Journal of Magnetic Resonance 192 (2008) 294–301

Contents lists available at ScienceDirect

Journal of Magnetic Resonance

journal homepage: www.elsevier.com/locate/jmr

Fractional volume integration in two-dimensional NMR spectra: CAKE, a Monte Carlo approach

Rocco Romano [a,*], Debora Paris [b], Fausto Acernese [a], Fabrizio Barone [a], Andrea Motta [b,*]

[a] Dipartimento di Scienze Farmaceutiche, Università degli Studi di Salerno, Via Ponte Don Melillo, I-84084 Fisciano, Salerno, Italy
[b] Istituto di Chimica Biomolecolare del CNR, Comprensorio Olivetti, Edificio A, Via Campi Flegrei 34, I-80078 Pozzuoli, Naples, Italy

ARTICLE INFO

Article history:
Received 26 October 2007
Revised 8 February 2008
Available online 21 March 2008

Keywords:
Volume integration
In-phase peaks
Peak symmetry
Monte Carlo
NMR
Structure calculations

ABSTRACT

Quantitative information from multi-dimensional NMR experiments can be obtained by peak volume integration. The standard procedure (selection of a region around the chosen peak and addition of all values) is often biased by poor peak definition because of peak overlap. Here we describe a simple method, called CAKE, for volume integration of (partially) overlapping peaks. Assuming the axial symmetry of two-dimensional NMR peaks, as it occurs in NOESY and TOCSY when Lorentz–Gauss transformation of the signals is carried out, CAKE estimates the peak volume by multiplying a volume fraction by a factor *R*. It represents a proportionality ratio between the total and the fractional volume, which is identified as a slice in an exposed region of the overlapping peaks. The volume fraction is obtained via Monte Carlo Hit-or-Miss technique, which proved to be the most efficient because of the small region and the limited number of points within the selected area. Tests on simulated and experimental peaks, with different degrees of overlap and signal-to-noise ratios, show that CAKE results in improved volume estimates. A main advantage of CAKE is that the volume fraction can be flexibly chosen so as to minimize the effect of overlap, frequently observed in two-dimensional spectra.

1. Introduction

NMR spectra can provide quantitative analysis of a sample, and a standard one-dimensional (1D) ^{1}H NMR spectrum is often used to obtain a reliable evaluation of peaks. However, as the complexity of the sample increases, resonance overlap becomes a serious problem that easily degrades the accuracy of the analysis, and two-dimensional (2D) NMR data are required to gain sufficient discrimination of resonances. Quantification of NMR spectra is also fundamental in the new emerging field of metabolomics/metabonomics [1,2] and in the structure and dynamics of proteins in solution [3]. This widespread requirement of deriving quantitative information from NMR data has prompted the need to find methods for accurate and precise integration procedures both for 1D and 2D spectra. This paper describes a new simple method for peak volume integration in 2D spectra, which appears to be particularly suited for overlapping peaks. Quantitative information in NMR spectra is brought by peak areas [4]. Two methods of peak integration are used: direct summation of spectral data points and peak parameter search by curve fitting. In the absence of a model for the peak shape, direct summation appears to be the only practical technique. It is not, however, adaptable to (partially) overlapping peaks, and introduces two kinds of systematic errors. One is due to the approximation caused by the assimilation of the integral of a continuous function with a finite sum [5]; the second one is caused by the parts of the peaks that are left outside of the integration range [6].

Ideally, an efficient integration method should be applicable even when in presence of peak overlap or artifacts. Many of the available NMR processing and analysis packages achieve volume integration by direct summation of all data points within a polygonal bounding the peak. This procedure requires a reliable definition of the peak area: the circling should be as large as possible to enable for a complete integration, but also small enough to minimize inclusion of artifacts (baseplane rolls, t_1 noise, tails of other peaks). As such, the idealized procedure appears to be restricted to well-resolved peaks. In automated protocols, a possible way to define the area integration makes use of the observation that the slope of a peak height decreases monotonically with the distance to the peak center, at which point it approximates zero [7,8]. A similar approach defines the peak integration area using an iterative region-growing algorithm [9–11], which recognizes all data points that are part of a given peak, and the integration is performed on a user-defined threshold level. This procedure works quite satisfactorily even for overlapping peaks, as long as the peak maxima are visibly resolved and therefore recognizable by the peak-picking procedure. In a different approach, the peaks are fitted by a set of reference peaks defined by the user [12–14]. In order

* Corresponding authors. Fax: +39 081 8041770.
E-mail addresses: rromano@unisa.it (R. Romano), andrea.motta@icb.cnr.it (A. Motta).

1090-7807/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2008.03.010

to obtain accurate line shapes and integrals in one dimension, it is necessary to apply a non-linear curve-fitting procedure [4,15]. Although this protocol is probably best suited in cases where peaks strongly overlap, it hinges on the careful definition of suitable reference peaks and selection of initial fitting parameters by the user.

A general approach for peak integration would be to exploit the peak symmetry as a criterion to evaluate the peak volume. Symmetry considerations have previously been used for pattern recognition in 2D NMR spectroscopy [16], and only rarely for the analysis of in-phase peaks as in NOESY and TOCSY experiments. The program AUTOPSY used symmetry for automated peak-picking in multi-dimensional NMR spectra of proteins [17]. Here we propose CAKE, a novel integration method based on peak symmetry. After a 2D Lorentz–Gauss time-domain filtering, the spectral lines are converted into Gaussian lines, therefore presenting a cylindrical or elliptical symmetry. By assuming the vertical axial symmetry of individual peaks (a peak with a unique center corresponds to its maximum), the volume is obtained by multiplying a selected volume fraction by a factor *R*, which represents a proportionality ratio between the total and the fractional volume, optimized by Monte Carlo techniques. This *minimalistic* approach warrants that the fractional volume can be chosen so as to minimize the effect of overlap in complex NMR spectra. When applied to simulated and experimental 2D in-phase peaks with different degrees of overlap, CAKE (Monte CArlo peaK volume Estimation) obtains an unbiased volume estimation. It is shown that, compared with the direct summation procedure, the fractional volume approach yields rather good estimates of the peak volumes, even for significant overlap, as long as a single contour level and its center arising from a single peak can be detected.

2. Materials and methods

2.1. NMR data collection

The mixture of tripeptides Ala-Phe-Ala (AFA) and pyroGlu-His-Pro (thyrotropin-releasing hormone, TRH), was prepared by dissolving appropriate amounts in 0.5 ml of $^1H_2O/^2H_2O$ (95/5 v/v) to yield for each peptide a concentration of 0.10 mM. Salmon calcitonin (sCT) was dissolved in $^1H_2O/^2H_2O$ (95/5 v/v) to obtain a concentration of 1.5×10^{-3} M. Perdeuterated sodium dodecyl sulfate (SDS, Cambridge Isotope Laboratories, Woburn, MA) was added as a solid, maintaining its concentration well above the critical micelle concentration, with a peptide–SDS molar ratio of about 1:100. ^{1}H NMR spectra, recorded at 295 K and pH 7.4, were acquired on a Bruker DRX-600 spectrometer operating at 600 MHz, equipped with a TCI cryoprobe™ fitted with a gradient along the *Z*-axis. Spectra were referenced to sodium 3-(trimethylsilyl)-[2,2,3,3-$^2H_4$]propionate. Homonuclear 2D clean TOCSY spectra [18] were recorded by standard techniques and incorporating the excitation sculpting sequence [19] for water suppression. Five-hundred and twelve equally spaced evolution time-period t_1 values were acquired, averaging four transients of 2048 points, with 6024 Hz of spectral width. Time-domain data matrices were all zero-filled to 4096 in both dimensions, yielding a digital resolution of 2.94 Hz/pt. Prior to Fourier transformation, time-domain filtering was applied with a Lorentz–Gauss window to both t_1 and t_2 dimensions. The TOCSY experiment was recorded with a spin-lock period of 64 ms, achieved with the MLEV-17 pulse sequence [20].

2.2. Software

NMR data processing and baseline correction were obtained with the program XWINNMR (Bruker, Biospin GmbH, Ettlingen, 2003). Standard peak integration was carried out with the programs XWINNMR and MestRe-C [21], in which integrated volumes are computed as the sum of all digital intensities within a rectangular box and a tunable ellipse bounding a peak, respectively. CAKE software was written in MATLAB language and was implemented in the graphical environment of MATLAB 7.1.

2.3. The fractional peak method

2.3.1. Line shapes in two-dimensional NMR

In high-resolution NMR the frequency-domain line shapes are closely approximated by a Lorentzian function. Neglecting coherence transfer echoes, the signal envelope of a 2D experiment can be assumed to have a biexponential form [16]

$$s^{(e)}(t_1,t_2) = s^{(e)}(0,0)\exp(-\lambda^{(e)}t_1)\exp(-\lambda^{(d)}t_2) \tag{1}$$

[with rates $\lambda = 1/T_2$ in the evolution (e) and detection (d) periods]. Such time-domain envelope, decaying exponentially in both dimensions, lacks cylindrical symmetry about the origin $t_1 = t_2 = 0$. After a 2D Fourier transformation, the corresponding 2D absorption peak shows a Lorentzian shape, whose sections, taken parallel to either axis yield pure 1D absorption Lorentzian line shapes. The asymptotic decay is proportional to $(\Delta\omega_1^{(e)})^{-2}$ and $(\Delta\omega_2^{(d)})^{-2}$ on sections parallel to one of the frequency axes, while it is proportional to the inverse fourth power in the bisecting planes [with $(\Delta\omega_1^{(e)})$ and $(\Delta\omega_2^{(d)})$, frequency offset in evolution (e) and detection (d) periods with respect to resonances $\omega_1^{(e)}$ and $\omega_2^{(d)}$]. This lack of cylindrical or elliptical symmetry has been called "star effect", and can be removed by a 2D Lorentz–Gauss transformation [16], which yields a 2D absorption mode peak shape with cylindrical or elliptical symmetry.

By using a Lorentz-to-Gauss weighting function

$$h(t_1,t_2) = \exp(+\lambda_1 t_1)\exp(+\lambda_2 t_2)\exp(-\sigma_1^2 t_1^2/2)\exp(-\sigma_2^2 t_2^2/2) \tag{2}$$

[with $\lambda_1 = \lambda^{(e)}$, $\lambda_2 = \lambda^{(d)}$ (enhancement resolution parameters) and $\sigma_1 = \sqrt{\frac{\lambda_1}{t_{1\max}}}$ and $\sigma_2 = \sqrt{\frac{\lambda_2}{t_{2\max}}}$ (Gaussian parameters), being $(t_{1\max}, t_{2\max})$ the point at which the weighting function reaches his maximum value]. The envelope of Eq. (1) becomes

$$s^{e}(t_1,t_2) = s^{e}(0,0)\exp(-\sigma_1^2(t_1^2/2))\exp(-\sigma_2^2(t_2^2/2)). \tag{3}$$

After a 2D transformation, a Gaussian line shape is obtained

$$S(\omega_1,\omega_2) = s^{(e)}(0,0)\left(\frac{2\pi}{\sigma_1\sigma_2}\right)\exp\left(-\frac{\Delta\omega_1^2}{2\sigma_1^2}\right)\exp\left(-\frac{\Delta\omega_2^2}{2\sigma_2^2}\right). \tag{4}$$

The contours are circular for $\sigma_1 = \sigma_2$ and elliptical for unequal widths. It is important to underline that 2D Lorentz–Gauss transformation is useful only if the dispersive components in peaks with mixed phase are suppressed, and this can be achieved with pure phase spectra (i.e. either pure 2D absorption or pure 2D dispersion peaks) [16]. It must also be emphasized that the elliptical symmetry of Gaussian signals is obtained only in phase-sensitive displays, and if the absolute amplitude of a Gaussian signal is calculated, a peak shape is obtained which features again a star effect.

In most practical applications, the complete analytical expression for a discrete Fourier transform NMR spectrum is a sum of complex, non-Lorentzian functions [4,22]. However, a true Lorentzian spectrum is obtained if the acquisition time t_2 is large, compared to the relaxation time of the slowest decaying resonance j ($t_2 \geqslant 1/R_{2j}$), and the sweep width is large compared to the relaxation rate R_{2j}, as well as the frequency range of the spectrum (that is, the difference between the slowest and the fastest decaying nuclei) [23]. Nevertheless, this discrete Fourier transform spectrum requires correction of a pseudobaseline stemming from the first point of the FID and of a frequency-dependent phase distortion of the spectrum (for details see Refs. [4,22]).

Accordingly, a phased, baseplane corrected unsaturated resonance line in solution is closely approximated by a Lorentzian function. Convolution of the time-domain with exponential, sine, cosine functions, does not alter the line shape after transformation [24], and preserves the frequency of its maximum. This shape has been useful in peak fitting procedures applied to experimental data [23]. As stated above, a 2D Lorentzian line lacks cylindrical or elliptical symmetry, which can be achieved by a 2D Lorentz–Gauss transformation. Gaussian filtering transforms a Lorentzian frequency-domain function of width ω_0 into a Gaussian frequency-domain function of width $\rho\omega_0$, where ρ is typically less than unity, and it has been found that $\rho = 0.66$ is usually close to optimum [25].

Bearing in mind the power of Lorentz–Gauss transformation and the symmetry of the Gaussian line, the CAKE algorithm aims at integrating a peak relying upon its axial symmetry, even when in drastic overlapping conditions. The idea is that the volume can be estimated by integrating a non-overlapping fraction of the peak obtaining a reasonable approximation of volume in cases where cross-peaks overlap. Therefore the major assumption in this study is that the Lorentzian signal is transformed into a Gaussian line by a Lorentz-to-Gauss transformation, which for in-phase peaks of TOCSY and NOESY spectra is well-suited to maximize signal-to-noise ratio [16].

Fig. 1a shows the contour plot of a Gaussian peak. The arbitrary angle $A\widehat{O}B$ (a "slice" selected in a non-overlapping region and centered on the center of mass), defines the area A_{F_i} of a peak fraction for each i-th level bound curve; such an angle identifies a fractional volume V_F in the three-dimensional representation. Because of the axial symmetry, for any i-th level the fractional volume V_F relates to the total volume V_T as the fractional area A_{F_i} of the i-th level relates to the corresponding total area A_{T_i} of the same level.

From the equation

$$V_T = \frac{A_{T_i}}{A_{F_i}} \cdot V_F, \tag{5}$$

true for each couple of level bound areas, if $R_i = \frac{A_{T_i}}{A_{F_i}}$, the total volume of a peak can be obtained by multiplying a fractional volume by the corresponding R_i factor.

Eq. (5) is completely independent from the sequence of levels arbitrarily chosen to represent the cross-peaks. The i-th level should be selected so as to belong to a single peak: this is always possible for isolated peaks while for overlapping ones this is true for the higher levels.

It is common experience that experimental 2D peak shapes are quite close to an ellipse. Therefore, Eq. (5) is still valid if the right angle $A\widehat{O}B$ delimits 1/4 of the ellipse by lying on the semimajor and the semiminor axes (Fig. 1c). In particular, by defining the ellipse eccentricity as $e = \sqrt{1 - \frac{b^2}{a^2}}$, where b and a are the semiminor and the semimajor axes (assuming $b < a$), $0 \leqslant e \leqslant 1$ and $e = 0$ in the case of a circle. More generally, it can be demonstrated that Eq. (5) applies with a good approximation to eccentricity $e \leqslant 0.5$, which corresponds to a difference $< 10\%$ between axes, and a circle well approximates the ellipse. For eccentricity $e > 0.5$, Eq. (5) can be safely used if one of the semiaxes is the bisector of the polygonal $A\widehat{O}B$. The advantage of this approach becomes apparent for overlapping Gaussian peaks. Here, the integration is biased by the presence of the overlapping region that affects both volumes. In contrast, the "slice" $A\widehat{O}B$ of peak 1 (Fig. 1b), selected in a non-overlapping region, has very little contribution, if any, from peak 2, and therefore its fractional volume can mostly be attributed to peak 1. The same is true for $C\widehat{K}D$ slicing peak 2 (Fig. 1b), whose fractional volume can mostly be attributed to peak 2. Therefore, if we integrate the volume fraction identified by $A\widehat{O}B$ and calculate the corresponding R_1 constant, it should be possible to estimate the unbiased volume of each peak. From Fig. 1b, the second most internal (highest) level of peak 1, essentially arises from peak 1, and the effect of peak 2 on that level is negligible. Consequently, the R_1 constant can be obtained from the ratio between the total area (A_{T_1}) and the fractional area (A_{F_1}) of that level, A_{T_1}/A_{F_1}. Analogously, for peak 2 the fractional volume identified by $C\widehat{K}D$ can be considered, and its second highest level can be chosen to obtain the respective factor R_2 (Fig. 1b).

2.3.2. The R factor estimation

In order to estimate the R factor for a selected fraction of a peak, an internal level attributable to the peak has to be chosen. Denoted by A_T the total level area and by A_F the fractional level area, the ratio $R = A_T/A_F$ can be obtained by a Hit-or-Miss Monte Carlo technique [26,27]. It generates uniformly distributed random points in a known bound area (or volume) containing an unknown area (or volume) of interest, and counts the number of *hits* (or points) contained in the unknown area, with respect to the total number of points in the known bound area. The Hit-or-Missing reasoning is that, if points are random and uniformly distributed, the ratio between the number of points in the unknown area with respect to the total number points in the known bound area will correspond to the ratio between the unknown area and the known bound area. This allows an estimate of the unknown area (or volume).

Let us denote by (lx_i, ly_i), with $i = 1, 2, \ldots, N$, the vertex coordinates of the polygonal P_{level} relative to a contour level, by (c_x, c_y) the

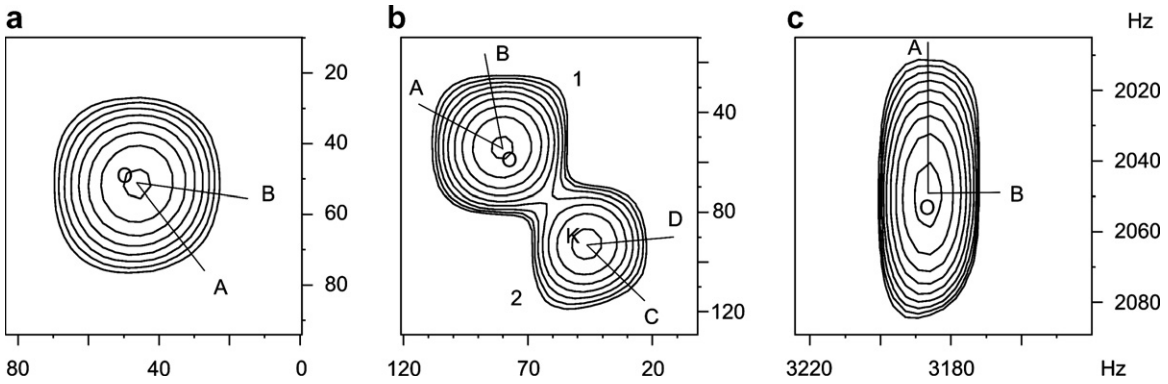

Fig. 1. Contour plots of simulated isolated (a) and overlapping (b) Gaussian peaks. In (a), the arbitrary angle $A\widehat{O}B$ defines a fraction of the peak area, selected in a non-overlapping region, and centered on the center of mass. In (b), $A\widehat{O}B$ and $C\widehat{K}D$ select a fraction of peaks 1 and 2, respectively. (c) Experimental Gaussian cross-peak. The right angle $A\widehat{O}B$ selects a fractional area corresponding to 1/4 of the total area.

coordinates of its center point, and by α_1, α_2 two rays with their common origins in (c_x, c_y). The fractional area A_F is therefore defined by the intersection of the polygon P_{level} and the area delimited by the rays. Furthermore, let us denote by lx_{min} and lx_{max} the minimum and maximum lx_i coordinates, and by ly_{min} and ly_{max} the minimum and maximum ly_i coordinates, respectively. Two pseudo random numbers x_r and y_r are now uniformly extracted in the intervals $[lx_{min}, lx_{max}]$, and $[ly_{min}, ly_{max}]$, respectively. The extraction is continued until a number N_{A_T} of points (x_r, y_r) is internal to the polygonal P_{level}. If an extracted point (x_r, y_r) is also inside the area A_F, then the number of fractional hits N_{A_F} is augmented by one. Of course, being the (x_r, y_r) pairs uniformly extracted in the rectangle $[lx_{min}, lx_{max}] \times [ly_{min}, ly_{max}]$, the ratio $R = A_T/A_F$ will be estimated by the ratio $R = N_{A_T}/N_{A_F}$.

2.3.3. The Monte Carlo integration

In principle, any method is suitable to integrate the selected fractional volume. However, the simple sum can be biased because of the small region and the limited number of points within the selected area. Accordingly, the Monte Carlo Hit-or-Miss technique appears to be more suitable. As a known volume we assume a quadrilateral base prism (of height h equal to the maximum of the peak) that contains the factional volume V_F to be integrated. Points (x_r, y_r) inside the prism base are randomly generated, and in correspondence to each base point, a pseudo random number P_h is uniformly extracted in the interval $[0, h]$. Such extracted point P_h is compared with the interpolated peak value $p(x_r, y_r)$, corresponding to the extracted (x_r, y_r) base point. If P_h is less then $p(x_r, y_r)$, the extracted point is considered internal to the fractional volume V_F, and counted as a *hit*. At the end, the fractional peak volume is given by the ratio between the number of *hits* and the total number of P_h extracted points multiplied for the known prism volume.

Let us denote by (px_i, py_i), with $i = 1, 2, 3, 4$, the vertex coordinates of the quadrilateral P_{base}, which is the base of a prism of height h and that contains the fractional volume V_F (in particular, $px_1 = c_x$, and $py_1 = c_y$, while other two points are chosen on the α_1 and α_2 rays). Furthermore, let px_{min} and px_{max} be the minimum and maximum px_i coordinates, and py_{min} and py_{max} the coordinates corresponding to the minimum and maximum py_i, respectively. Two pseudo random numbers x_r and y_r are uniformly extracted in the intervals $[px_{min}, px_{max}]$ and $[py_{min}, py_{max}]$, respectively. The extraction is continued until a number $N_{P_{base}}$ of points (x_r, y_r), internal to the quadrilateral of base P_{base}, is obtained. For any point internal to the quadrilateral of base P_{base}, a cubic interpolation gives the peak $p(x, y)$ values in the point (x_r, y_r), and another pseudo random number ρ is uniformly extracted in the interval $[0, 1]$. If $\rho \cdot h \leqslant p(x_r, y_r)$, that is, if $\rho \cdot h$ is a point internal to the fraction volume V_F, the number of volume hits N_V is augmented by one. If V_P is the prism volume, calculated by the software, then the fractional volume V_F is $V_F = N_V/N_{P_{base}} \cdot V_P$.

3. Results and discussion

3.1. Simulations

3.1.1. Bias vs. overlapping

In order to test the algorithm, we applied CAKE to simulated overlapping peaks of known volume. In particular, two Gaussian peaks had center (x_i, y_i), equation $A_i \exp\left[-\frac{(x-x_i)^2+(y-y_i)^2}{2\sigma_i^2}\right]$, volume $V_i = 2\pi\sigma_i^2 A_i$ and a half-height width $\zeta_i = \sqrt{2\sigma_i^2 \ln 2}$, $i = 1, 2$, and addition of Gaussian noise. Denoting $d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ the distance between the peak centers, it is possible to define the parameter $\eta \equiv \frac{\zeta_1+\zeta_2}{d}$ as an index of the overlap, such that a large value corresponds to strong overlap.

Setting the amplitude $A_1 = 50.0$ and the dispersion $2\sigma_1^2 = 2.0$ to obtain $V_1 = 100\pi$, the A_2 and $2\sigma_2^2$ values were changed so as to keep the volume V_2 constant ($V_2 = 100\pi$), with the overlap index being $0.8 \leqslant \eta \leqslant 1.5$. The contour plots of the simulated peaks are reported in Fig. 2 for $\eta = 0.8$ (peak 1), and $\eta = 1.5$ (peak 2). CAKE integration was compared with the standard one, obtained by summing the amplitudes of all data points within a polygonal bounding the peak. In order to establish the best number of extractions N_P in the Hit-or-Miss determination of R, and the best number of extractions $N_{P_{base}}$ in the Hit-or-Miss determination of the fractional volume, simulations were conducted in the extreme limit of $\eta = 1.5$ (Fig. 2, peak 2). Fig. 3 reports the percentage of bias vs. the number of extractions N_P, for different $N_{P_{base}}$ values ranging from 100 to 1000 (right column in Fig. 3). As it can be seen, results become unbiased for $N_P \geqslant 1500$, while, except for $N_{P_{base}} = 100$ (square symbol), the dependence on $N_{P_{base}}$ is negligible. Accordingly, the values $N_P = 2000$, and $N_{P_{base}} = 500$ appear to be a good compromise between computing time and accuracy. The results of the simulations are reported as percentage of bias vs. the degree of overlap for a signal-to-noise ratio (SNR) of 34.9 ± 3.0 (Fig. 4a) and 56.1 ± 4.7 (Fig. 4b). The standard integration (filled squares) was carried out by bounding the peak with an ellipse, while for the CAKE integration (filled circles) we used $N_P = 2000$, and $N_{P_{base}} = 500$. In both cases, each integration was repeated 10 times. In Fig. 4a ($SNR = 34.9 \pm 3.0$), the standard method gives unbiased integration values only for low overlap index $\eta \leqslant 0.9$. (Fig. 2, peak 1), to become totally biased for $\eta \geqslant 1.0$. In contrast, CAKE always performs better, especially in the range $1.0 \leqslant \eta \leqslant 1.3$, which represents different degree of overlap commonly found in 2D spectra. Overall, the fractional method appears to be unbiased in the whole $0.8 \leqslant \eta \leqslant 1.5$ range, that is for strongly overlapping peaks and in the presence of a low signal-to-noise ratio ($SNR = 34.9 \pm 3.0$). Fig. 4b reports the same simulations with a $SNR = 56.1 \pm 4.7$. The standard method performs well for $\eta \leqslant 0.9$, with a general trend very similar to that observed for lower SNR (Fig. 4a). In contrast, the fractional method shows a general reduction of the bias percentage, with values generally lower than those obtained in the previous simulation. Taken together our results suggest that, regardless of the SNR, the CAKE method performs always better than the standard one.

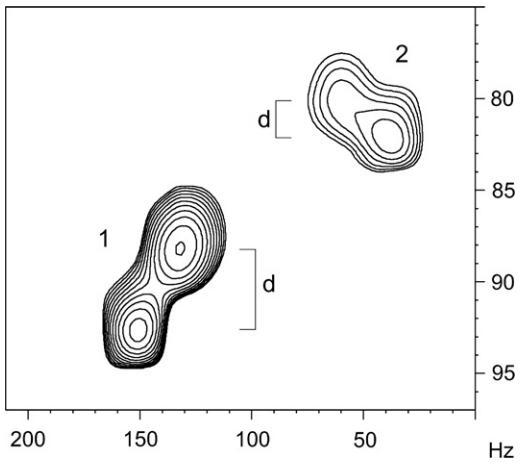

Fig. 2. Contour plot of two Gaussian peaks with different degree of overlap (η): peak 1, $\eta = 0.8$ and peak 2, $\eta = 1.5$. For the definition of η see text. d is the distance between peak centers.

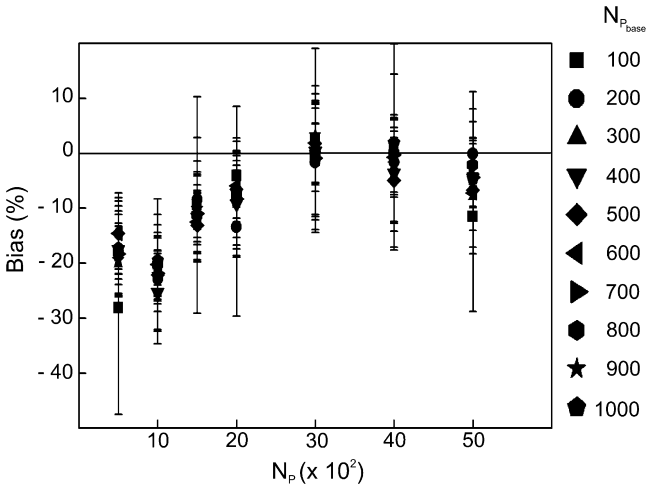

Fig. 3. Percentage (%) of bias as a function of the number of extractions (N_P) to estimate the R factor. For each N_P we tested several $N_{P_{base}}$ values to estimate the volume fraction, and they are indicated with corresponding symbols on the right.

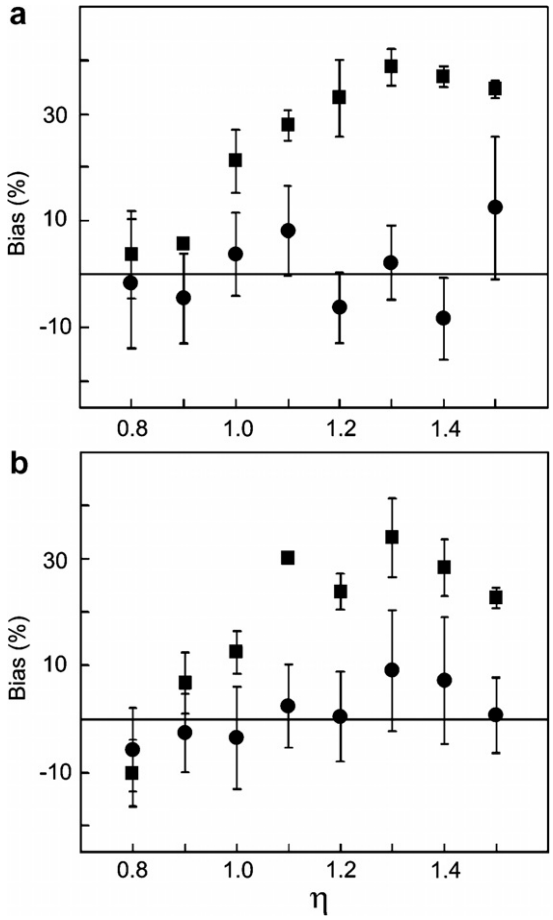

Fig. 4. Simulation results expressed as percentage of bias in volume estimation vs. the degree of overlap (η). Integration was achieved with the standard (■) and the CAKE (●) methods at different signal-to-noise ratios. (a) SNR = 34.9 ± 3.0; (b) SNR = 56.1 ± 4.7.

3.1.2. Bias vs. eccentricity

Since experimental 2D-peak shapes are close to elliptic, we tested CAKE on a simulated ellipse of known volume. In particular, we considered peaks of equation $S_i(\omega_1,\omega_2) = A_i\left(\frac{2\pi}{\sigma_{1i}\sigma_{2i}}\right)\exp\left(-\frac{\Delta\omega_1^2}{2\sigma_{1i}^2}\right)\exp\left(-\frac{\Delta\omega_2^2}{2\sigma_{2i}^2}\right)$, volume $V_i = A_i$ and contour of eccentricity $e_i = \sqrt{1-\frac{\min(\sigma 1i,\sigma 2i)}{\max(\sigma 1i,\sigma 2i)}}$, with addition of Gaussian noise. Integration was carried out in two ways. The fractional area was firstly selected randomly (i.e. avoiding any symmetry), and, secondly, symmetrically with respect to any of the semiaxes of the elliptic peak. The random choice (Fig. 5a) produced a scattered bias distribution between 0 and 20% for $0.8 \leqslant e \leqslant 0.74$, with a maximum of 25% for $e = 0.78$. For $0.8 \leqslant e \leqslant 0.9$, which corresponds to a ratio between semiaxes in the range of $0.45 \leqslant b/a \leqslant 0.60$, the average bias is 5%. This result appears to be relevant as the b/a value corresponds to the experimental elliptic shapes usually found in 2D spectra.

The symmetry selection of the fractional area (Fig. 5b) shows a bias $\leqslant 10\%$ for all eccentricity values, with the maximum at $e = 0.78$ reduced to 12%. For $0.8 \leqslant e \leqslant 0.9$ the average bias is very similar to that found for the random selection (Fig. 5a).

In conclusion, it is suggested that, for elliptical peaks, slicing should be done symmetrically with respect to one of the semiaxes, even though for $0.8 \leqslant e \leqslant 0.9$, that is for most of the experimental 2D peaks, the bias is essentially independent from the selection.

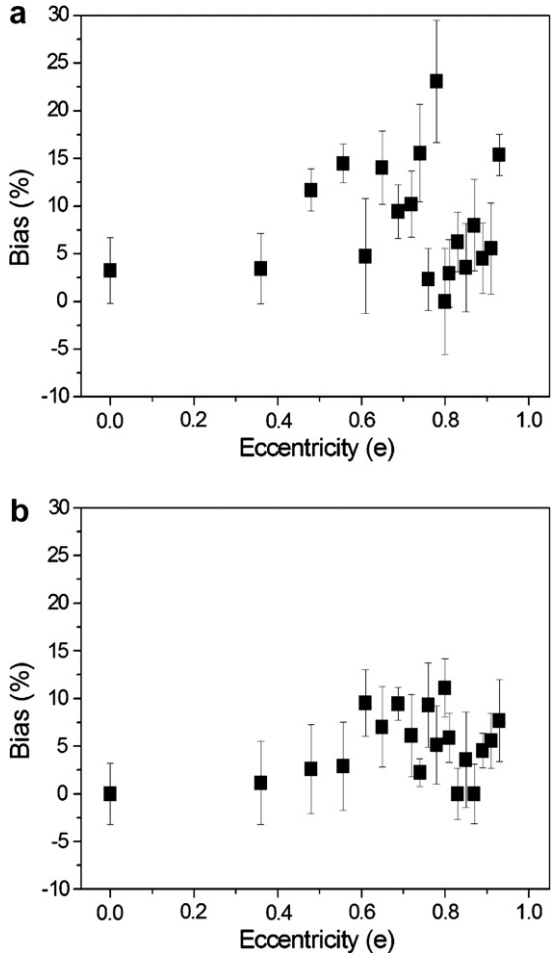

Fig. 5. CAKE integration of simulated elliptic peaks expressed as percentage of bias in volume estimation vs. contour eccentricity (e). In (a) the fractional area was chosen in a non-symmetric way with respect to the semimajor and the semiminor axes of the elliptic peak. In (b) the fractional area was chosen in a symmetric way with respect to the semimajor and semiminor axes of the elliptic peak. In both cases the SNR = 69.5 ± 3.2.

3.2. Experimental results

CAKE was initially tested on a TOCSY spectrum of a mixture of two tripeptides, AFA and TRH. In order to have an internal reference we selected pairs of peaks, each of them stemming from a single spin system, such that they have similar intensity within each pair but one peak overlaps with others. In particular we chose pairs that exemplify the correlations between the γCH_2 (labeled 1 in Fig. 6b), and between α and β protons of AFA *Phe*^{2} (labeled 2 in Fig. 6a), and TRH *His*^{2} (labeled 3 in Fig. 6a). The magnitude of a given TOCSY peak (governed by mixing coefficients $a_{lk}(\tau_m)$ for transfer of magnetization through the spin system from spin I_l to spin I_k) depends on the topology of the spin system, the coupling constants between pairs of spins, the efficiency of the isotropic mixing sequence employed, and the relaxation rate during the mixing pulse. Although the robustness of the integration method does not depend upon the experiment type or the intensity of the chosen peak, we looked for pairs in which the peaks are expected to have similar intensity but one of them overlaps with others. Accordingly, we selected the AMX spin system of the two aromatic residues (Fig. 6a) in *AFA* and *TRH*. From relaxation measurements (not shown) at two different spectrometer frequencies, we estimated for both peptides similar correlation times and relaxation rates; furthermore, the measured $^3J_{\alpha\beta}$ and $^3J_{\alpha\beta'}$ values in each spin system were identical, therefore excluding differences in the peak intensity due to different coupling constants; finally, the single $^2J_{\gamma\gamma'}$ value for the γCH_2 protons of the *TRHpyroGlu* warrants a similar intensity for the two peaks within each pair.

The selected peaks were integrated with standard and CAKE methods, and the results are reported in Fig. 6c as the Difference percentage of volume for each cross-peak pair. For the CAKE integration we selected the most internal level belonging to a single peak, which had elliptical symmetry with eccentricity $e > 0.75$. The values obtained with CAKE for the three peak pairs are all within 10%, giving an unbiased estimation of the difference percentage of the volumes in each pair. In contrast, the standard method estimates for each peak pair values $>35\%$ for pairs 1 and 2, and $\approx 25\%$ for pair 3. Surprisingly, the CAKE approach gives for the pair 1, which lies on the TOCSY diagonal, about zero volume difference, supporting robustness for the method, also in the presence of elliptical symmetry.

Extension of CAKE to larger polypeptides was tested on sCT, a hormone of 32 amino acids. Fig. 7a and b report the TOCSY expansions of sCT in aqueous SDS, with the AMX spin systems of *Asn*^{3} (peaks 1 and 1′), *Cys*^{7} (peaks 2 and 2′) and *Tyr*^{22} (peaks 3 and 3′) (Fig. 7a), and the β protons of *Leu*^{4} (peaks 4 and 4′) and *Arg*^{24} (peaks 5 and 5′) (Fig. 7b). According to the above considerations, the selected peaks were integrated with the standard and the CAKE methods and the results are reported in Fig. 7c. As above, the CAKE values are all unbiased, giving an estimation of the Difference percentage $<10\%$. On the contrary, the standard method gives for peak pairs 1–4 values of bias between 30% and 40%, while pair 5

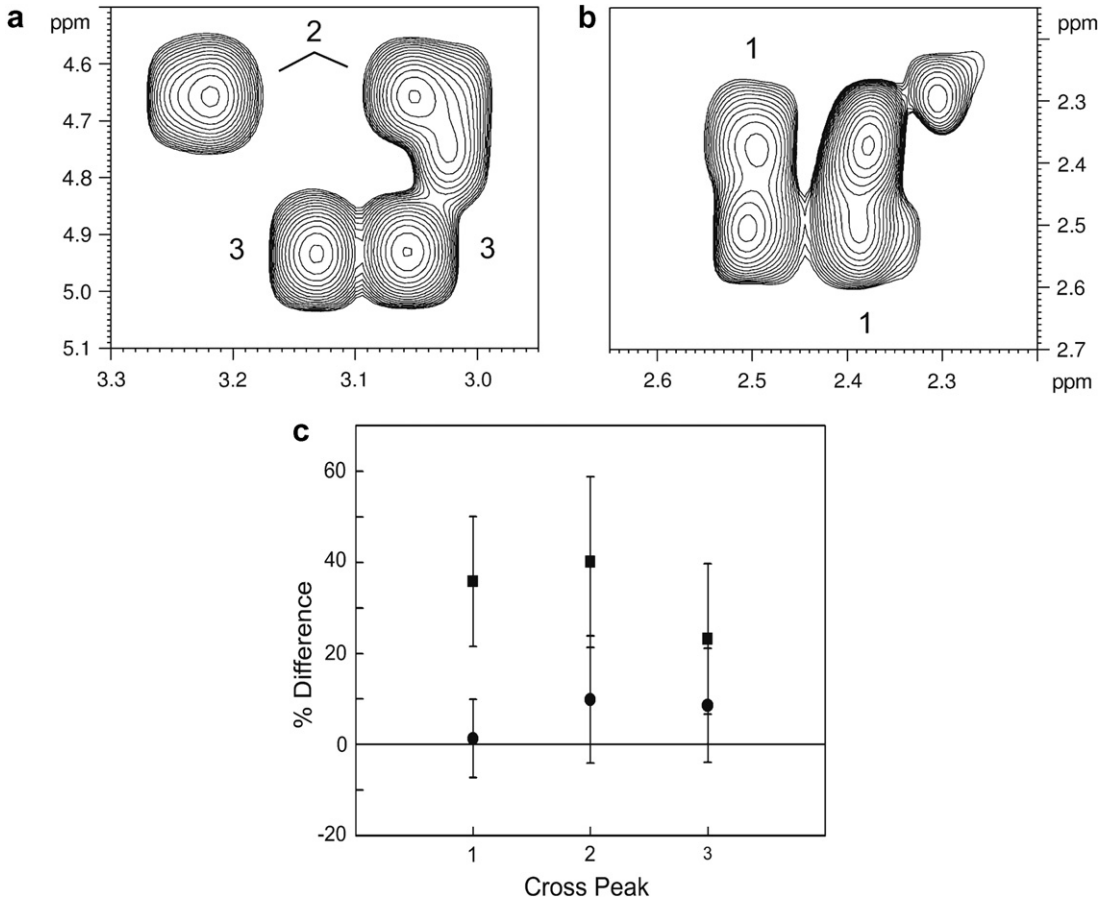

Fig. 6. TOCSY spectrum of the mixture of AFA and THR tripeptides, acquired at 300 K with 64 *ms* mixing time. Expansions (a) and (b) report peaks originating from γCH_2 protons of the TRH pyroGlu [labeled 1 in (b)], and α and β protons of AFA *Phe*^{2} [labeled 2 in (a)], and TRH *His*^{2} [labeled 3 in (a)]. (c) Difference percentage (%) of volume in the selected cross-peak pairs 1, 2 and 3 labeled as in (a) and (b). Filled squares and circles refer to the standard and CAKE integration methods, respectively.

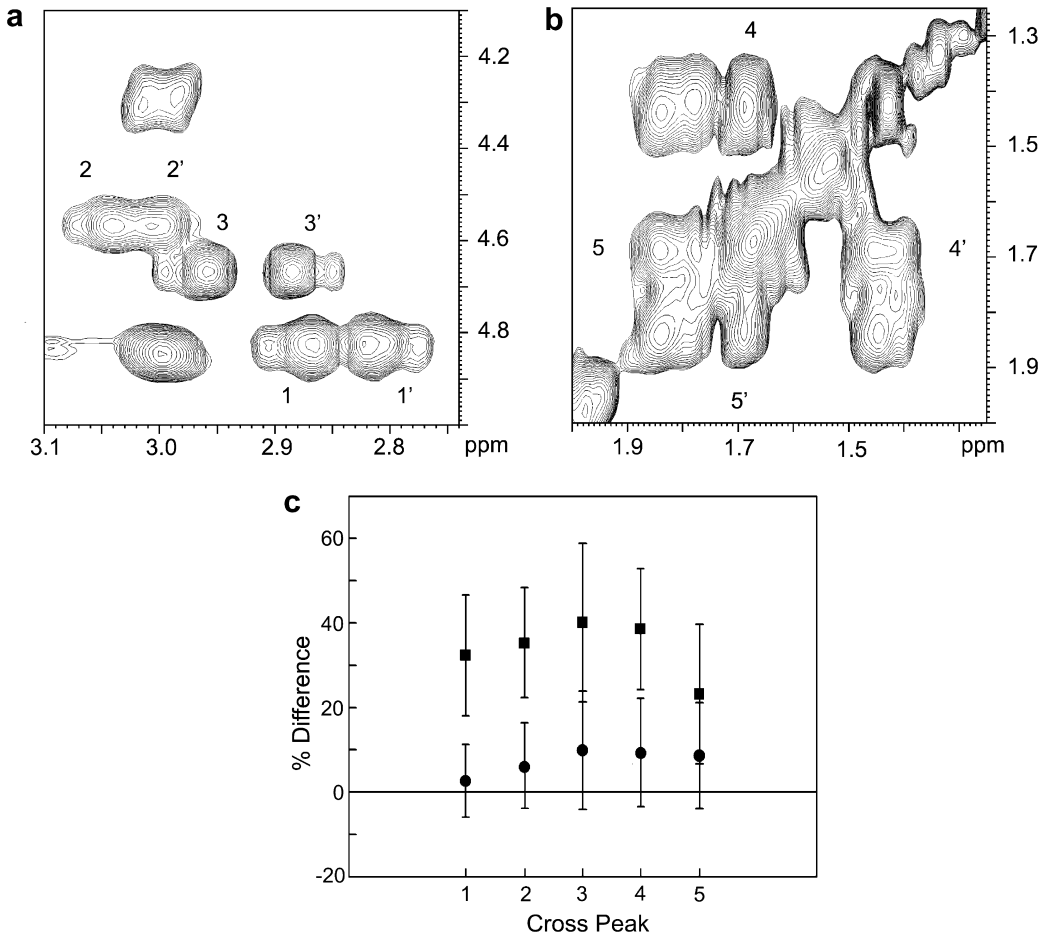

Fig. 7. TOCSY spectrum of sCT in aqueous SDS acquired at 300 K with 70 *ms* mixing time. Expansion (a) shows peaks originating from Asn^3 (peaks 1 and 1′), Cys^7 (peaks 2 and 2′) and Tyr^{22} (peaks 3 and 3′). Expansion (b) depicts the β protons of Leu^4 (peaks 4 and 4′) and Arg^{24} (peaks 5 and 5′). (c) Difference percentage (%) of volume in the selected cross-peak pairs 1–5 labeled as in (a) and (b). Filled squares and circles refer to the standard and CAKE integration methods, respectively.

shows a bias of ca. 20%. Taken together, CAKE applies efficiently to small tripeptides as well as to larger polypeptides. We are currently testing CAKE on NOESY spectra of sCT in aqueous SDS, and we have estimated an average time of ca. 1 min for the integration of each cross-peak. Therefore for an average of 10–20 NOE effects per residue, we estimated an average analysis time of about 10–20 min per amino acid.

3.2.1. Bias vs. digital resolution

The dependence of CAKE on digital resolution was investigated by integrating the peak pair 2 (Fig. 6c) at different digital resolution (0.5, 1.1, 2.2, 4.3 and 8.6 Hz/pt), and integration was carried out for each value with standard and CAKE methods (Fig. 8). The volume of pair 2 overlapping peak (located at $\omega_1 = 4.75$ ppm and $\omega_2 = 3.05$ ppm, Fig. 6c) was compared to the volume of the corresponding single peak at $\omega_1 = 4.75$ ppm and $\omega_2 = 3.22$ ppm at its maximum digital resolution, taken as reference. The values obtained with CAKE are all within 2%, giving an unbiased estimation of the % Difference up to 8.6 Hz/pt. On the contrary, the standard method estimates values $>10\%$ already at 2.2 Hz/pt to become $\approx 25\%$ at 8.6 Hz/pt. This finding can be explained by considering that a low resolution drastically reduces the number of points within an area identified by the *i*-th level, which, in turn, is itself poorly defined. Therefore, the sum of points done by standard methods is obviously biased. On the contrary, the Hit-or-Miss technique used in CAKE does not sum the existing points included in a level bound area, but generates random points and counts the number of "hits" (or points) that are included in the unknown area. Since a cubic interpolation (see Section 2.3.3) is used as a decisional mean to establish if the extracted point can be considered

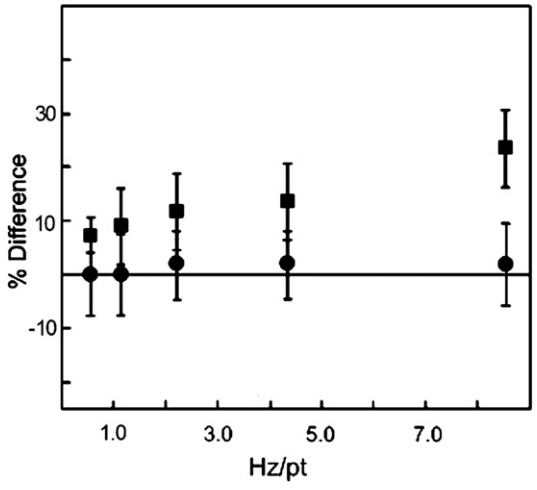

Fig. 8. Difference percentage (%) of volume determination at different resolution for cross-peak 2, as labeled in Fig. 6. The digital resolution was ca. 0.5, 1.1, 2.2, 4.3 and 8.6 Hz/pt. Filled squares and circles refer to the standard and CAKE integration methods, respectively.

a "hit", a low digital resolution could, in principle, affect the peak profile. However, with CAKE we were able to correctly integrate peaks with digital resolution up to ca. 30 Hz/pt.

4. Conclusions

Quantification of NMR spectra is fundamental both in metabolomics/metabonomics and in the structure determination of biomolecules. However, quantification of peaks is often hampered by the degeneracy of the NMR resonance frequency, a factor that aggravates with the increasing size of macromolecules and the number of metabolites. Here we have presented the CAKE approach that uses the symmetry of a single in-phase peak (a peak with a unique center corresponding to its maximum) to calculate its volume. It is obtained by multiplying the fractional volume by the *R* factor, a proportionality ratio between the total and the fractional volume, both evaluated with Monte Carlo techniques. Therefore, the peak volume can be estimated by integrating a known fraction of the peak, and the fractional volume can be chosen so as to minimize the effect of overlap in complex NMR spectra. Strictly speaking CAKE applies to Gaussian peaks showing cylindrical or elliptic symmetry. However, an NMR spectrum is closely approximated by Lorentzian functions, which in its 2D shape show the so-called "star effect". It can be easily removed by 2D Lorentz-to-Gauss transformation, which is routinely used for in-phase experiments, like TOCSY and NOESY. Therefore, the major assumption in this study is that the Lorentzian signal is converted into a Gaussian line by a Lorentz-to-Gauss transformation, which is routinely applied in 2D data manipulation. Integration of simulated and experimental 2D in-phase peaks with different degree of overlap shows that CAKE works well even for strongly overlapping peaks. The main advantage of CAKE is its simplicity as difficulties in its use are comparable to those presented by methods that sum all data points in a defined area. In fact, the user only has to select a peak slice not overlapping with other peaks therefore avoiding the guess of the total contour shape of the peak. Furthermore, CAKE does not require any time-consuming fitting of the peaks to functional forms, and therefore it can be easily incorporated as a subroutine in any NMR processing software. Tests on tripeptides and on sCT have shown that CAKE is a powerful method for volume integration. We are currently applying it to NOESY spectra of calmodulin, a calcium-binding protein of 148 amino acids, and the results will be reported in due course. The substantial independence of CAKE on digital resolution and SNR warrants that it can be safely used for peak integration in three-dimensional spectra. Because of its inherent simplicity the software can be extended to automated integration of three- and possibly higher-dimensionality NMR spectra.

Acknowledgments

This work was supported in part by a grant CNR/MIUR—Legge 449/97. We thank Dominique Melck (ICB-CNR, Pozzuoli) for technical assistance with NMR experiments, and Emilio P. Castelluccio (ICB-CNR, Pozzuoli) for computer maintenance.

References

[1] J.L. Griffin, Review. The Cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball?, Philos Trans. R. Soc. Lond. B. Biol. Sci. 361 (2006) 147–161.
[2] J.C. Lindon, E. Holmes, J.K. Nicholson, Metabonomics in pharmaceutical R&D, FEBS J. 274 (2007) 1140–1151.
[3] J. Cavanagh, W.J. Fairbrother, A.G. Palmer 3rd, N.J. Skelton, M. Rance, Protein NMR Spectroscopy: Principles and Practice, 2nd ed., Elsevier Academic Press, Burlington, MA, USA, 2007.
[4] J.J. Led, H. Gesmar, Quantitative information from complicated nuclear magnetic resonance spectra of biological macromolecules, Methods Enzymol. 239 (1994) 318–345.
[5] G.H. Weiss, J.E. Kiefer, J.A. Ferretti, Accuracy and precision in the estimation of peak areas: the effects of apodization, Chemometr. Intell. Lab. Syst. 4 (1988) 223–229.
[6] C. Rischel, Fundamentals of peak integration, J. Magn. Reson. A 116 (1995) 255–258.
[7] V. Stoven, A. Mikou, D. Piveteau, E. Guitettet, J.Y. Lallemand, PARIS, a program for automatic recognition and integration of 2D NMR signals, J. Magn. Reson. 82 (1989) 163–169.
[8] H. Shen, F.M. Poulsen, Toward automated determination of build-up rates of nuclear overhauser effects in proteins using symmetry projection operators, J. Magn. Reson. 89 (1990) 585–588.
[9] S. Glaser, H.R. Kalbitzer, Automated recognition and assessment of cross peaks in two-dimensional NMR spectra of macromolecules, J. Magn. Reson. 74 (1987) 450–463.
[10] K.P. Neidig, H.R. Kalbitzer, Improved representation of two-dimensional NMR spectra by local rescaling, J. Magn. Reson. 88 (1990) 155–161.
[11] M. Geyer, K.P. Neidig, H.R. Kalbitzer, Automated peak integration in multidimensional NMR spectra by an optimized iterative segmentation procedure, J. Magn. Reson. B 109 (1995) 31–38.
[12] W. Denk, R. Baumann, G. Wagner, Quantitative evaluation of cross-peak intensities by projection of two-dimensional NOE spectra on a linear space spanned by a set of reference resonance lines, J. Magn. Reson. 67 (1986) 386–390.
[13] T.A. Holak, J.N. Scarsdale, J.H. Prestegard, A simple method for quantitative evaluation of cross-peak intensities in two-dimensional NOE spectra, J. Magn. Reson. 74 (1987) 546–549.
[14] C. Eccles, P. Guntert, M. Billeter, K. Wüthrich, Efficient analysis of protein 2D NMR spectra using the software packageEASY, J. Biomol. NMR 1 (1991) 111–118.
[15] H. Gesmar, P.F. Nielsen, J.J. Led, Simple least-squares estimation of intensities of overlapping signals in 2D NMR spectra, J. Magn. Reson. B 103 (1994) 10–18.
[16] R.R. Ernst, G. Bodenhausen, A. Wokaun, Principles of Nuclear Magnetic Resonance in One and Two Dimensions, Clarendon Press, Oxford, 1987.
[17] R. Koradi, M. Billeter, M. Engeli, P. Güntert, K. Wüthrich, Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY, J. Magn. Reson. 135 (1998) 288–297.
[18] C. Griesinger, G. Otting, K. Wüthrich, R.R. Ernst, Clean TOCSY for proton spin system identification in macromolecules, J. Am. Chem. Soc. 110 (1988) 7870–7872.
[19] T.-L. Hwang, A.J. Shaka, Water suppression that works: excitation sculpting using arbitrary waveforms and pulse field gradients, J. Magn. Reson. 112 (1995) 275–279.
[20] A. Bax, D. Davis, MLEV-17 based two-dimensional homonuclear magnetization transfer spectroscopy, J. Magn. Reson. 65 (1985) 355–360.
[21] J.C. Cobas, F.J. Sardina, Nuclear magnetic resonance data processing. MestRe-C: a software package for desktop computers, Concepts Magn. Reson. 19A (2003) 80–96.
[22] J.J. Led, H. Gesmar, Application of the linear prediction method to NMR spectroscopy, Chem. Rev. 91 (1991) 1413–1426.
[23] H. Gesmar, J.J. Led, F. Abildgaard, Improved methods for quantitative spectral analysis of NMR data, Prog. NMR Spectrosc. 22 (1990) 255–288.
[24] J.C. Lindon, A.G. Ferrige, Digitisation and data processing in Fourier transform NMR, Prog. NMR Spectrosc. 14 (1980) 27–66.
[25] G.A. Pearson, Optimization of Gaussian resolution enhancement, J. Magn. Reson. 74 (1987) 541–545.
[26] R.Y. Rubinstein, Simulation and the Monte Carlo Method, Wiley, New York, NY, 1981.
[27] M.H. Kalos, P.A. Whitlock, Monte Carlo Methods, Basics, vol. 1, Wiley, New York, NY, 1986.