Journal of Biomolecular NMR **29:** 505–516, 2004.

NMR spectroscopic filtration of polypeptides and proteins in complex mixtures

Senapathy Rajagopalan[a], Charles Chow[a], Vinodhkumar Raghunathan[b], Charles G. Fry[a] & Silvia Cavagnero[a,*]

[a]*Department of Chemistry, University of Wisconsin-Madison, 1101 University Ave., Madison, WI 53706, U.S.A.;* [b]*Current Address: Department of Chemistry, University of Washington, Seattle, WA 98195, U.S.A.*

Received 29 December 2003; Accepted 9 April 2004

Key words: complex mixture, diffusion, diffusion-edited spectroscopy, HSQC, NMR filtration

Abstract

Due to the inherent complexity of the natural biological environment, most studies on polypeptides, proteins and nucleic acids have so far been performed *in vitro*, away from physiologically relevant conditions. Nuclear magnetic resonance is an ideal technique to extend the *in vitro* analysis of simple model systems to the more complex biological context. This work shows how diffusion-based spectroscopic selection can be combined with isotopic labeling to tackle and optimize the NMR analysis of specific macromolecules in multicomponent mixtures. Typical media include cell-free systems containing overexpressed proteins, lysates and proteolytic mixtures. We present a few variants of diffusion-edited HSQC pulse sequences for the selective spectroscopic detection of protein and polypeptide resonances within complex mixtures containing undesired species of smaller molecular weight. Due to diffusion-based filtering, peak intensities of fast diffusing small molecules are attenuated more than peaks due to large molecules. The basic sequence, denoted as PFGSTE-HSQC, combines translational diffusion-ordering with two dimensional heteronuclear single quantum correlation spectroscopy. The GCSTE-HSQC and BPPSTE-HSQC sequences include bipolar gradients and are therefore suitable for both diffusion-based filtering and determination of diffusion coefficients of individual mixture components. Practical applications range from protein stability/folding investigations in physiologically relevant contexts to prescreening of tertiary fold and resonance assignments in structural genomics studies. A few applications of diffusion-edited HSQC to an *E. coli* cell lysate containing the ^{15}N-labeled B domain of *streptococcal* protein G (GB1), and to a ^{15}N-labeled N-acetylglycine/apomyoglobin mixture are presented. In addition, we provide specific guidelines for experimental setup and parameter optimization.

Abbreviations: PFGSTE – pulse field gradient stimulated echo; GCSTE – gradient-compensated stimulated echo; BPPSTE – bipolar pulse pair stimulated echo; HSQC – heteronuclear single quantum coherence; GB1 – B domain of *streptococcal* protein G (T2E mutant).

Introduction

Most investigations on biomolecular structure and function are traditionally carried out *in vitro*. Nuclear magnetic resonance is an ideal technique to extend this type of analysis, based on isolated biochemical components and simple model systems, to the more diverse and physiologically relevant biological context. This specific context, whose role is presently poorly understood, may well have significant effects on thermodynamic stability, kinetic behavior and overall biological function. In order to properly tackle the above research area, it is desirable to take advantage of a spectroscopic tool, which enables the selective

*To whom correspondence should be addressed. E-mail: cavagnero@chem.wisc.edu

detection of the specific molecule of interest in the context of a complex mixture.

NMR is one of the best presently available techniques for the analysis of complex multi-component samples (Dixon and Larive, 1999). In addition to being ideally suited to high resolution structural analysis, NMR is a non-invasive method, which yields accurate information without the need to resort to the physical separation of individual mixture components. In cases where signal overlap is severe, spectral resolution is enhanced by increasing spectral dimensionality. The elegance of the purely NMR-based approach to mixture analysis is in contrast with the traditional analytical tools employed for this purpose, which have typically included either chromatography or electrophoresis (Manabe, 2000; Wu and MacCoss, 2002), or the combined use of chromatography and either mass spectrometry (GC/MS and HPLC/MS (Tatematsu et al., 1976)) or NMR (HPLC/NMR (Watanabe et al., 1979; Pullen et al., 1995)) or both (Shockcor et al., 1996). A common feature of the multi-technique approaches is the need to physically separate individual solution components to facilitate the subsequent spectral analysis. In addition, these methods tend to be time-consuming and require the use of multiple instruments. Most importantly, separation-based approaches prevent the study of the structural properties of individual molecules within the mixtures, which may be important in the study of complex systems. The above drawbacks make the purely NMR-based approach highly preferable, especially in cases where it is of interest to examine the molecular behavior of one species in the presence of others. The growing interest in the study of biomolecules within their most physiologically relevant environment best illustrates this case. This topic has recently been gaining considerable interest within the structural biology community, especially in relation to the study of the effect of molecular crowding on protein conformation (Minton, 2000; Flaugh and Lumb, 2001; Morar et al., 2001; Patel et al., 2002) and the investigation of protein structural properties in cell-like environments (Serber and Dötsch, 2001; Serber et al., 2001a; Shimizu et al., 2001; Guignard et al., 2002; Brüeggert et al., 2003; Morita et al., 2003).

NMR analysis of individual mixture components requires the selective spectroscopic detection of the species of interest over the remaining molecules. This selection can be achieved by following three primary approaches, i.e., isotopic labeling (Serber et al., 2001b; Shimizu et al., 2001), relaxation-based filtering (Rabenstein and Isab, 1979; Rabenstein et al., 1988) and diffusion-based filtering. Hahn's classic work on spin echoes and stimulated echoes (Hahn, 1950a, b), and the following investigations by Carr and Purcell further elaborating on the subject (Carr and Purcell, 1954), provided early evidence that the extent of translational diffusion influences echo amplitudes in the presence of magnetic field gradients. Soon thereafter, the replacement of continuous gradients by pulsed field gradients (Stejskal and Tanner, 1965) contributed to the growing popularity of diffusion-based NMR spectroscopy. The most popular application of this tool has been the determination of self-diffusion coefficients, leading to information about molecular sizes and shapes, both in purified samples and in complex mixtures (Woessner, 1961; Stejskal and Tanner, 1965; Tanner, 1970; Stilbs, 1981).

The analysis of diffusion coefficients in mixtures has been greatly facilitated by the introduction of an additional 'diffusion dimension' to the processed spectrum, leading to the prolific field of diffusion-ordered spectroscopy (DOSY), introduced by Johnson and coworkers (Morris and Johnson, 1992, 1993). Discriminating distinct mixture components by DOSY is quite difficult in crowded NMR samples unless the peaks belonging to the species to be identified have large differences in diffusion coefficients or are spectrally well resolved. Towards the latter goal, a number of recent implementations have effectively combined DOSY with other resolution enhancing sequences. These include DOSY-COSY (Wu et al., 1996b), DOSY-TOCSY (Birlirakis and Guittet, 1996), and DOSY-NOESY (Gozansky and Gorenstein, 1996).

Diffusion-based spectroscopy has been employed to selectively monitor ligand binding within mixtures (Lin and Shapiro, 1996; Hajduk et al., 1997; Lin et al., 1997; Gounarides et al., 1999; Hodge et al., 2001) and protein self-association (Altieri et al., 1995). In addition, sequences such as DRYCLEAN have been utilized for solvent suppression purposes (VanZijl and Moonen, 1990).

Two and three dimensional pulse sequences involving the combined use of DOSY and heteronuclear coherence selection include DOSY-INEPT and DOSY-DEPT (Wu et al., 1996a), and DOSY-HMQC (Barjat, 1998). For protein work, 2D DOSY-INEPT and DOSY-DEPT yield limited spectral resolution and the 3D DOSY-HMQC sequence suffers from suboptimal lineshapes inherent to HMQC-type experiments (Bax et al., 1990; Norwood et al., 1990). 2D sequences combining longitudinal-eddy-current delay

pulses (LED) (Wu et al., 1995) and heteronuclear single quantum correlation focused on either determining differences in diffusion coefficient of small citric acid derivatives (Parkinson et al., 1998), or following peptide self association and folding (Buevich and Baum, 2002). In general, the combination of diffusion-editing with heteronuclear correlation spectroscopy, typically needed for straightforward assignments and structure determination in biomolecular NMR, is still largely unexplored.

This work employs a combination of selective isotopic labeling of the species of interest (over the remaining mixture components), diffusion-based filtering and HSQC spectral editing to simplify the structural analysis of macromolecules in solutions containing significant amounts of spectrally undesirable small molecules. The peak intensities of the small molecule components are minimized while the intensities of proteins and other macromolecules are selectively retained. Towards this end, we introduce three novel pulse sequences, which improve upon past related efforts and are optimized for the analysis of proteins and other macromolecules within mixtures. These sequences are collectively denoted as 'diffusion-edited HSQC' and are easily implemented on modern NMR spectrometers.

Materials and methods

Protein expression/purification

The *Sperm whale* apomyoglobin gene was subcloned into the pET blue-1 vector (Novagen) by PCR amplification following standard molecular biology techniques (Chow et al., 2003). ^{15}N-labeled apomyoglobin was overexpressed in *E. coli* in ^{15}N enriched M9 minimal medium. Cell growth, lysis and purification were done according to published procedures (Eliezer and Wright, 1996; Cavagnero et al., 1999). The ^{15}N-labeled B domain of *streptococcal* protein G (GB1, T2E mutant) was prepared by similar protocols, except that cell growth was carried out in LB medium until time of induction. Cells were then transferred to ^{15}N enriched M9 medium as described (Serber et al., 2001a), and induced with 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) at 0.6–0.8 optical density at 600 nm. Growth was continued at 37 °C for about two hours.

NMR sample preparation

An N-acetylglycine-apomyoglobin mixture was prepared by dissolving the two ^{15}N labeled molecules in 10 mM phosphate buffer. Final apomyoglobin and N-acetylglycine concentrations were 180 μM and 3 mM, respectively. The solution pH was adjusted to 5.9. An NMR sample containing 1 mM GB1 in *E. coli* cells was prepared starting from gently pelleted cells coming from GB1 expression. Cell growth and expression were performed as described above. The cell pellet was resuspended in minimal medium lacking vitamins and ^{15}N ammonium sulfate but containing an equivalent amount of Na_2SO_4 to compensate for ionic strength losses, to a final 30% v/v slurry concentration (pH 6.9). After the NMR experiments, the sample was gently centrifuged and the supernatant tested. The presence of the residual GB1 ^{1}H,^{15}N–HSQC signals in the supernatant indicated that part of the cells had undergone lysis.

NMR spectroscopy

All experiments were performed on a Varian INOVA-600 MHz NMR spectrometer equipped with a Varian triple resonance ^{1}H{^{13}C, ^{15}N} triple axis gradient probe. Square gradients were employed in all experiments. The experiments involving the N-acetylglycine-apomyoglobin mixture and the GB1 cell lysate were performed at 37 °C. Relaxation delays were set to 1.5 s. GARP (Shaka et al., 1985) was used for ^{15}N decoupling during acquisition for the experiments in Figures 2 and 4.

Diffusion coefficients were measured by running 1D versions of GCSTE- or BPPSTE-HSQC on an 18 mM sample of pure aqueous ^{15}N-labeled N-acetylglycine (also labeled by ^{13}C at its acetyl carbonyl carbon) with diffusion-encoding gradient (g_0) values ranging from 0 to 60 G cm^{-1}. Δ was set to 21.5 and 22.7 ms in GCSTE- and BPPSTE-HSQC, respectively. Modulating g_0 gradient strength at constant Δ ensures constant relaxation effects throughout the experiment. Gradient calibration was done according to published methods (Antalek, 2002), using the known HDO diffusion coefficient in D_2O at 25 °C (Mills, 1973). For D measurements, no decoupling was applied during acquisition to avoid sample overheating and the consequent possible formation of convection currents (Goux et al., 1990; Mau and Kohlmann, 2001). Convection currents may increase effective translational diffusion rates, thereby causing errors in D values. Alternatively, in order to minimize

sensitivity losses arising from signal splitting in the coupled spectrum, low power WURST decoupling can be employed (Barjat, 1998). This option, which was not necessary in our test D measurements on pure N-acetylglycine, becomes highly desirable in the case of complex mixtures where sensitivity and spectral resolution are likely to be an issue. An additional option is incorporation of the double stimulated spin echo sequence (Jerschow and Muller, 1997). Diffusion coefficients based on GCSTE-HSQC were calculated by fitting acetylglycine peak intensities (I) at each gradient strength (g_0) to the relation

$$I = I_0 \exp\left[-\gamma^2 D\delta^2(\Delta - \frac{\delta}{3})(g_0)^2\right], \quad (1)$$

where I is the expected signal intensity at any given diffusion delay (Δ) and diffusion-encoding gradient (g_0) strength values. I_0 denotes the signal intensity when g_0 is set to zero, δ is the duration of g_0, γ is the ^{1}H nuclear gyromagnetic ratio and D is the translational diffusion coefficient. The corresponding expression used for measurements based on BPPSTE-HSQC is

$$I = I_0 \exp\left[-\gamma^2 D(2\delta)^2\left(\Delta - \frac{2\delta}{3} - \frac{\sigma}{2}\right)(g_0)^2\right], \quad (2)$$

where σ is the time interval between the two pulse field gradients forming each bipolar pair (Wu et al., 1995). D values were calculated from the slope of a $\ln(I/I_0)$ *vs* g_0^2 plot.

NMRPipe (Delaglio et al., 1995) and NMRView (Johnson and Blevins, 1994) were used for data processing. Curve fittings were done by Kaleidagraph (Synergy Software) and Excel (Microsoft).

Results

All three versions of the diffusion-edited HSQC pulse sequences, shown in Figure 1, are based on the sensitivity-enhanced HSQC experiment employing pulse field gradients for coherence selection (Kay et al., 1992). In order to achieve molecular size-dependent spectroscopic filtration, the first two ^{1}H pulses in the HSQC sequence have been replaced by diffusion-encoding sequences. The g_1 gradients serve the purpose of spatially dephasing undesired transverse single quantum and antiphase coherence during the diffusion delay Δ. In addition, they assist with solvent suppression.

The first pulse sequence, shown in Figure 1a, is denoted as PFGSTE-HSQC since diffusion-encoding is achieved by a pulse field gradient stimulated echo (PFGSTE) (Tanner, 1970). Alternatively, a gradient-compensated stimulated echo (GCSTE) (Pelta et al., 1998) can be utilized to achieve diffusion-encoding. The resulting GCSTE-HSQC pulse sequence is shown in Figure 1b. In this case, the presence of a bipolar gradient pair sandwiching the second and the third 90° ^{1}H pulse serves the purpose of suppressing eddy currents. These currents are a well known source of lineshape distortion and errors in diffusion coefficient measurements. The third pulse sequence, BPPSTE-HSQC (Figure 1c), includes two additional 180° 1H pulses, each of them sandwiched between bipolar gradient pairs. The additional pulses have been inserted in the intervals before the second and after the third 90° ^{1}H pulse. This sequence is otherwise very similar to PFGSTE-HSQC. Diffusion-based signal attenuation is proportional to the square of the pulse field gradient area, defined as the product of diffusion encoding/decoding gradient strength and gradient duration, as shown in Equation 1. An advantage provided by BPPSTE-HSQC is that the total available diffusion gradient area is four times as large as in the other pulse sequences presented here, when identical g_0 strengths and durations (i.e., δ) are used. A direct implication of the above is that BPPSTE-HSQC allows a greater extent of diffusion-based spectroscopic filtering than PFGSTE-HSQC and GCSTE-HSQC. In addition, since the presence of bipolar gradient pairs suppresses eddy currents, BPPSTE-HSQC is also suitable to determine diffusion coefficients. The performance of all the diffusion-edited HSQC variants discussed above is very similar in terms of signal-to-noise ratio.

In summary, all the above sequences perform comparably in terms of spectroscopic filtration of small molecules, with BPPSTE-HSQC being slightly preferable when very strong diffusion gradient amplitude and duration combinations are needed to reach appropriate levels of filtration.

In the case of the PFGSTE-HSQC and GCSTE-HSQC sequences, the first 90° ^{1}H pulse generates transverse single quantum coherence. The following chemical shift and J-coupling evolution gives rise to both transverse and antiphase coherences, which are subsequently encoded by the first g_0 gradient. The second 90° ^{1}H pulse transfers some of the resulting coherences into longitudinal magnetization (I_z) and longitudinal two-spin order (I_zS_z). These terms are then 'stored' and remain unchanged until the third ^{1}H 90° pulse, while all the remaining transverse (I_x) and antiphase (I_xS_z) coherences are dephased by the

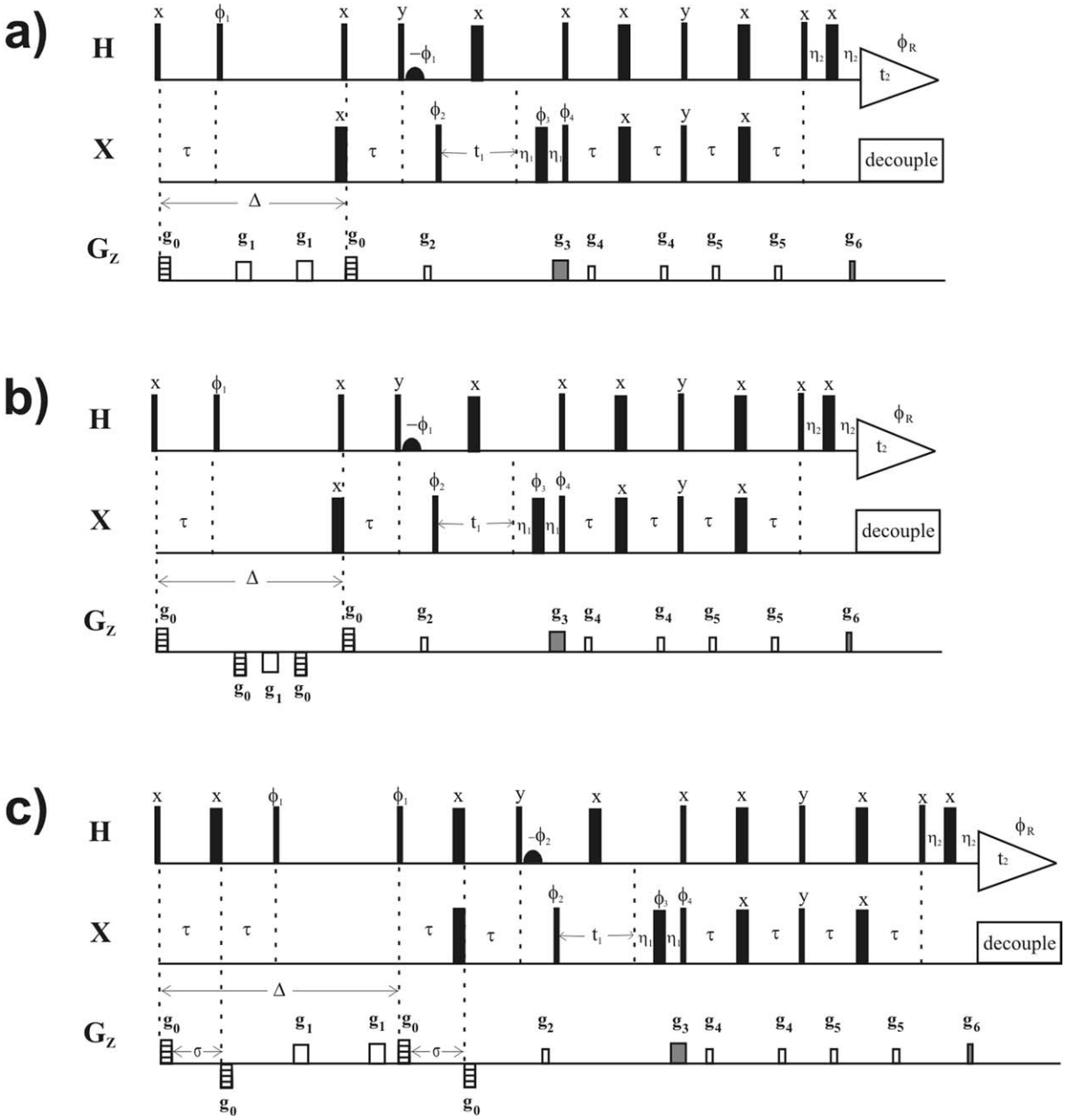

Figure 1. (a) The PFGSTE-HSQC pulse sequence. The delay τ was set to $1/4J_{HX}$ and delays η_1 and η_2 were set to 2.7 ms and 345 μs, respectively. The phase cycling is: ϕ_1 = x, $-$x; ϕ_2= x, ϕ_3 = 2(x), 2(y), 2($-$x), 2($-$y); ϕ_4 = x, ϕ_R = x, 2($-$x), x. The g_0 gradients (duration: 2.2 ms) provide diffusion encoding. The g_3 and g_6 coherence selection gradients were set such that $g_3/g_6 = -\gamma_H/\gamma_X$. The carrier frequency is centered on the solvent and a selective sinc-shaped pulse has been used for 'flip-back'solvent suppression (Grzesiek, 1993). Quadrature detection in the indirect dimension was achieved by collecting two separate data sets with ϕ_4 phase alternating between x and $-$x and concurrent g_1 gradient sign toggling. The two data sets were stored separately and then manipulated as described by Kay et al. (1992) to obtain purely absorptive line shapes. Data were processed according to the STATES method (States et al., 1982). Gradient strengths and durations were g_1 = 55 G cm^{-1}, 2.2 ms; g_2 = 25.3 G cm^{-1}, 1 ms; g_3 = 29.9 G cm^{-1}, 2.5 ms (for X = ^{15}N); g_4 = 4.6 G cm^{-1}, 0.5 ms; g_5 = 6.9 G cm^{-1}, 0.5 ms; g_6 = 29.7 G cm^{-1}, 0.25 ms (for X = ^{15}N). The delay τ was 2.4 ms (for X = ^{15}N). (b) 'Gradient compensated' version (Pelta et al., 1998) of the PFGSTE-HSQC pulse sequence (GCSTE-HSQC). The phase cycling is: ϕ_1 = 8(x, $-$x), 8(y,$-$y); ϕ_2 = 2(x), 2($-$x); ϕ_3 = 4(x), 4(y), 4($-$x), 4($-$y); ϕ_4 = x, ϕ_R = 2(x, 2($-$x), x, $-$x, 2(x), $-$x), 2($-$y, 2(y), $-$y, y, 2($-$y), y). All delays and gradient strengths were as in (a). (c) 'Bipolar pair' version of the PFGSTE-HSQC sequence (BPPSTE-HSQC). The phase cycling is: ϕ_1 = 8(x), 8($-$y); ϕ_2 = x, $-$x; ϕ_3 = 2(x), 2(y), 2($-$x), 2($-$y); ϕ_4 = $-$x, ϕ_R = x, 2($-$x), x. The delay σ was set to 200 μs. All other delays and gradient strengths were as in (a).

g_1 gradient(s). The first 180° X pulse changes the sign of the stored longitudinal two-spin order to $-I_zS_z$. This allows refocusing of antiphase I_xS_z coherence (occurring just before the fourth ^{1}H 90° pulse). The third 90° ^{1}H pulse converts the above longitudinal magnetization (I_z) and longitudinal two-spin order ($-I_zS_z$) into transverse single quantum and antiphase coherences, which then undergo chemical shift and J-coupling evolution. The second diffusion-encoding gradient refocuses some of the resulting coherences to generate pure antiphase coherence (I_xS_z) just before the fourth ^{1}H 90° pulse. The degree of coherence refocusing depends on the extent of translational diffusion experienced by the nuclei during the diffusion delay Δ (Johnson, 1999). The fourth ^{1}H 90° and the first ^{1}X 90° pulse then convert I_xS_z to I_zS_y, thus completing the INEPT transfer. The subsequent pulses and delays mirror those of the standard sensitivity enhanced HSQC experiment and are common to all sequences. Therefore the intensity of the detected signal is modulated by the diffusion coefficient D. The leading features of the above discussion also apply to the BPPSTE-HSQC sequence except that, in this case, the first X 180° pulse allows evolution of scalar couplings, leading to I_xS_z antiphase coherence right before the fourth ^{1}H 90° pulse.

The GCSTE- and BPPSTE-HSQC sequences have been tested by measuring the diffusion coefficient of a simple and readily available isotopically enriched small molecule with non-exchangeable amide protons. N-acetylglycine was chosen for this purpose, and its diffusion coefficient was found to be $8.67 \pm 0.25 \times 10^{-10}$ m^2 s^{-1} and $9.29 \pm 0.38 \times 10^{-10}$ m^2 s^{-1} at 25 °C by BPPSTE- and GCSTE-HSQC, respectively. The accuracy of this value was tested by determining the diffusion coefficient of the same sample by the standard BPPSTE proton pulse sequence (Wu et al., 1995). This measurement yielded a diffusion coefficient of $8.37 \pm 0.40 \times 10^{-10}$ m^2 s^{-1}. The BPPSTE-HSQC and BPPSTE values are identical within experimental error, while the GCSTE-HSQC value is slightly higher than the reference diffusion coefficient obtained by BPPSTE.

In order to assess the effectiveness of the above pulse sequences to selectively filter out spectral contributions due to small molecules, the simplest sequence, PFGSTE-HSQC, was tested on a mixture containing a 17:1 molar ratio of N-acetylglycine and apomyoglobin. The resulting spectra, with diffusion-encoding gradient g_0 set to either (a) 0.023 G cm^{-1} or (b) 60 G cm^{-1}, are shown in Figure 2. The strong resonance arising from N-acetylglycine is dramatically attenuated at higher gradient strengths while the protein resonances experience a much more moderate decrease in intensity. This pulse sequence is therefore effective at selectively filtering out peaks due to small molecules in heteronuclear correlation experiments. The degree of signal suppression for small molecules can be qualitatively assessed by comparing the first row ^{1}H traces of the PFGSTE-HSQC spectra, collected with g_0 set to either 0.023 G cm^{-1} (Figure 2a) or 60 G cm^{-1} (Figure 2b).

A more detailed assessment of the extent of signal suppression experienced by the apomyoglobin and N-acetylglycine resonances was obtained by a systematic comparison of peak intensities as a function of both diffusion-encoding gradient strength and diffusion delay. The corresponding plots (Figure 3) show that the intensity of the N-acetylglycine amide proton resonance decreases more rapidly than the apomyoglobin resonances as diffusion-encoding gradient strength or diffusion delay increase. A relaxation delay of 1–1.5 s was found to be optimal (at 600 MHz) to ensure efficient spectroscopic filtration of resonances due to small molecules while preserving relatively fast data acquisition. As shown in Figure 3, the use of this delay value does not significantly affect the overall curve shape but it has a small effect on peak intensities. However, we found that diffusion coefficients are not affected, within experimental error, by the use of either a short or a longer (about $5\times$ T_1) relaxation delay.

The PFGSTE-HSQC sequence was also tested on a complex biological mixture, i.e., an *E. coli* cell lysate resulting from overexpression of GB1 (Figure 4). The ^{1}H,^{15}N-HSQC spectrum of the crude lysate looks very similar to that of the purified protein (K.H. Gardner, unpublished data). In addition, some spurious resonances are reproducibly present (circled peaks). These arise from endogenously produced isotopically labeled small molecules. The circled peaks, visible in the PFGSTE-HSQC spectrum at low gradient strengths (Figure 4a), selectively disappear when strong diffusion-encoding gradients are applied (Figure 4b).

An additional stringent test for molecular size-dependent selectivity in signal filtration is provided by plotting the percent intensity attenuations by all the resonances in Figure 4 (data not shown). These attenuations were calculated as $(I_{0.023}\text{-}I_{60}) \times 100/I_{0.023}$, where I denotes the peak intensity. The subscript refers to the specific g_0 value in G cm^{-1} used in the experiment. The average percent intensity attenuation for

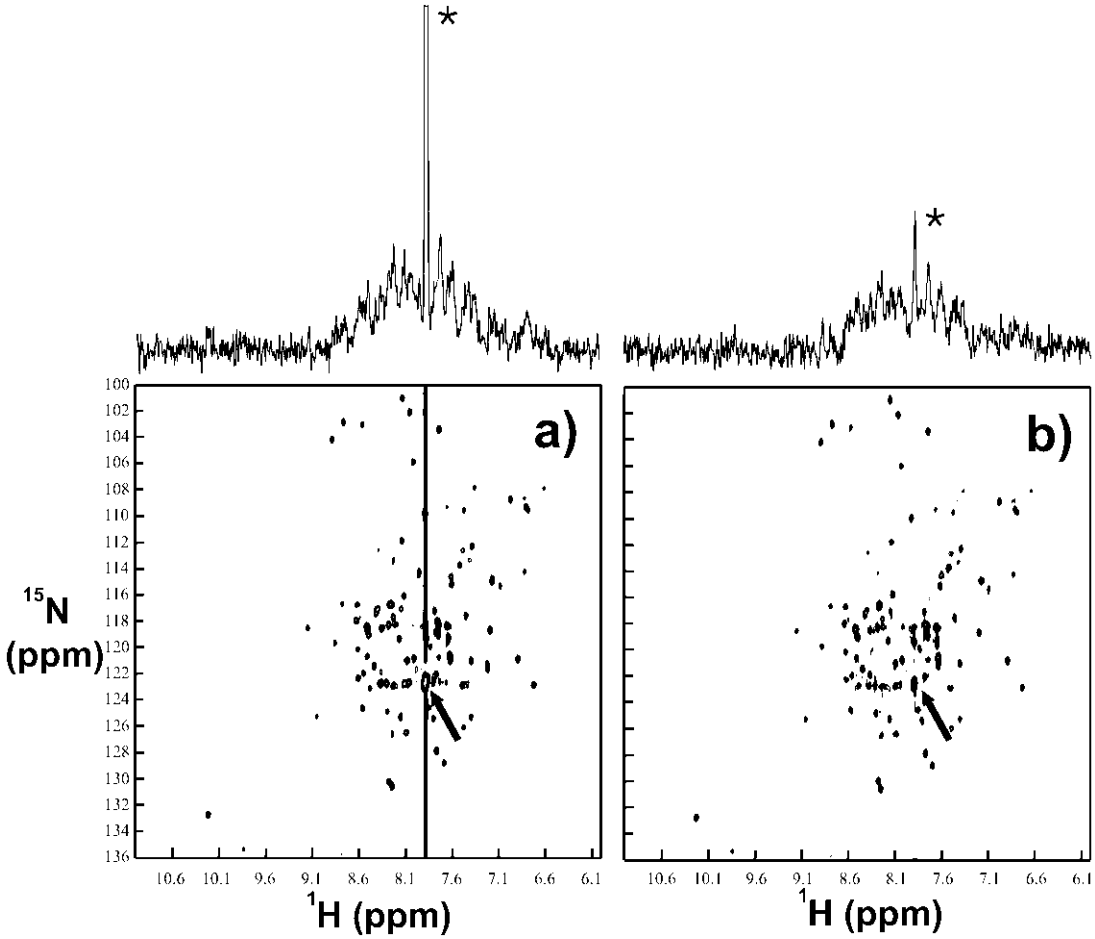

Figure 2. PFGSTE-HSQC spectra and corresponding 1D first row ^{1}H traces for an N-acetylyglycine-apomyoglobin mixture. The diffusion encoding gradient g_0 is set to either (a) 0.023 G cm^{-1}, or (b) 60 G cm^{-1}. The diffusion delay Δ is set to 20.4 ms in both experiments. Spectral widths were set to 9000 and 2200 Hz in the direct and indirect dimensions, respectively. Data were acquired as 512(t_1) × 4096(t_2) total points, apodized with an unshifted gaussian function in both dimensions and zero filled to 1024(t1) × 8192(t2) complex points. The arrow and asterisk denote the N-acetylglycine resonance in the 2D spectrum and ^{1}H trace, respectively.

the resonances arising from fast diffusing non-GB1 species was 92%, i.e., a significantly greater value than the average protein peak attenuation. Under optimized conditions, the difference between the percent intensity attenuations of a globular protein and a small molecule can be up to about 70%. A useful diagnostic feature, which may be exploited to selectively identify protein peaks, is the nearly-uniform level of percent intensity attenuation experienced by the resonances belonging to the protein. The standard deviation for the GB1 peak intensities was smaller than 2%. Unexpectedly, two GB1 amide backbone resonances (i.e., T11 and E19) showed an 8% higher percent intensity attenuation relative to the other protein peaks. T11 is located in the loop connecting the first two β-strands while E19 is close to the C-terminal portion of the second β-strand. Backbone dynamics studies show that these residues have no unusual T_1, T_2, steady-state NOEs or order parameters (Idiyatullin et al., 2003). Rapid chemical exchange by T11 and E19 amide protons with a cellular small molecule component may explain the above effects (Johnson, 1999), although no evidence is currently available to support this intriguing hypothesis.

In order to obtain further insights on the dependence of signal intensities on diffusion-encoding gradient strength and diffusion delay, we have generated theoretical plots illustrating the expected signal intensities for the diffusion-edited HSQC sequences (Figure 5). The plots are based on Equation 1. This

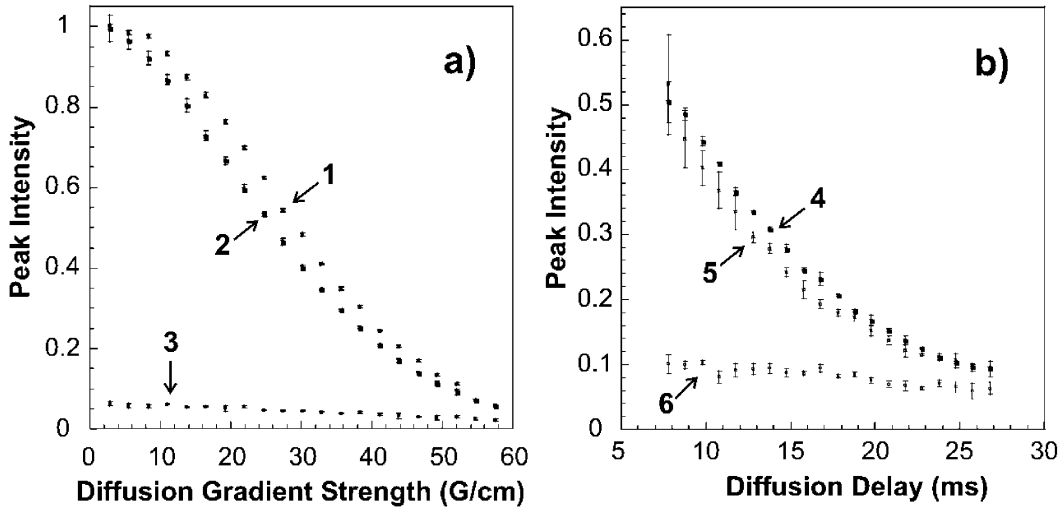

Figure 3. PFGSTE-HSQC intensities of N-acetylglycine (traces 1, 2, 4 and 5) and apomyoglobin amide proton resonances (traces 3 and 6) in the mixture of Figure 2 as a function of (a) diffusion gradient strength and (b) diffusion delay. Representative values for apomyoglobin were determined by summing the amide proton peak intensities of A127, L61, R118, H36 and F43 residues. The relaxation delay for this experiment was set to either 1.5 s (plots 2, 3, 5 and 6) or 12 s (plot 4) and 15 s (plot 1). Error bars indicate standard deviations for two or three independent experiments.

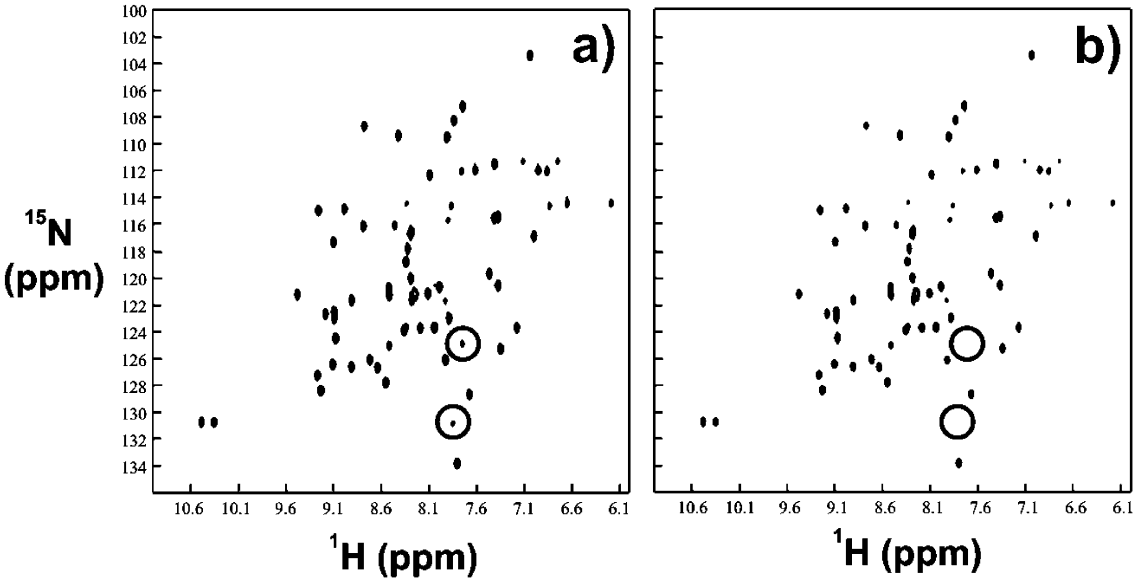

Figure 4. PFGSTE-HSQC spectra of the B domain of *Streptococcal* protein G (GB1) in an *E. coli* cell lysate. The diffusion encoding gradient g_0 is set to either (a) 0.023 G cm^{-1} or (b) 60 G cm^{-1}. The diffusion delay Δ is set to 55.4 ms in both experiments. Spectral widths were set to 9000 and 3500 Hz in the direct and indirect dimensions. Data were acquired as 256(t1) $\times$ 4096(t2) total points, apodized with a gaussian function in both dimensions and zero filled to 512 (t_1) $\times$ 8192(t_2) complex points. The circled peaks present in (a) are reduced to noise level in (b).

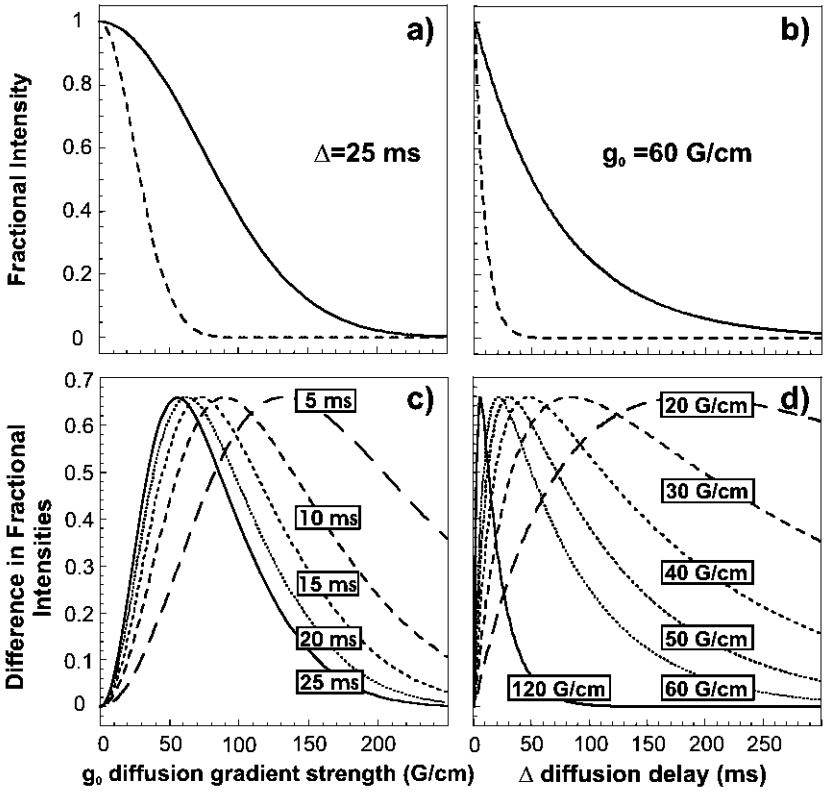

Figure 5. Theoretical I/I_0 fractional intensities for N-acetylglycine (- - -) and apomyoglobin (——) as a function of (a) diffusion-encoding gradient strength and (b) diffusion delay. Values were calculated according to Equation (1), which applies to PFGSTE-HSQC and GCSTE-HSQC. Gradient durations δ were set to 2.2 ms. Difference in theoretical I/I_0 values for N-acetylglycine and apomyoglobin (c) as a function of g_0 for different values of the diffusion delays Δ, and (d) as a function of Δ for different g_0 gradient amplitudes.

equation does not take relaxation into account and applies to the PFGSTE-HSQC and GCSTE-HSQC sequences. Figure 5 (panel a and b) shows the expected fractional intensities (I/I_0) for N-acetylglycine and apomyoglobin as a function of diffusion-encoding gradient strength and diffusion delay, respectively. The representative published D value for apomyoglobin at 22 °C, i.e., 1.12×10^{-10} m^2 s^{-1} (Papadopoulos et al., 2000) and the N-acetylglycine D value determined here were used to generate the plots. The graph shows that small molecule peak intensities decrease more steeply relative to large molecule peak intensities at increasing g_0 and Δ values. On the other hand, N-acetylglycine and apomyoglobin I/I_0 values are identical at very low and very high g_0 gradient strengths and diffusion delay Δ durations. The expected differences between the fractional intensities of N-acetylglycine and apomyoglobin as a function of diffusion-encoding gradient strengths and diffusion delays are plotted in Figure 5c and 5d, respectively. From these plots, it is evident that, for each diffusion-encoding gradient strength, there is an optimal diffusion delay Δ_{OPT} which maximizes the difference between the I/I_0 values due to small and large molecules. Therefore, a proper selection of diffusion delay at each diffusion-encoding gradient strength is crucial to optimize the quality of spectroscopically filtered HSQC spectra emphasizing resonances due to large molecules.

An optimized diffusion delay (Δ_{OPT}) can be theoretically determined for each diffusion-encoding gradient strength by maximizing the function representing the difference between the fractional intensities of small and large molecules with respect to the diffusion delay. This function can be expressed as

$$\left[\frac{I_L}{I_{0L}} - \frac{I_S}{I_{0S}}\right] = \exp\left(-\gamma^2 D_L \delta^2 (\Delta - \frac{\delta}{3}) g_0^2\right) - \exp\left(-\gamma^2 D_S \delta^2 (\Delta - \frac{\delta}{3}) g_0^2\right). \quad (3)$$

The large and small molecules are denoted by the 'L' and 'S' subscripts, respectively. Alternatively, the

fractional intensity difference function can be maximized with respect to diffusion-encoding gradient for a fixed Δ value. The optimal diffusion delay Δ_{OPT} for the PFGSTE and GCSTE HSQC sequences is

$$\Delta_{OPT} = \frac{\ln (D_S/D_L)}{\gamma^2\delta^2 g_0^2 (D_S - D_L)} + \frac{\delta}{3}. \quad (4)$$

In the case of BPPSTE-HSQC, an additional term, $\sigma/2$, is added to the right hand side of the above equation and δ is replaced by 2δ. The maximal achievable fractional intensity difference for a small and large mixture components can be achieved by substituting the optimized Δ-g_0 pair from Equation 4 into Equation 3. This leads to

$$\left(\frac{I_L}{I_{0L}} - \frac{I_S}{I_{0S}}\right)_{MAX} = \left(\frac{D_L}{D_S}\right)^{\frac{-D_L}{D_L - D_S}} \left[1 - \left(\frac{D_L}{D_S}\right)\right]. \quad (5)$$

The above relation shows that the maximum difference in intensity depends only on the ratio of the diffusion coefficients for the two species as long as an optimal Δ-g_0 pair is used. Rational criteria to select optimal Δ-g_0 pairs are discussed in the Supporting Information.

Discussion

The pulse sequences introduced in this study are suitable to simplify the analysis of complex *in vitro* mixtures by combining heteronuclear correlation and translational diffusion-based spectral editing. The intensities of resonances due to small molecules are minimized while resonances due to large molecules are selectively preserved. Viable applications include proteomics and structural genomic studies involving the screening of expression conditions and NMR spectral quality in lysates and cell-free systems containing novel proteins. The sequences presented here are also applicable to studies targeting the conformational and thermodynamic properties of bioactive macromolecules in a physiologically relevant context. Viable media include tissue and biopsy homogenates, cell lysates, cell-free systems and conventional multicomponent solutions. In all cases spectral lineshapes for the species of interest have to be sufficiently sharp in spite of the possible presence of magnetic susceptibility gradients generated by inhomogeneous cellular components (Springer, 1994; Wieruszeski et al., 2001). This issue is highly dependent on the sample nature and needs to be carefully evaluated in each case. Respectable spectral quality has been obtained in heteronuclear correlation experiments of complex mixtures presented here (Figure 4) and elsewhere (Serber and Dötsch, 2001; Serber et al., 2001a, b; Guignard et al., 2002; Brüeggert et al., 2003; Morita et al., 2003).

The presence of NMR peaks not belonging to the protein of interest in complex biological samples has been documented before (Ou et al., 2001; Serber et al., 2001b; Morita et al., 2003) and attributed to the presence of products of amino acid metabolism (Ou et al., 2001; Serber et al., 2001b). This can complicate preliminary NMR analysis and assignments of proteins in complex media. Therefore it is desirable to minimize the intensity of these resonances and, if possible, eliminate them.

In order to maximize the degree of size-dependent selectivity in signal suppression upon spectroscopic filtration, an optimized Δ-g_0 pair must be used. This pair typically comprises a large diffusion-encoding gradient g_0 and small diffusion delay Δ or vice versa (Figure 5, panels c and d). The first option, involving use of a large g_0, is generally to be preferred since signal losses due to $-I_z$ and $-I_zS_z$ T_1 relaxation during the diffusion delay Δ are minimized.

In general, molecular diffusion coefficients are directly proportional to temperature and inversely proportional to the viscosity of the medium, regardless of molecular shape (van Holde et al., 1998). In case no significant changes in shape or size take place as a result of viscosity or temperature changes, the diffusion coefficient ratios in Equation (5) are to be regarded as temperature- and viscosity-independent. This is an important conclusion, which implies that, under the above conditions, the efficiency of spectroscopic filtration is not expected to decrease in complex biological media, which are typically more viscous than conventional solutions. For instance, the viscosity of cell lysates (Williams et al., 1997), subcellular compartments (Dayel et al., 1999) and animal tissues (Lebihan et al., 1991) is between 1.2- to 2-fold larger than the value obtained in pure water at physiologically relevant temperatures. As seen in the Results section, size-dependent spectroscopic filtration was easily achieved for an *E. coli* GB1 lysate at 37 °C. In more heavily viscous mixtures or at low temperatures, however, linewidth broadening resulting from very long correlation times may limit the general applicability of NMR analysis unless TROSY-type (Pervushin et al., 1997; Riek et al., 1999) sequences were to be employed. These experiments work optimally if the species of interest have deuterated carbon and fully protonated amide nitrogens. *In situ* generation of such species may pose special practical challenges in the

case of complex biological mixtures. For instance, the conventional selective backbone amide reprotonation, which follows production of fully deuterated proteins, may need untolerably harsh conditions.

In practice, the degree of small molecule spectroscopic filtration within complex mixtures can be optimized by judicious choices of acquisition parameters. It is generally convenient to start by setting g_0 to the largest value allowed by the NMR spectrometer. The matching optimal diffusion delay can then be determined from Equation 4, if D values for the small and large molecules can be estimated. Additional guidelines on parameter optimization are available as Supporting Information.

In addition to the pulse sequences presented here, other related diffusion-based sequences involving heteronuclear correlation have been reported in the literature and used for purposes different from spectroscopic filtration of mixtures. Implementations by Parkinson et al. (1998) and Buevich *et al.* (Buevich and Baum, 2002) have taken advantage of the in-phase single quantum coherence (I_y) generated at the end of a longitudinal-eddy-current delay (LED) sequence (Wu et al., 1995) by replacing the first ^{1}H pulse, in either the basic or sensitivity-enhanced HSQC, with LED. The resulting pulse sequences are overall longer than PFGSTE- GCSTE- and BPPSTE-HSQC and provide some additional time ($\geq \frac{1}{2J_{HX}}$) for transverse relaxation to take place. This is not ideal for the spectroscopic filtration of protein samples, which usually have short T_2 values. The sequences introduced here are more specifically tailored to the spectroscopic analysis of proteins in mixtures than LED-based HSQC experiments.

The concept of diffusion-edited heteronuclear correlation and its use in small molecule NMR filtration can in principle be extended towards the development of triple resonance analogs of HSQC such as diffusion-edited HNCA, HNCO and related sequences. However, due to the anticipated overall signal losses introduced by the diffusion-editing pulses (we experienced sensitivity drops ranging from 60% to 85% from regular to diffusion-edited HSQC) this may not be a convenient implementation. A more straightforward procedure, in the course of triple-resonance-based resonance assignments, is to simply exclude from triple resonance analysis the peaks that have been identified as impurities or undesired species by independently performed diffusion-edited HSQC analysis. The signal losses experienced in diffusion-edited HSQC, relative to regular HSQC, result from the fact that, during the stimulated echo portion of the sequence, only part of the overall magnetization is stored (as gradient-encoded $\pm I_z$ and $\pm I_z S_z$) while the remaining terms are lost by being wiped out by the g_1 gradient(s). Additional potential contributions to signal losses may arise from T_1 relaxation of $-I_z$ and $-I_z S_z$ during the diffusion delay. The above is not a problem provided that the concentration of the macromolecule of interest within the complex mixture is not too low.

Acknowledgements

We would like thank Dr Kevin H. Gardner for the generous gift of the plasmid encoding GB1 (T2E mutant). The National Science Foundation (grant #0215368), the Research Corporation (Research Innovation Award) and the Milwaukee Foundation (Shaw Scientist Award) are gratefully acknowledged for supporting this research.

Supplementary material

Practical guidelines on experimental setup and parameter optimization are available free of charge via the Internet at http://www.kluweronline.com/issn/0925-2738.

References

Altieri, A.S., Hinton, D.P. and Byrd, R.A. (1995) *J. Am. Chem. Soc.*, **117**, 7566–7567.
Antalek, B. (2002) *Conc. Magn. Reson.*, **14**, 225–258.
Barjat, H., Morris, G.A. and Swanson, A.G. (1998) *J. Magn. Reson.*, **131**, 131–138.
Bax, A., Ikura, M., Kay, L.E., Torchia, D.A. and Tschudin, R. (1990) *J. Magn. Reson.*, **86**, 304–318.
Birlirakis, N. and Guittet, E. (1996) *J. Am. Chem. Soc.*, **118**, 13083–13084.
Brüeggert, M., Rehm, T., Shanker, S., Georgescu, J. and Holak, T.A. (2003) *J Biomol. NMR*, **25**, 335–348.
Buevich, A.V. and Baum, J. (2002) *J. Am. Chem. Soc.*, **124**, 7156–7162.
Carr, H.Y. and Purcell, E.M. (1954) *Phys. Rev.*, **94**, 630–638.
Cavagnero, S., Dyson, H.J. and Wright, P.E. (1999) *J. Biomol. NMR*, **13**, 387–391.
Chow, C.C., Chow, C., Raghunathan, V., Huppert, T.J., Kimball, E.B. and Cavagnero, S. (2003) *Biochemistry*, **42**, 7090–7099.
Dayel, M.J., Hom, E.F.Y. and Verkman, A.S. (1999) *Biophys. J.*, **76**, 2843–2851.
Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) *J. Biomol. NMR*, **6**, 277–293.
Dixon, A.M. and Larive, C.K. (1999) *Appl. Spectr.*, **53**, 426A–440A.
Eliezer, D. and Wright, P.E. (1996) *J. Mol. Biol.*, **263**, 531–538.

Flaugh, S.L. and Lumb, K.J. (2001) *Biomacromology*, **2**, 538–540.
Gounarides, J.S., Chen, A.D. and Shapiro, M.J. (1999) *J. Chromatogr. B*, **725**, 79–90.
Goux, W.J., Verkruyse, L.A. and Salter, S.J. (1990) *J. Magn. Reson.*, **88**, 609–614.
Gozansky, E.K. and Gorenstein, D.G. (1996) *J. Magn. Reson. B*, **111**, 94–96.
Grzesiek, S. and Bax, A. (1993) *J. Am. Chem. Soc.*, **115**, 12593–12594.
Guignard, L., Ozawa, K., Pursglove, S.E., Otting, G. and Dixon, N.E. (2002) *FEBS Lett.*, **524**, 159–162.
Hahn, E.L. (1950a) *Phys. Rev.*, **80**, 297–298.
Hahn, E.L. (1950b) *Phys. Rev.*, **80**, 580–601.
Hajduk, P.J., Olejniczak, E.T. and Fesik, S.W. (1997) *J. Am. Chem. Soc.*, **119**, 12257–12261.
Hodge, P., Monvisade, P., Morris, G.A. and Preece, I. (2001) *Chem. Commun.*, **3**, 239–240.
Idiyatullin, D., Nesmelova, I., Daragan, V.A. and Mayo, K.H. (2003) *Protein Sci.*, **12**, 914–922.
Jerschow, A. and Muller, N. (1997) *J. Magn. Reson.*, **125**, 372–375.
Johnson, B.A. and Blevins, R.A. (1994) *J. Biomol. NMR*, **4**, 603–614.
Johnson, C.S. (1999) *Prog. Nucl. Magn. Reson. Spectrosc.*, **34**, 203–256.
Kay, L.E., Keifer, P. and Saarinen, T. (1992) *J. Am. Chem. Soc.*, **114**, 10663–10665.
Lebihan, D., Moonen, C.T.W., VanZijl, P.C.M., Pekar, J. and Despres, D. (1991) *J. Comp. Ass. Tomogr.*, **15**, 19–25.
Lin, M. and Shapiro, M.J. (1996) *J. Org. Chem.*, **61**, 7617–7619.
Lin, M., Shapiro, M.J. and Wareing, J.R. (1997) *J. Am. Chem. Soc.*, **119**, 5249–5240.
Manabe, T. (2000) *Electrophoresis*, **21**, 1116–1122.
Mau, X.-A. and Kohlmann, O. (2001) *J. Magn. Reson.*, **150**, 35–38.
Mills, R. (1973) *J. Phys. Chem.*, **77**, 685–688.
Minton, A.P. (2000) *Curr. Opin. Struct. Biol.*, **10**, 34–39.
Morar, A.S., Wang, X. and Pielak, G.J. (2001) *Biochemistry*, **40**, 281–285.
Morita, E.H., Sawasaki, T., Tanaka, R., Endo, Y. and Kohno, T. (2003) *Protein Sci.*, **12**, 1216–1221.
Morris, K.F. and Johnson, C.S. (1993) *J. Am. Chem. Soc.*, **115**, 4291–4299.
Morris, K.F. and Johnson, C.S.J. (1992) *J. Am. Chem. Soc.*, **114**, 3139–3141.
Norwood, T.J., Boyd, J., Heritage, J.E., Soffe, N. and Campbell, I.D. (1990) *J. Magn. Reson.*, **87**, 488–501.
Ou, H.D., Lai, H.C., Serber, Z. and Dotsch, V. (2001) *J. Biomol. NMR*, **21**, 269–273.
Papadopoulos, S., Jurgens, K.D. and Gros, G. (2000) *Biophys. J.*, **79**, 2084–2094.
Parkinson, J.A., Sun, H.Z. and Sadler, P.J. (1998) *Chem. Comm.*, **8**, 881–882.
Patel, C.N., Noble, S.M., Weatherly, G.T., Tripathy, A., Winzor, D.J. and Pielak, G.J. (2002) *Protein Sci.*, **11**, 997–1003.
Pelta, M.D., Barjat, H., Morris, G.A., Davis, A.L. and Hammond, S.J. (1998) *Magn. Reson. Chem.*, **36**, 706–714.
Pervushin, K., Riek, R., Wider, G. and Wüthrich, K. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 12366–12371.
Pullen, F.S., Swanson, A.G., Newman, M.J. and Richards, D.S. (1995) *Rapid Comm. Mass Spectrosc.*, **9**, 1003–1006.
Rabenstein, D.L. and Isab, A.A. (1979) *J. Magn. Reson.*, **36**, 1969–1992.
Rabenstein, D.L., Millis, K.K. and Strauss, E.J. (1988) *Anal. Chem.*, **60**, 1380A–1391A.
Riek, R., Wider, G., Pervushin, K. and Wütrich, K. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 4918–4923.
Serber, Z. and Dötsch, V. (2001) *Biochemistry*, **40**, 14317–14323.
Serber, Z., Keatinge-Clay, A.T., Ledwidge, R., Kelly, A.E., Miller, S.M. and Dötsch, V. (2001a) *J. Am. Chem. Soc.*, **123**, 2446–2447.
Serber, Z., Lewidge, R., Miller, S.M. and Dötsch, V. (2001b) *J. Am. Chem. Soc.*, **123**, 8895–8901.
Shaka, A.J., Barker, P.B. and Freeman, R. (1985) *J. Magn. Reson.*, **64**, 547–552.
Shimizu, Y., Inoue, A., Tomari, Y., Suzuki, T., Yokogawa, T., Nishikawa, K. and Ueda, T. (2001) *Nat. Biotech.*, **19**, 751–755.
Shockcor, J.P., Unger, S.E., Wilson, I.D., Foxall, P.J.D., Nicholson, J.K. and Lindon, J.C. (1996) *Anal. Chem.*, **68**, 4431–4435.
Springer, C.S. (1994) In *NMR in Physiology and Biomedicine*, Gillies, R.J., Ed., Academic Press, San Diego, pp. 75–99.
States, D.J., Haberkorn, R.A. and Ruben, D.J. (1982) *J. Magn. Reson.*, **48**, 286–292.
Stejskal, E.O. and Tanner, J.E. (1965) *J. Chem. Phys.*, **42**, 288–292.
Stilbs, P. (1981) *Anal. Chem.*, **53**, 2135–2137.
Tanner, J.E. (1970) *J. Chem. Phys.*, **52**, 2523–2526.
Tatematsu, A., Miyazaki, H., Suzuki, M., Miyazaki, H. and Suzuki, M. (1976) *KAKTAF*, **29**, 891–896.
van Holde, K.E., Johnson, W.C. and Ho, P.S. (1998) *Principles of Physical Biochemistry*, Prentice-Hall, NJ.
VanZijl, P.C.M. and Moonen, C.T.W. (1990) *J. Magn. Reson.*, **87**, 18–25.
Watanabe, N., Niki, E. and Shimizu, S. (1979) *JEOL News*, **15A**, 2–5.
Wieruszeski, J.M., Bohin, J.P. and Lippens, G. (2001) *J. Magn. Res.*, **151**, 118–123.
Williams, S.P., Haggie, P.M. and Brindle, K.M. (1997) *Biophys. J.*, **72**, 490–498.
Woessner, D.E. (1961) *J. Chem. Phys.*, **34**, 2057–2061.
Wu, C.C. and MacCoss, M.J. (2002) *Curr. Opin. Mol. Ther.*, **4**, 242–250.
Wu, D., Chen, A. and Johnson, C.S. (1995) *J. Magn. Reson. A*, **115**, 260–264.
Wu, D.H., Chen, A.D. and Johnson, C.S. (1996a) *J. Magn. Reson. A*, **123**, 215–218.
Wu, D.H., Chen, A.D. and Johnson, C.S. (1996b) *J. Magn. Reson. A*, **121**, 88–91.