J Biomol NMR (2007) 37:265–275
DOI 10.1007/s10858-006-9136-4

ARTICLE

A general algorithm for peak-tracking in multi-dimensional NMR experiments

P. Ravel · G. Kister · T. E. Malliavin · M. A. Delsuc

Received: 7 September 2006 / Accepted: 12 December 2006 / Published online: 10 February 2007

Abstract We present an algorithmic method allowing automatic tracking of NMR peaks in a series of spectra. It consists in a two phase analysis. The first phase is a local modeling of the peak displacement between two consecutive experiments using distance matrices. Then, from the coefficients of these matrices, a value graph containing the a priori set of possible paths used by these peaks is generated. On this set, the minimization under constraint of the target function by a heuristic approach provides a solution to the peak-tracking problem. This approach has been named GAPT, standing for General Algorithm for NMR Peak Tracking. It has been validated in numerous simulations resembling those encountered in NMR spectroscopy. We show the robustness and limits of the method for situations with many peak-picking errors, and presenting a high local density of peaks. It is then applied to the case of a temperature study of the NMR spectrum of the Lipid Transfer Protein (LTP).

Keywords Peak tracking · Protein · Graph theory

Introduction

NMR spectroscopy is the tool of choice for the study of molecules in solution. Compared to other spectroscopic techniques, it is carried out directly in solution with no modification of the system under study; it does not require a complex preparation step such as crystallization; and finally, it allows straightforward analysis of interactions, as well as the gathering of detailed structural information. In certain multi-dimensional spectra, such as an HSQC or HNCO, one spectral line corresponds to one protein residue, and any modification of the characteristics of a peak is the sign of a molecular event.

In protein studies, it is common practice to perform a series of NMR spectra under varying chemical or physical conditions, such as pH, ligand concentration, temperature or external pressure. On such a series, a common task consists in analyzing the peak displacements and modifications during the experiment. This can be a complex task, as many peaks may be present in the spectra. Moreover, because of fortuitous line overlap and additional spurious signals, NMR spectra of large systems tend to be crowded. Consequently, peak detection is not always obvious. Certain peaks are overlooked by peak detection methods, while non-existing peaks are noted as detected. The method presented in the present work aims at assisting the

P. Ravel · M. A. Delsuc
CNRS UMR5048, Centre de Biochimie Structurale, 34090
Montpellier, France

P. Ravel · M. A. Delsuc
INSERM U554, 34090 Montpellier, France

P. Ravel · M. A. Delsuc
Université Montpellier 1 et 2, 34090 Montpellier, France

G. Kister
Laboratoire de Physique Industrielle, Faculté de Pharmacie,
Montpellier, France

T. E. Malliavin
Unité de Bioinformatique structurale, Institut Pasteur, 25-28
rue du docteur Roux, 75724 Paris Cedex 15, France

P. Ravel (✉)
Faculté de Pharmacie, 15 avenue Charles Flahault, 34 000
Montpellier, France
e-mail: ravel@univ-montp1.fr

spectroscopist in this task. Additionally, the trajectory of each peak is monitored and conveniently reported for further analysis. Such an analysis may also be a means to overcome chemical shift ambiguities by adding a new dimension (ie. the physical parameter, temperature or pressure) to the spectrum. This approach was shown (Malliavin et al. 2001) to be promising in the frame of automatic spectral assignment.

In the case of a rapid equilibrium, the characteristics of an NMR signal are a weighted sum of the characteristics of the species in equilibrium. If the varying external parameter affects this equilibrium, the net effect is a peak displacement. In certain types of multidimensional spectra, peaks correspond to atoms connected by one-bond linkages, and the information gathered by such spectra is local. In this kind of spectra, the shift in peak position in each dimension reflects perturbation on the associated spin and is expected to be proportional to this perturbation and thus, roughly linear. Consequently, the analysis we developed is based on the assumption of small linear peak displacements, (Zuiderweg 2002; Akasaka 2003).

Former work on peak tracking mainly consist in comparing a reference spectrum with a perturbed test spectrum. In a first attempt for NOESY spectra comparison, Pielak et al. (1988) used a simple approach where overlaid and difference spectra are computed in order to determine the unchanged test peaks. In more recent work (Williamson et al. 1997; Muskett et al. 1998), peaks picked in the test spectrum are matched against their nearest neighbor in the reference spectrum. However, no allowance is made for large displacements or for experimental artifacts in the test spectrum. Ross et al. (2000) used a different approach: starting from a large number of test spectra obtained with different ligands and a reference spectrum; they detected the spectral modifications by a principal component analysis of the bucketed spectra. Recently, Peng et al. (2004) proposed the APET/PROPET techniques, as an extension of previous work (Kumar et al. 1998; Accelrys 2002). In their work, through careful peak-picking from the reference spectrum and considering peak displacement as well as peak shapes, they use a tree search and simulated annealing optimization method, to search the peak mapping corresponding to a global minimum of the spectral difference. An approach to overlapped peaks through filtering unmatched test peaks based on the statistics of peak shapes is also proposed. The authors finally demonstrated their approach on the titration of a SH3 domain by a peptide ligand. However, PROPET implements a progressive spectral pair comparison, which does not allow a more comprehensive analysis of peak overlap or of trajectory details.

The following approach is proposed here: from each spectrum acquired during the variation experiment, a list of peaks is generated. From these peak lists, peak paths are generated in a two step procedure. The first step is local, it relies on creating matrices of transition distances between two successive peak lists. For each peak of a given list, all the neighboring peaks in the next list are defined. And, in order to avoid a combinatorial explosion, only the restriction of the geometric components of these matrices in relation to a critical neighborhood calculated statistically is considered. The second step is global; from the coefficients of the transition matrices, a graph containing a set of possible paths followed by the peaks is generated. A score is associated with each path of the graph. This score measures the quality of the paths based on the various hypotheses made, and an optimization is performed to maximize the overall score of the set of paths. At this stage, additional peaks are created in order to handle paths with missing peaks. From the list of all the paths collected in step 2, each observed peak is either noted as belonging to a probable path, or as artifactual. When several peak paths have crossing trajectories, potential peak overlaps are considered, and a volume additivity law is used to determine which peak are actually overlapped.

A program called General Algorithm for Peak Tracking (GAPT) was developed which implements this algorithm. A simulator was also included in the program, and was used to extensively test the program and to validate statistical assumptions. Finally, the program was applied to a temperature variation followed by recording ^{15}N-HSQC on the Lipid Transfer Protein (LTP).

Methods

NMR

The protein studied is the ns-LTP2, a 67 residues protein, from *triticum Aesticum* (Pons et al. 2003). A ^{15}N labeled sample was produced in *P. Pastoris* and purified as previously described in (de Lamotte et al. 1999). A 2 mM sample, with 1.2 equivalents of lysophosphatidyl glycerol myristoyl was prepared in 90% H_2O 10%D_2O at pH 3.5. The NMR experiments were acquired on a Bruker Avance 600 spectrometer, equipped with a TXI probe operating at 599.93 MHz. The temperature was varied from 300 K to 310 K by 2.5 K steps, and five experiments were obtained from this variation. For each temperature, the sample was allowed to settle for 30 min after the temperature change, and an ^{15}N-HSQC spectrum was acquired. 128 FID were acquired, with 8 scans. Peak picking was performed by detecting all local maxima above a given threshold. Detected peaks were first filtered to remove obvious artifacts, then each peak was fitted to a Gaussian 2D line to precisely determine peak coordinates, widths and amplitudes. The peak volume was also used. It was defined as the product of the intensity by the widths. Datasets were processed and peaks were picked using the *Gifa* NMR processing software (Pons et al. 1996).

Peak-tracking

The problem is generalized as follows: a sample is subject to a varying physical parameter P, and M different NMR experiments are recorded using a value $\Pi(k), k \in [1, M]$. It is assumed that the NMR spectra, of spectral dimension n, contain the same number N of actual signals evolving under the varying experimental conditions. Each NMR spectrum is analyzed with a standard peak picker, such as AUTOPSY (Koradi et al. 1998), Sparky (Goddard and Kneller 1999), or *Gifa* (Pons et al. 1996) and a peak list $L(k)$, containing $N(k)$ peaks is produced for each spectra of the series. Due to experimental errors, the $N(k)$ are usually different from N. The details of the peak picking process are considered outside the scope of this study.

It is considered that each peak is characterized by n spectral coordinates f_i and by n' other parameters α_i such as amplitude, phases or widths. The ith peak of the kth peak list is noted:

$$X_i(k) = (f_1, \ldots, f_n, \alpha_1, \ldots, \alpha_{n'}) \tag{1}$$

The tracking of the displacement and of the modifications of a peak from one experiment to another determines a peak path c, defined as a list containing one entry in each peak list $L(k)$. It will be noted

$$c = (X_u(1), \ldots, X_v(k), \ldots, X_w(M)) \tag{2}$$

Note that the index of the peak X may vary from one list to another since the peak picking operation is highly dependent on the data.

All the possible peak paths found during the analysis are handled as a whole as the set:

$$C = \{c_q\}. \tag{3}$$

This set will be detailed below.

Figure 1 presents the standard situations which may arise when trying to build peak paths. These situations are the simplest ones, and the reality is generally a mixture of them.

For each peak present or absent in the peak lists, we define a status as follows :

- True positive peak, i.e.; a peak meaningful to the experimenter, (Fig. 1a).
- False negative peak: a peak meaningful to the experimenter but missing in one peak list (Fig. 1b).
- False positive peak: a peak resulting from an artifact in the experiment, (Fig. 1c).
- Overlapping peaks: several peak paths share the same peak, (Fig. 1d, e).

For each peak list, it is then possible to define the following relationship between the length of the peak list $N(k)$, the number of false positive peaks N_+ (k), the number of false negative peaks N_- (k), the number of overlapping peaks N_o (k), and the actual number of expected signals N:

$$N = N(k) - N_+(k) + N_-(k) + N_o(k) \tag{4}$$

Step 1: Local approach

The tracking of a peak X_i (k) (Eq. 1) from the list $L(k)$ to the list $L(k + 1)$ consists in determining the following peaks X_j $(k + 1)$ which are related to the initial peak, X_i (k). The $N(k + 1)$ peaks must be considered as potential followers for each peak of the $L(k)$ list. A transition matrix, measuring the displacement length for each possibilities from $L(k)$ to $L(k + 1)$, is used to reduce this complexity.

We call $T(k)$ the kth transition matrix between the peak lists $L(k)$ and $L(k + 1)$. The coefficients $T_{ij}(k)$ of the $T(k)$ matrix are equal to the Euclidean distance between the normalized spectral components $(f_1, \ldots, f_n)$ of the two peaks, $X_i(k), X_j(k+1)$ to which was added the variation of the physical parameter, $\Pi(k)$. From the knowledge of the matrix coefficients, it is possible to determine for each peak $X_i(k)$ a critical neighborhood, centered on the peak and of radius $\rho(k)$ which contains the most probable following peaks. The radius $\rho(k)$ allows the simplification of the study. A small $\rho(k)$ reduces the search space, but may generate errors by overlooking peaks.

In order to estimate $\rho(k)$ we will study the following statistical series $G(k)$

$$G(k) = \left\{ \min_{1 \leq j \leq N(k+1)} (T_{ij}(k)); 1 \leq i \leq N(k) \right\}, \quad 1 \leq k \leq M \tag{5}$$

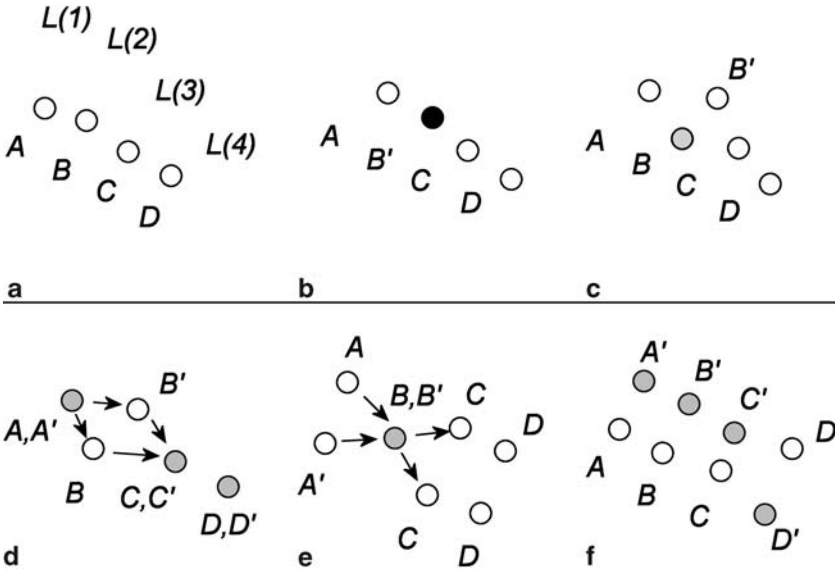


Fig. 1 Canonical situation observed during peak list analysis; examples are given with 4 peak lists: $L(1),\ldots,L(4)$ (**a**) True positive peak path, (A, B, C, D). (**b**) Path (A, B′, C, D) with a false negative peak B′ (missing peak). (**c**) Path (A, B′, C, D) with a false positive peak B′ next to a complete peak path (A, B, C, D). (**d**) Two paths (A, B, C, D) and (A, B′, C, D) with overlapping peaks A,C D. (**e**) Two paths (A, B, C, D) and (A′, B, C′, D′) with overlapping peak B. (**f**) Two paths crossing (A, B, C, D) and (A′, B′, C′, D′)

which describes the distributions of the minimum distance between two successive lists of peaks. The critical radius $\rho(k)$ is an upper bound of the confidence interval of the mean of the statistical series $G(k)$ and is defined as follows:

$$\rho(k) = \overline{d(k)} + 5\sigma(k); \quad 1 \leq k \leq M-1 \tag{6}$$

where $\overline{d(k)}$ and $\sigma(k)$ are respectively the average and the standard error of $G(k)$ (Eq. 5). A large number of numerical simulations were performed to evaluate the form of the probability law, supposing that the values of the geometric characteristic of the peaks follow the uniform law. The $G(k)$ series is not a normal law, and the coefficient 5 in the confidence interval insures that in 99.9% of the cases, the peak displacement from one experiment to the other is less than the critical radius.

In order to reduce the combinatorial complexity of the problem, the transition matrix $T(k)$ is modified to $\widetilde{T}(k)$ by applying the critical radius as a threshold value:

$$\widetilde{T}_{ij}(k) = \begin{cases} T_{ij}(k) & \text{if} \;\; T_{ij}(k) \leq \rho(k) \\ -1 & \text{otherwise} \end{cases} \tag{7}$$

The use of a threshold value for reducing the complexity of the problem is very efficient, however it is valid only under the hypothesis of small displacements

Step 2: Global approach

From the transition matrices $\widetilde{T}(k)$ (Eq. 7), peak paths are built by simply connecting peaks related by the transition matrices, and by following all the possible connections. However, because some relevant peaks are not detected by peak-picking, due to the peak picking uncertainty, certain peaks are isolated, and this will result in partial paths. Such isolated peaks are easily labeled by looking at the values of the transition matrices: if one line (respectively one column) contains only negative values, this is the sign that the corresponding peak has no detected predecessor (respectively follower). In order to suppress the isolated peaks, each one is duplicated as a virtual peak in the respective peak list, and the transition matrices are expanded accordingly. This procedure is repeated until no isolated peaks remain.

After this procedure, all the possible peak paths of C (Eq. 3) are built from the transition matrices with a recursive graph search algorithm (Gondran and Minoux 1985). Thanks to the virtual peak extension, all peak paths in C are defined from the first to the last experiment, and all detected peaks are included in at least one complete path. However, a given peak is usually shared by several peak paths in C, whenever this peak has several possible neighbors within the critical radius. This feature makes it possible to naturally handle fortuitous peak overlaps which may occur in one experiment, as long as the peaks are separated in another experiment.

Linearity constraint

At this point, the problem means choosing the subset of the most likely peak paths within C. It is thus necessary to score each path, taking into account altogether the invariance of the peak parameters such as volume or widths, the number of virtual peaks involved, as well as the linearity of the peak evolution.

First, for a given path c (Eq. 2) composed of M peaks (detected or virtual), a function $l(c)$ is built as follows:

$$l(c) = \frac{1}{(N^{*}-1)(N^{*}-2)}\left[\sum_{i=1}^{n'+1} \omega_i F_i(\widetilde{x}_i)\right]$$

where

$$F_i(\widetilde{x}_i) = \sum_{k=1}^{M-2} \|\widetilde{x}_i(k)\widetilde{x}_i(k+1)\|\|\widetilde{x}_i(k+1)\widetilde{x}_i(k+2)\| - \overrightarrow{\widetilde{x}_i(k)\widetilde{x}_i(k+1)} \cdot \overrightarrow{\widetilde{x}_i(k+1)\widetilde{x}_i(k+2)} \qquad (8)$$

where N^* is the number of detected peaks in the path, the ω_i are the weights associated with each peak parameters, and $\widetilde{x}_i(k)$ is a composite point containing the ith spectral parameter of the kth experiment, with the index i running on all the peak parameters including the spectral coordinates.

The rational behind this equation is that, for a linear trajectory, each parameter should evolve proportionally to the external perturbation. Departures from this linear behavior can be related to the angle α made by the $(k, k+1)$ and $(k+1, k+2)$ steps, as presented in Fig. 2.

With the function $l(c)$ defined in Eq. 8, the score of a path c increases when the peak departs from a regular and linear trajectory. This function is real positive. It is the sum of the elementary weighted scores $\omega_i F_i(\widetilde{x}_i(k))$ associated with the different spectral parameters (spectral position, amplitude, area, width). The weights ω_i are the variation coefficients computed from a statistical analysis of the $G(k)$ distribution, as presented above (Eqs. 5 and 6).

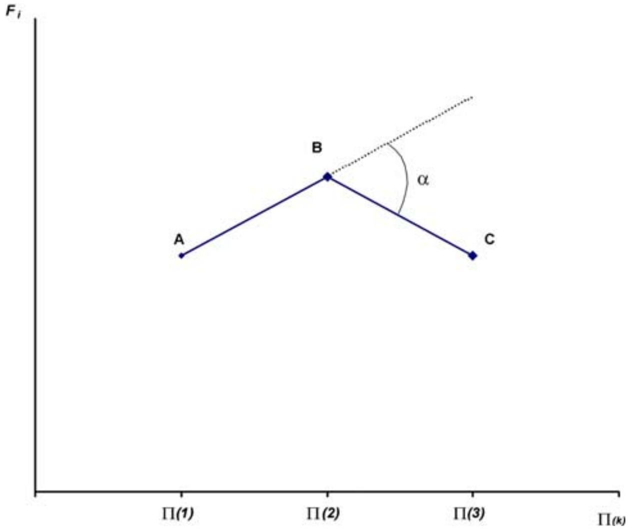


Fig. 2 Example of calculus of the elementary function score F_i, $i > 1$ applied to a three peaks path (A, B, C). In this case, $F_i = \|AB\|\|BC\| - \overrightarrow{AB} \cdot \overrightarrow{BC} = \|AB\|\|BC\|(1 - \cos(\alpha))$.. Whenever the path is linear ($\alpha \approx 0$) or whenever the AB or BC vector norm is small, F_i is minimal

In Eq. 8, $\widetilde{x}_i(k)$ is a composite value, built from the peak parameters as follows: for $i = 1, \widetilde{x}_1(k)$ is a point of $n + 1$ components built from the n spectral coordinates f_i, augmented with the value of the varied physical parameter $\Pi(k)$; for $i = 2, \ldots,, \widetilde{x}_i(k)$ is a 2-dimensional point, built from the ith peak parameter value augmented with the value of the varied physical parameter $\Pi(k)$. Thus $\widetilde{x}_i(k)(i > 1)$ is a 2-dimensional point for parameters such as peak volume or width, but is an $n + 1$-dimensional value for the spectral coordinate parameters ($i = 1$). Finally, paths with many virtual peaks are penalized due to the $1/N^*$ weight.

It should be noted that Eq. 8 introduces a generalized definition of a linear trajectory encompassing all the NMR peak features rather than just the spectral coordinates. By including the value of the perturbation $\Pi(k)$ to the peak parameter values, the function $l(c)$ enforces the fact that modifications of a peak parameter by a given external perturbation is proportional to the size of the perturbation, and bears the same proportionality from step to step.

To give an example, if the peak width is augmented by 10% with a 0.5 pH unit variation, it should change by 20% to a 1 pH unit variation and by 40% to a 2 pH unit variation.

Such linearity is common in NMR spectroscopy when weak or medium strength interactions are taking place. When a strong interaction is present, and a marked non-linear sigmoïdal titration curve is expected, this approach is no longer valid and is hence, likely to fail.

Scoring the linearity of the peak paths with the $l(c)$ function is quite efficient for separating true peak paths, however it is not sufficient when many false positive and false negative peaks are present in the initial peak lists.

Crossing peak paths

Due to spectral crowding, it is common to observe the crossing of two peak paths. Depending on the situation, the effect is observed either as two peaks in the list $L(k)$ overlap in $L(k + 1)$, or as one peak in $L(k)$ splits in two peaks in $L(k + 1)$ (Fig. 1d, e). In all cases, if the peaks are real, the peak volumes add. On the other hand, if some of the peaks are erroneous, the additivity will not hold, and both peak paths will be improved by the removal of the false peaks. Based on this additive property, two additional tests where added to detect crossing peak paths, these tests are used to modify the list of possible peak paths C.

Test for overlapping peaks

If two paths c_1 and c_2 share one or several peaks, all the peak volumes V of both paths are modified and defined as: $V(k) = (V_1(k) + V_2(k))/2$ if the peaks are separated and as: $V(k) = V_1(k)/2$ if the peak is shared. If the scores of these modified paths are significantly better than the scores of the initial ones, the shared peaks between c_1 and c_2 are considered as overlapping and the modified volumes are kept and the set C of all possible peak paths updated accordingly.

Test for false positive peaks

If two paths c_1 and c_2 share one peak, the shared peak is removed from the c_1 peak list, and replaced by a virtual peak. If the score of this modified path is significantly better than the score of the unmodified one, then the initial path is replaced by the modified one in C. This test is applied symmetrically to c_2 with respect to c_1.

For both tests, the score will be considered as significantly better by considering the impact of the tests on the part of the score function computed for the volumes. To be considered as positive, the test should improve this score by a value larger than the standard deviation of the volume scores computed from 80% of the best paths from the current list. This empiric rule seems to give satisfactory results.

Resolution of the minimization problem on graph C through an exhaustive method

At this stage, the problem is reduced to finding a set of N peak paths among the set of possible paths C for which the score $l(c)$ (Eq. 8) will be minimal. As the number of paths in C is not very large, the resolution of this minimization problem on C is tackled here by a best first search algorithm (Laveen et al. 1989).

This minimization is performed in three stages.

First, all the peak paths in the set C are sorted with respect to the value of the score function $l(c)$ regardless of peaks shared between paths.

Second, all peak paths involving shared peaks are re-evaluated, starting with lowest score value, and considering the two tests described above (test for overlapping peaks and test for false positive peaks). The combinations which yield the best score are conserved, and the set C is modified accordingly. At the end of this process, all the peak paths are corrected for shared peaks, and are sorted with respect to the value of the score function. The N first entries of this list are then considered as real peak paths.

Third, the status of all peaks is obtained by considering the N best peak paths obtained in C. All the peaks are considered as real, either true positive if the peak was present in the initial peak list $L(k)$ or true negative if the peak is virtual and was built during the optimization process. Some peaks may be shared by several paths, and are considered as overlapping peaks. Isolated peaks and sparse paths (paths consisting of only one or two non-virtual peaks) are considered as false positives and given an infinite score.

Applications

The program GAPT was developed to implement the algorithm presented in the preceding sections. It includes two distinct parts: a simulator and a peak tracker.

The simulator generates situations close to real experiments. The program takes as input the entire set of the parameters as defined in the methods section, namely, Π the external parameter describing the change in experimental conditions, M the number of experiments, N the number of expected paths, n' the number of parameters characterizing a peak as well as the evolution of the former in relation to Π. A probability law for each of these variables can be given. Finally, the generation of false positive and false negative peaks is monitored by two error rates $\delta_+ (k)$ and $\delta_- (k)$ which are defined respectively by the following ratio $N_+(k)/N$ and $N_-(k)/N$. After the peak generation phase, peak overlap is simulated by merging peaks located within a user defined distance from each other.

Thanks to a large number of simulations, it has been possible to estimate the statistical distributions of $G(k)$ (Eq. 5) and thus to calculate the critical radius $\rho(k)$ (Eq. 6). The test for overlapping peaks and the test for false positive peaks were also studied statistically with the aid of the simulator. The simulator was then used to generate a large number of experiments which were used to blind test the peak tracker.

The GAPT peak tracker implements the algorithm presented above. It takes input from peak tables coming from the simulator or from the analysis of an experiment. Along with the data, the user supplies the number N of expected peaks. At the end of the processing, GAPT provides the user with the paths associated with the most likely peak trajectories and the status of each experimental peaks. The choices taken by the algorithm at each step, are also returned to the user.

Two applications are detailed below. The first application is a simulation; the results show the performances and limits of the proposed method as a function of the complexity of problems under study. Four variable parameters are used: the crowding rate, the false positive rate, the false negative rate and the number of experiments M. The second application is the temperature variation study of the interaction of the ns-LTP1 protein liganded to a phospholipid.

Example of simulations

NMR experiments were simulated, taking into account the experimental hypothesis of small peak displacements along a quasi-linear trajectory. Several experiments with a varying external parameter were simulated, and noise was added to every peak parameters using a uniform law of amplitude σ. The simulations presented here were performed under the following conditions: two spectral coordinates (n = 2) and the volume (n' = 1) were used to characterize a peak.

Example 1: Evolution of the GAPT's performance according to crowding rate

The crowding rate δ_c is defined as the square of the average critical radius multiplied by the average number of peaks by peak lists:

$$\delta_c = \overline{\rho(k)}^2 \, \overline{N(k)}^2 \qquad (9)$$

This rate is proportional to the average of the probability law of the number of crossings between paths (Fig. 1f). Thus the larger the crowding rate, the more complex the resolution of the tracking problem. Figure 3 shows the efficiency of GAPT, according to the ratio of correctly found paths, as a function of crowding rate. In order to evaluate the stability of the results as a function of the size of the peak lists, the different confidence intervals are the result of simulation runs for different values of the number of peaks $N(k) = \{10, 20, \ldots, 70\}$ and of the noise level σ chosen in the range [1%,5%].

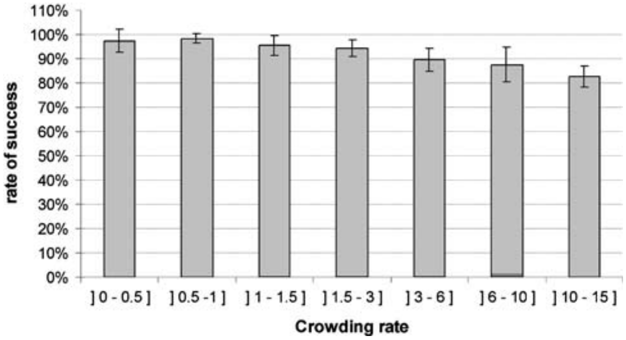


Fig. 3 Evolution of the GAPT performance according to the crowding rate. Five successive experiments were simulated (M = 5). For each experiment, seven different simulations with $N(k) = \{10, 20, \ldots, 70\}$ peaks in the peak lists were used. Noise was added to the peak parameters, with a varying noise level σ in the range [1%, 5%]. The rates of false positive, and false negative peaks δ_+ (k), δ_- (k) were fixed at 5%. The whole simulation was then repeated 15 times altogether. The simulation involved 1260 = 15*7*12 GAPT analyses. Error bars are shown at plus and minus the standard error of the mean of the rate of success

In Fig. 3, the rate of success is the ratio of the number of paths found among the N expected and is calculated in the following manner. The path is regarded as found if all its peaks have been identified; a partially found path is regarded as an error. Only the N first paths (among the total number of possible paths) classified by GAPT are taken into account in the computation of the ratio of the number of paths found.

The quality of the results decreases according to the crowding rate. Nevertheless, low crowding rates are frequently found in NMR experiments ($\delta_c < 3$). In this case the ratio of found paths varies from 90% to 100%, with a high proportion for results between 95% and 100%. The program performance deteriorates in the case of high crowding rates ($\delta_c > 3$), but remains higher than 80%.

In order to illustrate the complexity of the problem, we considered a difficult case. The number of peaks per peak list has been fixed at 110 and five peak lists were used. The crowding rate δ_c is 6.4. All the other parameters are the same as in the previous simulations. The simulation and the result of the GAPT analysis are shown on Fig. 4. The number of paths found by GAPT is 96 out of 110, showing the efficiency of GAPT.

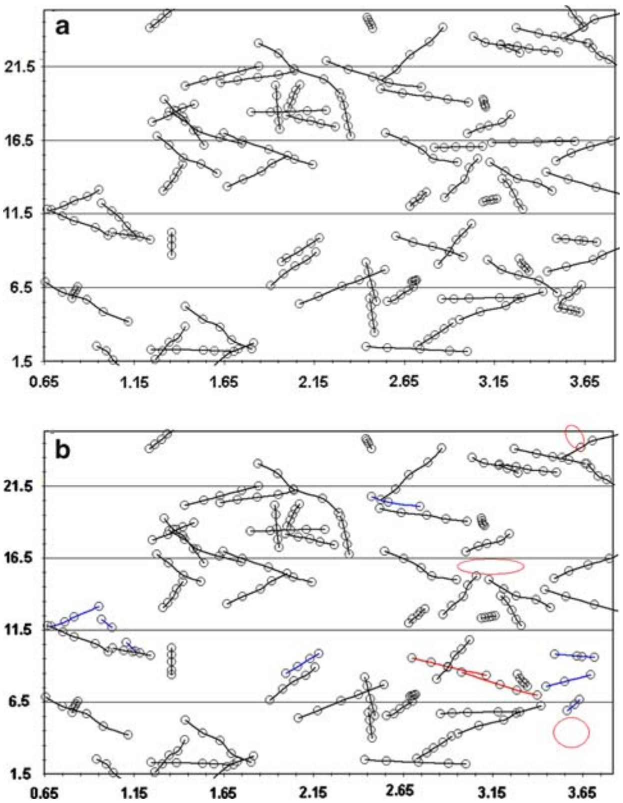


Fig. 4 Example of part of a simulation with a crowding rate δ_c = 6.4. The figure shows the spectral coordinates of the peaks for 5 peak lists with $N(k)$ = 110. Noise was added to the peak parameters, with a varying noise level σ = 2.5%. The rates of false positive, and false negative peaks δ_+ (k), δ_- (k) were fixed at 5%. **(a)** represents the path simulated by GAPT. **(b)** shows the paths computed by GAPT. Red circles are for missing paths. Red paths are for wrong paths and blue paths are for paths partially found

Example 2: Evolution of the GAPT performance according to the error rates δ_+ (k), δ_- (k)

One difficulty remains when computing a peak list from an NMR experiment: peaks detected by the peak-picker are considered as valid when their intensity is above a given threshold. If this threshold is too high, there is a risk of missing peaks (i.e. false negative peaks). On the other hand if this threshold is too low, the risk of false positive peaks is increased. This situation has been simulated, and Fig. 5 shows the rate of success of GAPT for varying error rates of false positive and false negative: δ_+ (k), δ_- (k). In order to be close to a typical NMR case, the crowding rate was in the range of [1.8, 2.2]. The number of peaks per peak list $N(k)$ was fixed to 70. The other parameters used for this simulation are the same as those in the preceding example.

For values of δ_+ (k), δ_- (k) less than 15%, the method is relatively stable. The rate of success is greater than 85%. Practically, GAPT performances are more sensitive to the number of false negatives. Indeed, for rates of false positives and negatives smaller than 15% and 5%, the rate of success is above 95%. This rate of success is in the range of [90%, 95%] for a rate of false negatives in range of [5, 11%]. It ranges between [85%, 90%] for a rate of false negatives ranging from [11%, 17.5%]. Even for a high rate of false positives (>20%), the rate of success is greater than 90% if the rate of false negatives is less than 5%.

Isolated false positives are easily found by GAPT. Nevertheless, if a false positive peak appears close to a true positive peak with a similar volume, the performances of GAPT are degraded. Moreover, the results are not as good for a high rate of negative peaks δ_- (k). Indeed, defining a path in accordance with a chosen score (Eq. 6) requires a minimum of three different (non-virtual) peaks. The probability of obtaining k non-virtual peaks in a path follows a binomial law of parameters:$(M,\ \delta_-\ (k))$. The proportion of meaningful paths that we defined diminishes as δ_- (k) increases. Additionally, a large size of the critical radius can lead to errors in missing some negative peaks.

Finally, the performances of GAPT were studied for different numbers of experiments (M = 3, 5, 7, 9) and for several error rates δ_+ (k), $\delta_-(k) = \{0\%, 15\%\}$. The results are displayed in Fig. 6

The results of GAPT are fairly stable although the complexity of the problem increases according to M. This satisfactory result is linked to the assumption of peak path linearity, which makes it possible to match more efficiently the paths with $M \geq$ 5. It can be observed that the program is less efficient for a large rate of false negatives, but that even in this case, a larger path may compensate partially the errors.

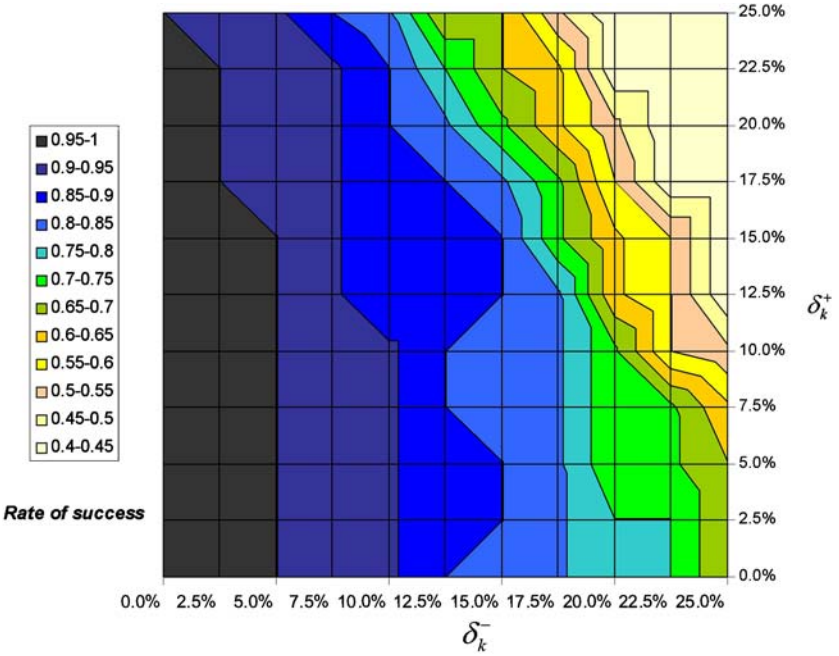
As a conclusion, for practical cases, the GAPT method performs very well for simulated NMR experiments with low peak detection threshold and is less efficient for experiments in which the threshold of peak detection is raised.

Application case of the LTP subjected to temperature variations

An example of a real application is proposed. It consists in the study of the protein LTP, in interaction with lysophospholipid, subjected to a temperature variation.

The temperature was varied from 300 K to 310 K and ^{15}N HSQC spectrum was acquired for each temperature. Figure 7 presents the HSQC of the complex at 300 K, along with the peaks detected by the peak picking program of *Gifa* (Pons et al. 1996).

Fig. 5 Evolution of the GAPT performance according to the error rates δ_+ (k), δ_- (k). Five successive experiments were simulated ($M = 5$). For each experiment, the rates of false positive and false negative peaks δ_+ (k), δ_- (k) were varied between 0 and 25 with 5 steps. The number of peaks per peak list $N(k)$ was fixed to 70. The noise level was fixed to 2.5. The crowding rate was in the range of [1.8, 2.2]. The whole simulation was then repeated 15 times altogether. The simulation involved 540 = 36*15 GAPT analyses. The chart presents the mean over 15 simulations



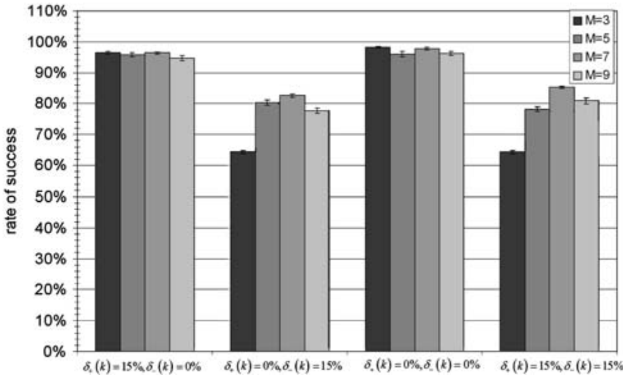


Fig. 6 Evolution of the GAPT performance according to the number of list of peak M and different error rates δ_+ (k), δ_- (k). Four successive experiments were simulated (M = 3,5,7,9). For each experiment, the rates of false positive and false negative peaks δ_+ (k), δ_- (k) were equaled to 0% or 15%. The number of peaks per peak list $N(k)$ was fixed to 70. The noise level was fixed to 2.5%. The crowding rate was in the range of [1.8, 2.2]. The whole simulation was then repeated 30 times altogether. The simulation involved 480 = 4*4*30 GAPT analyses. Error bars are shown at plus and minus the standard error of the mean of the rate of success

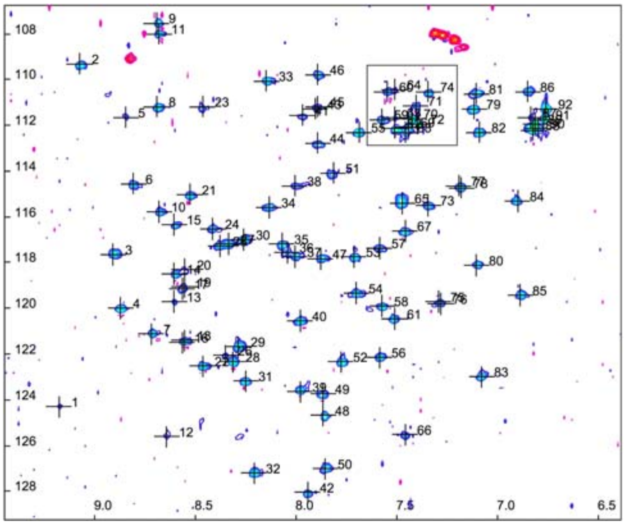


Fig. 7 ^{15}N HSQC NMR spectrum of wheat nsLTP-2 liganded to lysophosphatidyl myristoyl glycerol at 300 K. The crosses present the location of the detected peaks, detected as local maxima, and further fitted to a 2D Gaussian shape. The outlined area is studied in detail in the text and in Fig. 5

Five peak lists were obtained from the experiments for the following temperatures (300, 302.5, 305, 307.5, 310 K) as described in the methods section. Figure 7 displays the spectrum recorded at 300 K, along with the positions of the detected peaks. Only the peak coordinates and the peak volumes were used for the GAPT peak tracking. The peak volume was defined as the product of the intensity by the widths along each dimension.

The number of detected peaks in each experiment is about 90, to be compared with an expected number of signals of 84. For the set of the five experiments the global crowding rate δ_c is evaluated at 1.7.

GAPT performs very satisfactorily in this experiment. The most crowed region, framed in Fig. 7, is presented in Fig. 8. Locally this zone presents a crowding rate value of 2.7. The first and last experiments are displayed; in blue for 300 K and in pink for

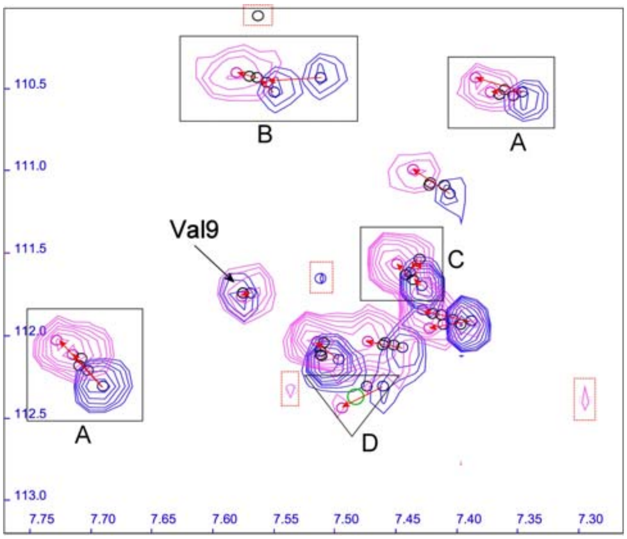


Fig. 8 Zoom of the outlined area in Fig. 5. The red arrows show paths. The circles represent the frequency fitted peaks. Blue circles are for the first experiment obtained at 300 K, pink circles for the last experiment obtained at 310 K and black circles for the other experiments [302.5, 307.5 K]. Green circles represent false negative peaks. i.e. missing peaks added by the algorithm

310 K. The positions of the peaks for the intermediate experiments (302.5; 305; 307.5 K) are represented by black circles. Applied to this zone, GAPT enumerates 78 possible paths. After solving the minimization problem by the exhaustive method, only 15 paths were retained. This particular zone summarizes the set of problems posed by the peak tracking issue. These problems were all identified by GAPT. Zones A contain peaks splitting before the next-to-last experiment and the last experiment. This splitting is certainly due to a gain in resolution afforded by the increase in temperature, and reflects the doublet fine structure of the signal. Zone B corresponds to a situation where the peaks merge between the first experiment and the second experiment. For zone C, the situation is more complex. The peak between the second and third experiment splits then merges again between the third and fourth experiment. Finally the triangular zone D presents an example of a path containing two false negatives for the experiments 305 and 307.5 K. A peak is present in the first two experiments and disappears in the following two, and finally reappears in the last. The expected positions of the missing peaks are represented with the help of a green circle. In studying the set of solutions, a path containing two false negatives was retained by GAPT. The other zones are complete paths of true positive peaks.

Additionally, in these paths certain peaks do not move. For this reason, the number of circles containing a path is not always equal to 5. It should be noted that Valine 9, noted in Fig. 8, presents very small displacements. This is consistent with the fact that it is the only signal arising from the protein backbone in this spectral region, and is probably less sensitive to temperature changes than the other signals, originating from amino acid side chains. Finally, the red dotted rectangles are isolated false positive peaks found by GAPT.

Conclusion

Several authors have previously proposed techniques concerning the tracking of NMR peaks from one multidimensional experiment to another (Pielak et al. 1988; Williamson et al. 1997; Kumar et al. 1998; Muskett et al. 1998; Ross et al. 2000; Accelrys 2002; Peng et al. 2004). While most of these approaches present quite efficient techniques for correlating the peak list from one NMR experiment to another, the method presented here departs from previous work by its ability to handle a complete series of experiments obtained through the variation of all the experimental parameters. In this method, the NMR spectra are first considered by pair, in the so called local approach and distances between peaks are considered within a critical neighborhood determined statistically. Then the whole series is considered, and peak trajectories are determined.

The peak trajectories are chosen so that a peak experiences regular displacements, proportional to the external perturbation, with a trajectory as linear as possible. The linearity of the displacement is a property which comes from (i) the assumption of equilibrium displacement imposed by the external perturbation (ii) the proximity in space of the n spins involved in the n dimensional spectroscopy. With the chosen function score $l(c)$, GAPT is able to select paths with minimum curvature, and the curvature value is evaluated during the selection process. Departure from linearity is thus estimated, and can be used as a biologically meaningful parameter. This property was used for instance by Kitahara et al. (2001, 2002) as a marker of the deformation of the HN bound in the peptide backbone in pressure experiments.

The chosen approach makes it possible to solve difficult cases such as missing peaks, strongly overlapped peaks, or spurious artifacts. This leads to a very robust approach. It was shown by extensive simulations that, if a generous peak-picking is used, a less than 10% error rate is possible in the recovery of the actual peak displacement, even in the case of highly crowded and noisy spectra.

The program presented here can be applied to many typical spectroscopic situations where one wishes to monitor the displacement of a peak: titration experiments, temperature or pressure variation, etc. Titration experiments presenting a strong sigmoïdal behavior may be not analyzed suitably through this approach as the linearity assumption is broken, however a mild sigmoïdal behavior can be analyzed as long as small incremental steps are performed in the titration experiments. The present program, called GAPT, has been implemented in Visual Basic on a Windows-based PC. It takes its input from textual files, and outputs a textual as well as graphical report of the analysis, along with a report of the different steps of the analysis. The program is freely available from authors upon request.

Acknowledgments This work was supported by the French ACI IMPBio, by university Montpellier I and by CNRS. The authors acknowledge C.Royer for her assistance with the English.

References

Accelrys Inc (2002) FELIX user guide, Version 2002, San Diego

Akasaka K (2003) Biochemistry 42:10875–10885

de Lamotte F, Klein C, Issaly N, Gautier MF, Boze H (1999). Biotech Techniques 13:351–354

Goddard TD, Kneller DG (1999) SPARKY 3. University of California San Francisco

Gondran M, Minoux M (1985) Graphes et algorithmes. Eyrolles, Paris

Kitahara R, Yamada H, Akasaka K (2001) Biochemistry 40:13556–13563

Kitahara R, Royer C, Yamada H, Boyer M, Saldana JL, Akasaka K, Roumestand C (2002) J Mol Biol 320:609–628

Koradi R, Billeter M, Engeli M, Guntert P, Wuthrich K (1998) J Magnet Reson 135:288–297

Kumar RA, Bhakta K, Szalma S, Donlan M, Carter B (1998) An integrated high throughput solution for SAR by NMR, 39th experimental nuclear magnetic resonance conference, Asilomar (USA), March 22–27, 1998, Proceedings p 258

Laveen NK, Levitt TS, Lemmer JF (Eds) (1989) Uncertainty in artificial intelligence, vol 3. North-Holland, Amsterdam

Malliavin TE, Barthe P, Delsuc MA (2001) Theor Chem Accts 106:91–97

Muskett F, Frenkiel TA, Feeney J, Freedman RB, Carr MD, Williamson RA (1998) J Biol Chem 273:21736–21743

Peng C, Unger S, Filipp F, Sattler M, Szalma S (2004) J Biomol NMR 29:491–504

Pielak GJ, Atkinson RA, Boyd J, Williams RJ (1988) Eur J Biochem 177:179–85

Pons JL, de Lamotte FD, Gautier MF, Delsuc MA (2003) J Biol Chem 16:14249–14256

Pons JL, Malliavin TE, Delsuc MA (1996) J Biomol NMR 8:445–452

Zuiderweg ERP (2002) Biochemistry 41:1–7

Ross A, Schlotterbeck G, Klaus W, Senn H (2000) J Biomol NMR 16:139–146

Williamson RA, Carr MD, Frenkiel TA, Feeney J, Freedman RB (1997) Biochemistry 36:13882–13889