a representation which accurately describes the data. PIQABLE's non-parametric estimates of number of maxima, maximum positions and inter-valley areas are available as initial parameter estimates for non-linear least squares fitting routines. Because the noise distribution is Gaussian, the values obtained provide the Maximum Likelihood estimates of model parameters. Suppose there are M peaks and each peak shape is described by parameters $\underline{a}_k$, then we are assuming in the frequency domain

$$D_i = B_i + \sum_{k=1}^{M} c_k P(\underline{a}_k) + E_i$$

or alternatively, in the time domain

$$d_i = b_i + \sum_{k=1}^{M} c_k p(\underline{a}_k) + e_i$$

where the small letter denotes Fourier transform of the corresponding component.

Fitting could be achieved either in the time or frequency domains. For data with Lorentzian peak shapes, both manual and automatic baseline estimates may include a significant portion of the peak tails in the baseline (7,9). PIQABLE corrects for this when performing fits. The advantages of using PIQABLE over

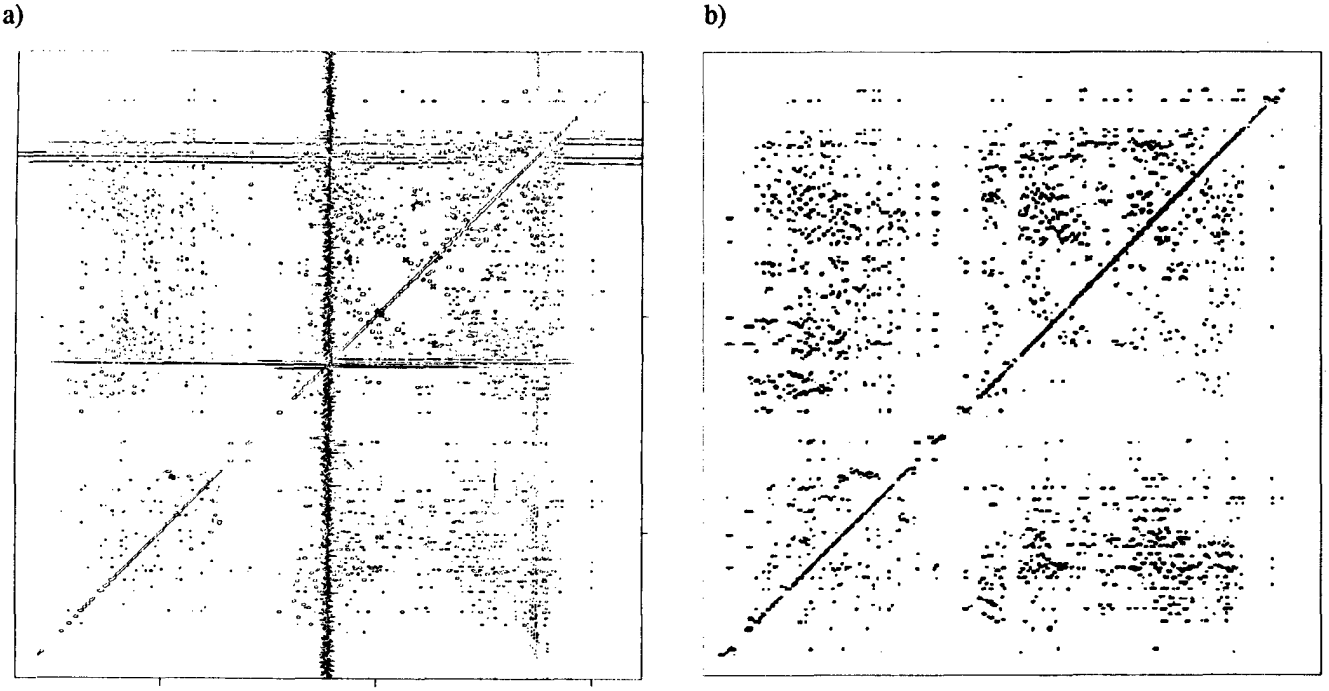

Figure 1. Ability of PIQABLE 2 to isolate regions of a 2-D ^{1}H NOESY spectrum which contain peaks: a) contour plot of 500 MHZ spectrum of ubiquitin and b) regions of the spectrum identified as containing peaks.

direct fitting procedures are the additional assumption of a significant baseline component and the use of non-parametric statistics to get automatic estimates of baseline and initial values of peak parameters. Prior knowledge concerning the number or relative positions of peaks can be included either in the initial peak detection stage or during the fitting procedure by assuming that some of the model parameters are known.

APPLICATION TO PEAK DETECTION IN 2-D SPECTRA OF PROTEINS

The two dimensional version of PIQABLE can separate bivariate peaks from univariate ridges and random noise. This is extremely useful for peak detection in 2-D ^{1}H spectra of proteins where each spectrum may contain several hundred bivariate peaks and low level univariate ridges. This is illustrated by the application to a 2-D ^{1}H NOESY spectrum of the 76 amino acid protein ubiquitin. These data have been used in structural studies using the Main Chain Directed Procedure (4,10). Figure 1a shows a contour plot of the spectrum and Figure 1b the regions detected as having statistically significant peaks. In this case, the same PIQABLE parameters were used throughout the spectrum. In practice, because the spectrum is heterogeneous in nature, it may prove more efficient to use region specific parameters. This is particularly true for areas where the spectrum departs from the standard PIQABLE assumptions; very close to the diagonal and along the solvent track. A comparison of PIQABLE2 peak detection with manual analysis is presented elsewhere in this volume (11). In addition to the obvious time saving and reproducibility of using an automatic technique, the accuracy of peak positions determined by PIQABLE2 is significantly better than that obtained manually. Future studies will apply PIQABLE2 to the analysis of J-correlated spectra and will investigate its natural extensions to consider 3-D data.

PRODUCTION OF METABOLIC IMAGES

The analysis of *in vivo* localized spectra acquired using chemical shift imaging (CSI) provides major challenges to quantitative procedures. These include the multi-dimensional nature of the datasets, the low signal to noise ratio and the distortion of peak shapes due to loss of initial time points in each FID. An important part of the analysis is finding a method of representing the results which can be readily interpreted. We have found that for multi-dimensional datasets with sufficient spatial resolution, images of the spatial distribution of individual metabolites are extremely useful.

We have used an adaptation of the PIQABLE algorithm to derive metabolic images from ^{31}P CSI data. To obtain accurate phasing parameters and prior information as to the approximate location of individual peaks, we make use of a non-localized ^{31}P FID which is acquired with enough acquisitions to give good S/N but otherwise the same parameters as the CSI data. Thus, the first step of the analysis is perform a Fourier transform of

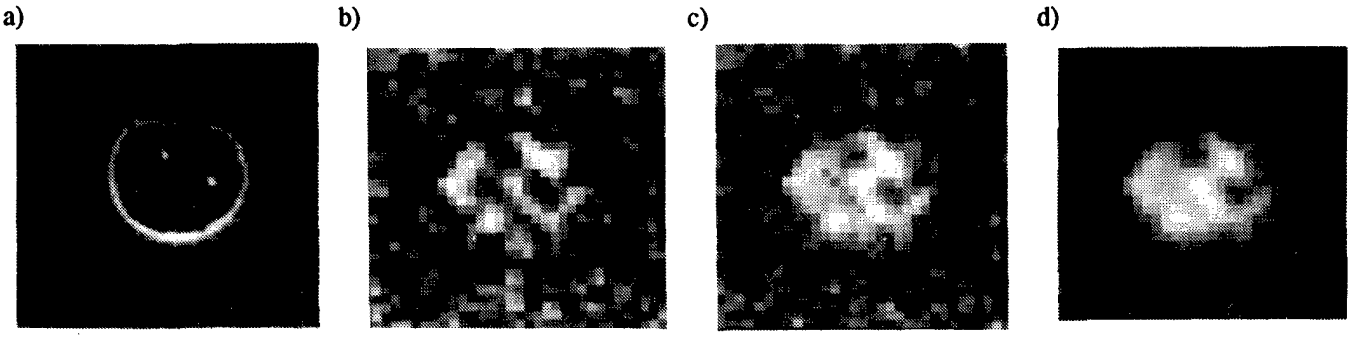

Figure 2. Different processing steps applied to CSI data to obtain metabolic images of inorganic phosphate from the forearm of a volunteer: a) proton image showing CSI voxel sizes, b) no filtering or early time point reconstruction, c) Wiener filtering only, d) Wiener filtering plus early time point reconstruction. The data were obtained from the forearm of a volunteer using a Siemen's Magnetom operating at 1.5T.

the non-localized FID and quantify it with PIQABLE. Although field inhomogeneity and variations in chemical shift of the metabolites can cause the linewidths and peak positions in individual voxels to differ from the values obtained from the non-localized data, the overall peak regions are still a good guide to regions in the CSI spectra which will contain peaks.

For a totally non-parametric analysis, the raw CSI data are Fourier transformed (in time and k-space) and individual spectra analyzed using PIQABLE with pre-defined peak regions. Peak heights or peak areas are then displayed as images. In practice, the CSI data are relatively low signal to noise and improved metabolic images can be obtained by k-space filtering and reconstructing the missing initial time points in each FID. We have obtained good results with the following procedures. First, approximate Wiener filters are determined which are tailored to balance the signal and noise in space voxels for the particular metabolic of interest. Data are then zero-filled to give a 32 by 32 image resolution and the spatial Fourier transforms applied. Second, the missing initial time points are reconstructed for each spatial voxel using a modification of PIQABLE. In this modification, initial peak parameters are estimated using a short baseline estimator which can follow the sinc wiggle introduced by the absence of early time points. The contribution of the missing time points to the frequency domain is estimated, assuming that the FIDs comprise a sum of exponentially decaying sine waves. This contribution is added onto the original spectrum and peak parameters re-estimated. As a substantial part of the sinc component has been removed, a longer baseline estimator can be used for the second quantification. In practice, stable peak estimates are obtained by iterating this procedure 3-4 times. Figure 2 shows examples of metabolic images obtained from a human forearm by using these processing steps on the low S/N inorganic phosphate (Pi) peak. The data were obtained by placing a 12.5 cm single turn coil around the arm and collecting a 16 x 16 2-D CSI dataset with nominal spatial resolution 1cm by 1cm. The drop-off in coil sensitivity was used for localization in the third dimension, giving estimated voxel sizes of approximately 8 ml. The CSI data were collected with 8 acquisitions per voxel, giving a total acquisition time of 34 minutes. Based upon a proton image collected just before the CSI data, the two areas within the arm which have no signal correspond precisely to the radius and ulna.

Encouraged by these results on such a low S/N peak, we have since improved the positioning and shimming to give the images shown in Figure 3. On the left is a proton image, in the middle is a Pi image and on the right is a phosphocreatine (PCr) image. Note the significant localized changes in intensity of both metabolites. Since acquiring these data, we have succeeded in obtaining good quality PCr images from CSI data with as low as 1 acquisition per k-space voxel i.e. collection of a CSI dataset in 4.5 minutes. With this time and spatial resolution comes the possibility of obtaining a time course of localized spectroscopy data; results of such experiments will be reported elsewhere (12). These studies demonstrate the feasibility of using PIQABLE-based automatic processing procedures to obtain ^{31}P metabolic images from the CSI data. Potential applications include the visualization of metabolic inhomogeneity in the region of a tumor and studies of localized changes in metabolite levels during specific muscle exercise.

ACKNOWLEDGMENTS

This work was supported by NIH grants CA06927 and CA41076 and by an NSF grant DIR-8904066 to SJN. We wish to think Dr. A. J. Wand for allowing us to use the 2-D spectrum of ubiquitin and Drs. J.S. Taylor, J. Murphy-Boesch and D. Vigneron for providing us with the CSI data.

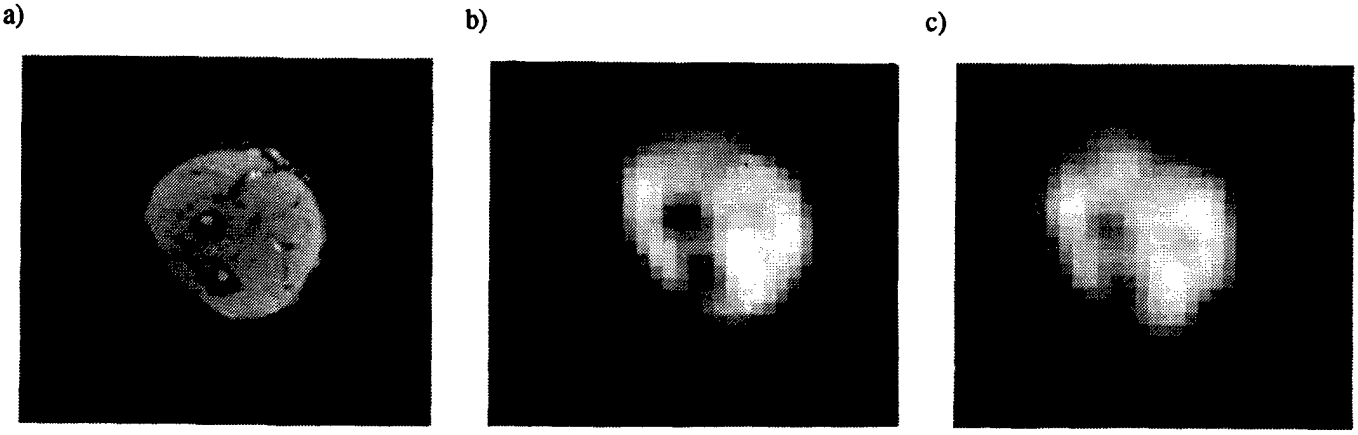

Figure 3. Images from a human forearm: a) proton image, b) metabolic image of Pi and c) metabolic image of PCr. The data were acquired as previously but with improved positioning.

REFERENCES

1. Meier U, Bodenhausen G. and R. R. Ernst J., Magn. Reson., 60, 161, (1984).
2. Pfandler P. and G. Bodenhausen. J. Magn. Reson., 70, 71, (1986).
3. Pfandler P. and G. Bodenhausen. J. Magn. Reson., 79, 99, (1988).
4. Wand A. J. and S. J. Nelson., In "NMR and X-Ray Crystallography: Interferences and challenges." M. C. Eher, editor. AIP, New York, in press.
5. Brown T.R., Kincaid B. M., and K. Ugurbil. Proc. Natl. Acad. Sci. USA, 79, :3523 (1982).
6. Nelson S. J. and T. R. Brown. J. Magn. Reson. 75, 229, (1987)
7. Nelson S. J. and T. R. Brown. J. Magn. Reson. 84, 95, (1989)
8. Marshall R. J. Computed Biomedicine Reson. 19, 319, (1986)
9. Meyer R. A., Fisher M. J., Nelson S. J. and T. R. Brown. NMR Biomedicine 1, 131, (1988)
10. DiStefano D. and A. J. Wand. Biochemistry 26, 7272, (1987)
11. Nelson S. J., Schneider S. and A. J. Wand. Refinement and Automation of the Main Chain Directed Assignment Procedure for the Analysis of 2-D ^{1}H spectra of proteins (this volume.)
12. Brown T. R., Taylor J. S., Murphy-Boesch J. Vigneron D, Heddema T., Kosinski E and S. J. Nelson Eighth Annual SMRM, Amsterdam Works in Progress 1031, (1989).