J Biomol NMR (2014) 60:51–66
DOI 10.1007/s10858-014-9851-1

ARTICLE

VirtualSpectrum, a tool for simulating peak list for multi-dimensional NMR spectra

Jakob Toudahl Nielsen · Niels Chr. Nielsen

Received: 17 February 2014 / Accepted: 1 August 2014 / Published online: 14 August 2014

Abstract NMR spectroscopy is a widely used technique for characterizing the structure and dynamics of macromolecules. Often large amounts of NMR data are required to characterize the structure of proteins. To save valuable time and resources on data acquisition, simulated data is useful in the developmental phase, for data analysis, and for comparison with experimental data. However, existing tools for this purpose can be difficult to use, are sometimes specialized for certain types of molecules or spectra, or produce too idealized data. Here we present a fast, flexible and robust tool, VirtualSpectrum, for generating peak lists for most multi-dimensional NMR experiments for both liquid and solid state NMR. It is possible to tune the quality of the generated peak lists to include sources of artifacts from peak overlap, noise and missing signals. VirtualSpectrum uses an analytic expression to represent the spectrum and derive the peak positions, seamlessly handling overlap between signals. We demonstrate our tool by comparing simulated and experimental spectra for different multi-dimensional NMR spectra and analyzing systematically three cases where overlap between peaks is particularly relevant; solid state NMR data, liquid state NMR homonuclear ^{1}H and ^{15}N-edited spectra, and 2D/3D heteronuclear correlation spectra of unstructured proteins. We analyze the impact of protein size and secondary structure on peak overlap and on the accuracy of structure determination based on data of different qualities simulated by VirtualSpectrum.

Keywords Software · Simulation of spectra · Solid state NMR · Intrinsically disorder proteins · Protein structure · Peak overlap

Introduction

Structures of proteins and nucleic acids have been determined by NMR spectroscopy for almost three decades now (Wüthrich 1986; Williamson et al. 1985; Driscoll et al. 1989). The fundamental processes of assignment of resonances to nuclei and assignments of NOE signals, to derive distance constraints, still remain challenging problems, particularly where there is significant signal overlap. This is especially true for large proteins, natural abundance studies, largely unstructured or intrinsically disordered proteins (IDPs) (Wright and Dyson 1999), and proteins studied by solid state NMR (ssNMR) (see references below). Software for simulating different NMR spectra on the fly based on derived or expected resonance assignments (and a structure model in case of through-space experiments such as NOESY (Jeener et al. 1979)), is valuable for validating both the candidate resonance assignments and/or the model structure. Software has been developed for the automation and semi-automatic/iterative analysis of the resonance assignment (Bartels et al. 1997; Zimmerman et al. 1997; Jung and Zweckstetter 2004; Malmodin et al. 2003; Moseley et al. 2001; Moseley and Montelione 1999) as reviewed in Guerry and Herrmann 2011, and has

Electronic supplementary material The online version of this article (doi:10.1007/s10858-014-9851-1) contains supplementary material, which is available to authorized users.

J. T. Nielsen (✉) · N. Chr. Nielsen
Department of Chemistry, Center for Insoluble Protein Structures (inSPIN), Interdisciplinary Nanoscience Center (iNANO), University of Aarhus, Gustav Wieds Vej 14, 8000 Aarhus C, Denmark
e-mail: jtn@chem.au.dk

N. Chr. Nielsen
e-mail: ncn@inano.au.dk

recently also been developed for solid state NMR (Moseley et al. 2010; Tycko and Hu 2010; Hu et al. 2011). Similar software also exist to handle the assignment of through space signals from both NOESY type experiments (Herrmann et al. 2002; Linge et al. 2001, 2003; Nilges 1995; Nilges et al. 1997; Rieping et al. 2007) and ssNMR experiments, such as DARR (Takegoshi et al. 2001; Fossi et al. 2005; Loquet et al. 2010). Both processes use lists of picked peaks as input and therefore a procedure for generating synthetic peak lists would be useful for evaluating the performance of the software. Such software makes it possible to assess the efficiency of an NMR experiment by simulating the corresponding peak list and evaluating its impact by including it in data sets for performing either automated resonance assignment or structure calculation, before performing experiments.

Software exist to simulate expected peaks as idealized signal positions (Schneider et al. 2013; Gradmann et al. 2012) and to generate the spectra from the signal positions for some special cases (Matsuki et al. 2007). In particular, software has been published to generate 1D ^{1}H NMR spectra for small molecules (Binev and Aires-De-Sousa 2004; Golotvin et al. 2007; Advanced Chemistry Development, Inc. (ACD/Labs), NMR predictors 2007) and 2D ^{1}H–^{1}H NOESY spectra for biomolecules (Gronwald and Kalbitzer 2004; Donne et al. 1995; Zhu and Reid 1995; Allard et al. 1997) and more general types of spectra as part of larger NMR data visualization software packages (Goddard; Delaglio et al. 1995; Vranken et al. 2005; Stevens et al. 2011). If the resonance line width is large compared to the resonance dispersion, some signals will overlap and fewer peaks compared to the number of signals will be observed. Therefore, the observed (merged) peak positions needs to be derived by performing peak picking on the generated spectra.

Here we present a stand-alone software, VirtualSpectrum, which combines the two procedures, bypassing the need for spectrum visualization software and peak picking, by generating an in-memory representation of the simulated spectrum and deriving the observed peak positions from this representation. VirtualSpectrum produces peaks by defining an analytic expression for the spectrum as a sum of Gaussian shape densities centered at the expected individual signal positions based on an underlying structural model. By definition here, a peak is observed for each local maximum above a certain threshold in the space of the virtual spectrum. If signals overlap on the scale of the line width, fewer maxima (peaks) than signals will be observed. VirtualSpectrum is a general tool applicable for most multi-dimensional NMR experiments, both in the liquid and the solid state, and produces assigned peak lists, with a quality representative of observed peaks. A representative quality means here that the peak lists not only contains peaks in the expected positions but also includes data artifacts, such as a tunable number of noise peaks, perturbed peak positions, and an option to include missing peaks in ssNMR experiments, originating from flexible regions of a protein. VirtualSpectrum is available at http://nmr.au.dk/software/.

Here we demonstrate the applicability of VirtualSpectrum by analyzing a few case studies for which overlap between peaks is particularly relevant. We study the grouping of aligned peaks into spin systems in ssNMR and quantify the amount of overlapping signals in different multi-dimensional ssNMR spectra. Understanding how different parameters effects the extent of overlap between peaks is important for designing and choosing between different NMR experiments. Here we analyze systematically the amount of overlapping peaks at different line widths and for different protein sizes and secondary structures in homonuclear ^{1}H and ^{15}N-edited spectra, and 2D/3D heteronuclear correlation spectra of intrinsically disordered proteins. Finally, we judge the impact of spectral resolution in structural studies by the use of simulated spectra and Cyana (Herrmann et al. 2002) structure calculations.

Methods

VirtualSpectrum calculates simulated peak lists based on provided resonance assignments and protein primary sequence. Most common NMR experiments can be simulated, and even more can be implemented, by providing the definition of the nuclei and residue positions involved in the experiment. Both through-bond and through-space transfers are implemented. For through-space experiments a structural model must be provided. If a X-ray structure, with B-factors, is provided it is possible to model signal attenuation due to mobility, in through-space ssNMR experiments.

Calculating the model signals

First a peak list is inputted to derive the observed peaks. Assume that the resonances of interest are assigned:

$$\Omega = (\omega_1, \omega_1, \ldots, \omega_N) \quad (1)$$

A model peak list is a set of peaks, p, having a vector coordinate and a height, i.e. satisfying:

$$p = (\omega, h) \in \Omega_{\mathrm{n}} \times \mathbb{R}^{+}, h = m(\mathbf{d}), \Omega \xrightarrow{m} \mathbb{R}^{+} \quad (2)$$

where p is a peak, which is defined by its height, h, and frequency (position), ω, and m is an underlying model, which takes the distance, d, as input, for the NMR experiment. The implementation here for the underlying model

is described in Eq. 3 below. The experiment type defines which atoms are connected in the experiment giving rise to cross peaks, which are observed at values determined by the resonance assignments of the corresponding atoms. Most common experiments can be simulated or can be defined by the user through definition of each transfer step. The procedure requires, for each axis of the peak coordinates, a definition of the *atom type* i.e. ^{1}H, ^{15}N, or ^{13}C, *residue order* 0 or −1 (="None" for through-space), denoting atoms for residues *i* and *i*-1, respectively, and in some cases specific atom (atom name) such as C′, Cα or Cβ. For example for NCACX, atom type = (^{15}N, ^{13}C, ^{13}C), residue order = (0, 0, 0), atom name = (N, Cα, None), where "None" denotes a non-specific atom.

VirtualSpectrum implements a general low-level theory model to calculate the height based on the distance between the atoms involved. The model signals can also be provided by third-party software such as e.g. CORMA (Keepers and James 1984) or IRMA2 (Boelens et al. 1988). The model height here is proportional to each of the transfer efficiencies of each magnetization step, f_i. For through-bond experiments this number is set to 1.0 and for through-space experiments a more advanced expression is used (Eq. 4). As implemented here, the model function, *m*, is defined for a set of atoms (*i*, *i* + *1*,…, *n*) assuming the distances, $\mathbf{d} = (d_{\mathrm{i,i+1}}, d_{\mathrm{i+1,i+2}}, \ldots, d_{\mathrm{n-1,n}})$, between the atoms involved in the individual transfer steps and optionally the atomic B-factors, $\mathbf{B} = (B_{\mathrm{i}}, B_{\mathrm{i+1}}, \ldots, B_{\mathrm{n}})$, obtainable from an X-ray structure are known if these are present in the input pdb file.

$$m(\mathbf{d},\mathbf{B}) = \prod_{i=1}^{n-1} f_i(d_{i,i+1}) \prod_{i=1}^{n} g_i(B_i) \tag{3}$$

The B-factors are used to model local dynamics influence on intensities as described in Eq. 6. For through-space transfer steps we use a phenomenological expressions for the intensity:

$$\begin{aligned} f_i(d) &= F_{ij} e^{-b_i}(1 - e^{-c_i d^{-6}}), b_i, c_i > 0 \\ &\cong \begin{cases} F_{ij} e^{-b_i} c_i d^{-6} & for \quad c_i d^{-6} \ll 1 \\ F_{ij} e^{-b_i} & for \quad c_i d^{-6} \gg 1 \end{cases} \end{aligned} \tag{4}$$

This is a two-parameter buildup-type model, which can be interpreted as a spin diffusion process defined by two constants (Macura and Ernst 1980); the correlation rate between two atoms multiplied by the mixing time, *c*, and a leakage rate to the surroundings multiplied by the mixing time, *b*. Thus, experiments with longer mixing times will have larger values of *b* and *c* (proportional to the mixing time in theory). In practice a cut-off distance, d_{max}, is used for which intensities corresponding to larger distances are set to zero.

The distance dependent expression is multiplied with a constant, F_{ij}, to model differences in transfer efficiency, which can account for properties involving other nuclei than the considered spin pair, *i* and *j*, such as spin diffusion and orientation dependence in solid state NMR. The randomized multiplier is defined as:

$$F_{ij} = F_0 L_{ij}, L_{ij} \sim logN(1, \sigma) \tag{5}$$

where L_{ij} is a random number drawn from a log-normal distribution with scaling parameter σ for each atom. F_0 is a constant, which is set to 1.0 for liquid state NMR. For ssNMR, F_0 is largest for intra-residue atom pairs and smaller for inter-residue contacts to model the influence of spin-diffusion in agreement with experimental data.

The g_{i}s are atomic scaling factors modeling the uneven excitation of spins in NMR experiments, which are very pronounced in ssNMR spectra.

$$g_i(B) = G_i \left(\frac{B_0}{B}\right)^{p_i} \tag{6}$$

B_0 and p_i are positive constants, near-zero values of *p* leads to decreasing significance of the atomic scaling factor (g_{i}). G_{i} is a constant.

The noise-level of a spectrum is modeled by rejecting every model peak with a height, *h*, below a specified minimum intensity threshold, h_{min}.

Deriving the observed peaks

Based on the model signals, an analytic expression is derived to represent the full spectrum at any chemical shift coordinate, the "virtual spectrum". The intensity, ρ, in the virtual spectrum is calculated as a sum of Gaussian curves, ρ_{i}, centered at the model signal positions, $\boldsymbol{\omega}_{\mathbf{i}}$, and proportional to the calculated model height, h_i:

$$\begin{aligned} \rho(\mathbf{x}) &= \sum_i \rho_i(\mathbf{x}), \\ \rho_i(\mathbf{x}) &= h_i \prod_{j=1}^{n} exp\left(-\frac{1}{2}\left(\frac{x_j - \omega_{j,i}}{\Gamma_{\mathrm{j}}}\right)^2\right) \end{aligned} \tag{7a}$$

where Γ_{j} is the *Gaussian line width* of the resonances in the jth dimension. The theoretical observed peak positions are the local maxima, $\mathbf{c}_0$, of ρ. The corresponding peak heights are defined as $h = \rho(\mathbf{c}_0)$. It is also possible to use a Lorentzian peak shape:

$$\rho_i(\mathbf{x}) = h_i \prod_{j=1}^{n} 1 \Big/ \left(1 + \left(\frac{x_j - \omega_{j,i}}{\Gamma_{\mathrm{j}}}\right)^2\right) \tag{7b}$$

In practice, the local maxima, $\mathbf{c}_0$, are found by performing a numerical optimization (Nocedal and Wright 2000) as implemented in the scipy python module (Oliphant 2007) (http://www.scipy.org/). The function

optimization is initialized, as parallel computations, with a starting guess equal to the position of all model signals one by one. Optimizations with different starting guesses near the same local maximum (as would be the case for two overlapping signals) yield almost equal solutions. The slightly different solutions are easily combined into observed peak positions using a clustering procedure. In practice, only a subset of the signals in Eq. 7a, 7b, which are near the starting guess, is included in the function optimization.

The observed peak position, **c**, is obtained after adding an error, e, to the initial position, $\mathbf{c}_0$:

$$\mathbf{c} = \mathbf{c}_0 + \mathbf{e}, \mathbf{e} = (e_1(h, \Gamma_1), \ldots, e_n(h, \Gamma_n)) \quad (8)$$

$$e_i(h, \Gamma_i) = z\left(e_{min} + k\Gamma_i\left(e_0 + 1/\max\left(m_0, \ln\left(\frac{h}{h_0}\right)\right)\right)\right) \quad (9)$$

where z is a random number drawn from the standard normal distribution and e_{min}, e_0, k, m_0 and h_0 are constants that are set by the user. The values used in the demonstration of the software are discussed in further detail below. The logarithmic expression leads to low-intensity peaks having larger errors added to the peak position. Both sources of error proportional to the line-width of the resonances and independent of the line-width are present in the expression describing the error. To illustrate the difficulties in finding an accurate position for overlapped signals, e_0 was set to 0 if only one single model peak had a significant contribution to the density in the virtual spectrum at the observed peak position (defined as the peak is *resolved*, see Eq. 11), and 0.3 alternatively.

The random numbers, z above and L_{ij}, in Eq. 5 imply that the generation of peak list can be repeated with different results in each new run. Furthermore, a small random number $l_0 \sim N(0, s_l)$ is added to each coordinate of the model signal positions, $\boldsymbol{\omega_i}$, before calculation of the virtual spectrum, to allow for more representative differences between observed symmetry-related peaks on either side of the diagonal. This procedure, indirectly, also models differences in the line-widths in the direct and indirect dimensions.

Noise peaks can be added to the peak list. The position, **c**, of the noise peak is derived by drawing a random position, $\boldsymbol{\omega}$, from the set of resonances $\boldsymbol{\Omega_n}$ (Eq. 1) and adding a random small shift $\boldsymbol{\varepsilon}$.

$$c = \omega + \epsilon, \epsilon \sim N(0, \Gamma) \quad (10)$$

The random additive shift can be drawn from a normal distribution with a scale parameter equal to the line width as above. The procedure above produces only noise peaks that align approximately with other observed peaks, as these are the only peaks that would interfere with a good assignment of the spectra.

The artificial data generated by VirtualSpectrum can be customized to test the impact of different quality parameters in subsequent tasks such as resonance assignments and structure calculation. The sensitivity can be increased by lowering the noise threshold, h_{min}, or decreasing the line width, Γ, of the resonances (Eq. 7a, 7b) leading to more peaks being picked due to decreasing overlap in the spectra. The data can be made noisier by adding more noise peaks. Changing the constants in Eq. 9 (e.g. decreasing e_{min}, e_0 or k) can increase the accuracy in the pick positions. Also, the excitation heterogeneity can be increased by choosing larger value of the powers p_i in Eq. 6.

Parameters used and a guide for choosing parameters for the simulation of NMR peak list with VirtualSpectrum

In the applications of VirtualSpectrum discussed in this publication, several parameters were used to produce simulated data as close to the experimental data as possible. Some parameters had fixed values, whereas some had different values depending on the systems studied, and finally, some were varied systematically. The parameter (p_i) used to simulate the excitation heterogeneity (Eq. 6) (only used for ssNMR data) used $p_i = 1.4$ and 0.35 for the first and subsequent transfers, respectively. Higher values mean a larger spread in the simulated excitation heterogeneity, for most cases values between 0 and 2 are good choices. B_0 is a reference B-factor and $B_0 = 10$ was used here and is the default. This value corresponds roughly to the average B-factor in crystal structure for the proteins studied here, but for other systems a different reference B-factor might be more appropriate. G_i (Eq. 6) was set to unity in all cases, except for ssNMR where $G_i = 0.8$ for aromatic carbons was used to model the attenuated intensities for those spins. With the current implementation of VirtualSpectrum it is only possible to choose non-unity values for aromatic carbons and carbonyl.

For the scale parameter, σ (Eq. 5), which controls the random number generation for each atom pair, $\sigma = 0.5$ was used, large values lead to a larger spread in the dynamic range of intensities. The intensity accounting for transfer efficiencies was set to $F_0 = 1.0$ (Eq. 5) for liquid state NMR. $F_0 = 0.1$ and 0.033 for ssNMR for sequential (adjacent residues) and medium range/long range (residue number difference > 3) transfers, respectively, and 1.2 for intra-residue. The higher value used for intra-residue cross peaks was chosen to model the observation in the ssNMR data presented here that, for the short mixing times used here, intra-residue correlations were highly abundant due to

efficient spin diffusion. For longer mixing times, larger values $0.5 < F_0 < 1.0$ should be used.

For all ssNMR and liquid state data, $s_l = 0.05$ ppm and 0.001 ppm, respectively, were used to simulate noise in the peak positions. If one wishes to simulate peak lists for cases where experimental data is available, this parameter can be estimated from the variation in peak positions along a peak strip corresponding to the same isolated resonance. The noise constants in Eq. 9: $h_0 = 0.75$, $k = 0.333$, $m_0 = 0.8$, and $e_{min} = 0.1$ for 3D homo-nuclear carbon, $e_{min} = 0.002$ for 3D liquid state experiments, otherwise $e_{min} = 0$ was used. For resolved signals (Eq. 11), $e_0 = 0.0$ (Eq. 9) was used whereas $e_0 = 0.3$ was used for non-resolved peaks to model greater uncertainty in picking overlapped peaks correctly. We argue that these are appropriate default values; choosing larger values for e_0 (or smaller values for e_{min}) will produce noisier data.

Through-space transfer (Eq. 4) were simulated with constants defined with the help of a *maximum intensity*, $I_0 = e^{-b}$, $0 < I_0 < 1.0$, and a *characteristic distance*, r_{min}, where a peak usually would be observed in 50 % of the cases when $h_{min} = 0.1$, where h_{min} is the noise level above which a peak is defined as observable. The effect of different choices of r_{min} is visualized in Figure S1. Values of r_{min} close to 5 Å should be used to model long mixing times together with lower values for I_0. Smaller values for r_{min} will lead to less intense peaks and a larger dynamic range in the intensities corresponding to peaks for short and longer distances. Since the height at the maximum of the strongest peaks in the spectrum is close to 1.0, h_{min} corresponds to the inverse of the signal to noise ratio, $SN = h_{min}^{-1}$. The following relationship was used to calculate c: $c = 0.1/I_0 * r_{min}^6$. All NOESY type spectra used, $r_{min} = 4.0$ Å, $I_0 = 0.4$, $SN = 75$. For the ssNMR data $r_{min} = 4.0$, 3.8, 3.3 Å, and $I_0 = 0.4$, 0.9, 0.9 were used for 2D ^{13}C–^{13}C through-space correlation (DARR-type), 3D NCACX, and NCOCX, respectively. The signal to noise ratios used were $SN = 15$, 12, 12, 8, and 25 for NCACO, NCACX, NCOCX, CONCA, and DARR-type, respectively, and $SN = 75$ for NOESY type spectra. Only the aliphatic side chain resonances were included in the NCACX and NCOCX spectra. The resonance line width (Eq. 7a, 7b) was varied systematically for all types of spectra and the particular values are described in the text and the related figure legends.

Results and discussion

VirtualSpectrum was applied to simulate peak lists for several types of multidimensional NMR spectra for different proteins. The simulated spectra, used to generate the simulated peak lists, were compared to the experimental spectra for some of the proteins to test the procedure. We have chosen to focus on three cases where overlap between signals are particularly significant: (1) solid state NMR (ssNMR) with ^{13}C-detection having relatively large ^{13}C and ^{15}N line widths, (2) homonuclear ^{1}H and ^{15}N-edited spectra for which overlap is especially significant in 2D ^{1}H–^{1}H NOESY spectra, (3) unstructured/intrinsically disordered proteins (IDPs) where the reduced chemical shift dispersion due to lack of regular structure leads to increased overlap of signals. Peak lists for different multidimensional spectra were simulated for the three proteins: the immunoglobulin binding domain B1 of streptococcal protein G (GB1) (Bouvignies et al. 2006; Gallagher et al. 1994), Ubiquitin (Zech et al. 2005; Igumenova et al. 2004; Vijaykumar et al. 1987), and Hen Egg White Lysozyme (Blake et al. 1965; Redfield and Dobson 1988) (HEWL). A subset of the spectra was also simulated for the intrinsically disordered cytoplasmic domain of human neuroligin-3 (hNL3-Cyt) (Wood et al. 2012; Paz et al. 2008). Protein structures used to model the through-space transfers were downloaded from the RCSB Protein Data Bank (PDB; http://www.rcsb.org/pdb/) (Berman et al. 2000) and chemical shifts assignments were retrieved from the BioMagResBank (BMRB; www.bmrb.wisc.edu) (Ulrich et al. 2008). The aim here was not to exhaustively characterize all possible aspects of the three case studies, but to provide proof-of-concept demonstration of the applicability of our software. We chose here to focus on analyzing the number of resolved signals, comparing various conditions and comparing between three different proteins and studying the alignment of peaks as well for ssNMR data. To exploit larger ranges of protein structure, virtual peak list was also generated for a set of invented protein sequences to analyze the degree of overlap. Finally, the impact of different spectral quality parameters on the accuracy in a protein structure calculation were tested, by performing structure calculations using the program Cyana (Herrmann et al. 2002), based on simulated peak lists.

Solid state NMR

Despite the challenges of solid-state protein NMR, increasingly more protein structures are being published (Castellani et al. 2002; Lange et al. 2005; Manolikas et al. 2008; Zech et al. 2005; Loquet et al. 2008; Thiriot et al. 2004). In solid state NMR, line widths are generally much larger due to proton dipolar couplings and/or structural heterogeneity, particular observed for fibril structures, often studied by ssNMR (Tycko 2006; Naito and Kawamura 2007). Another issue is the reduced sensitivity in ssNMR, compared to liquid state NMR that limits the number of dimensions possible in an experiment. Typical ssNMR experiments are 2D or 3D experiments correlating

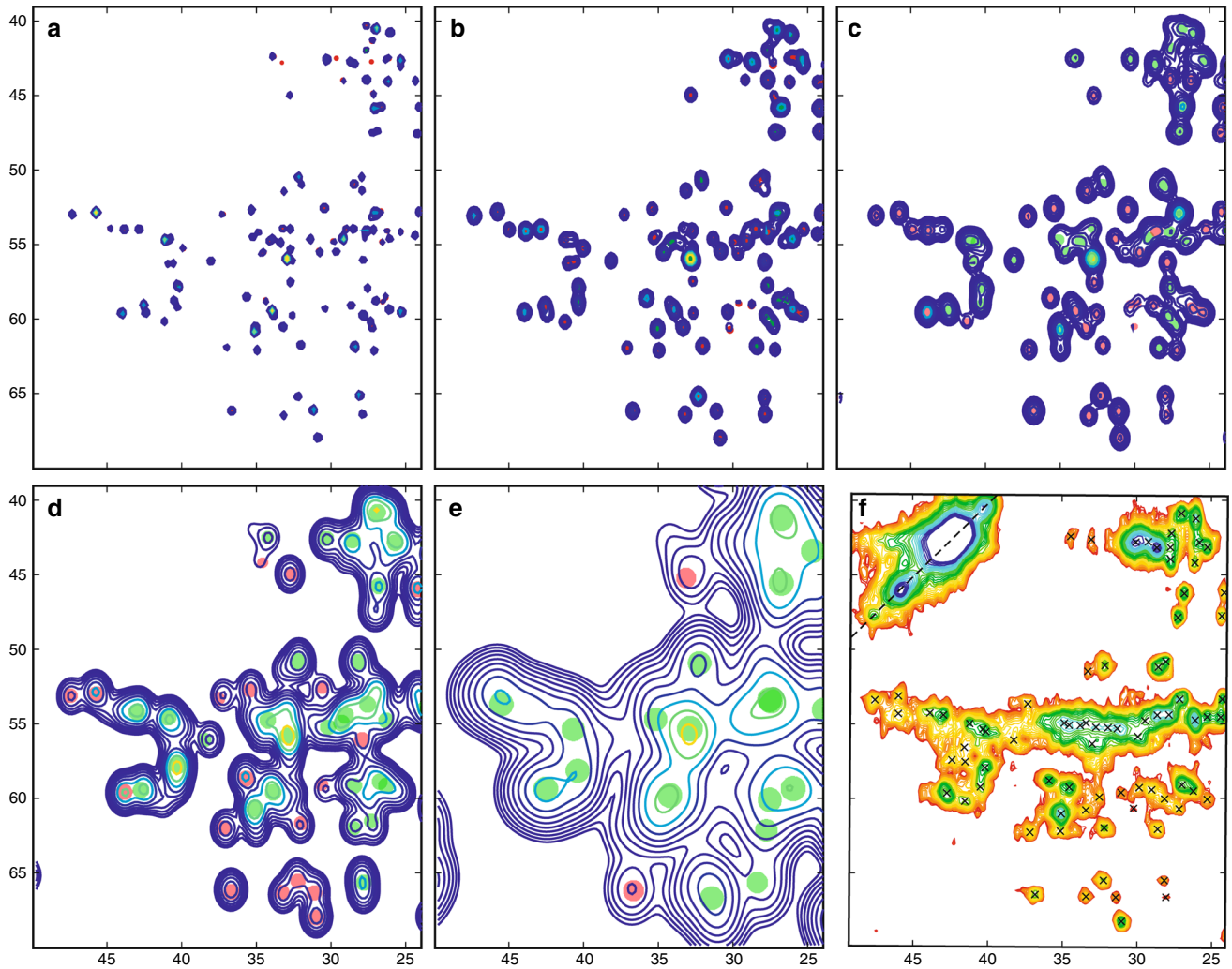

Fig. 1 Comparison of simulated and experimental spectra of Ubiquitin. **a**–**e** Excerpt showing a crowded region of 2D ^{13}C–^{13}C through space correlation spectra simulated by VirtualSpectrum, the intensity is depicted as a contour plot. The Gaussian line width, Γ (Eq. 7a, 7b), of the peaks in the direct dimension is 0.1, 0.2, 0.3, 0.5 and 1.0 ppm for **a**–**e** corresponding to a full width at half maximum height of 2.355Γ ppm. In the indirect dimension the above-mentioned line width was multiplied with 1.3. All other parameters are as described in the "Methods" section. The chemical shifts from BMRB id 7111 and the structure from pdb id 1ubq was used. The diagonal signals were not simulated. Overlapped and resolved peaks (Eq. 11) are highlighted with *green* and *red disks*, respectively. **f** Experimental 2D ^{13}C–^{13}C DARR spectrum (Takegoshi et al. 2001) with picked peaks used for quantitative comparison of the spectra shown as *black crosses*, the diagonal signal is highlighted with a *dashed line*. The spectrum was acquired with a mixing time of 20 ms. The sample consists of hydrated microcrystals recorded using a Bruker 700 MHz Avance II spectrometer at 12 kHz spinning using standard 4 mm double resonance Bruker probe. 80–100 kHz SPINAL-64 ^{1}H decoupling was applied during direct and indirect acquisition periods, and acquisition times of around 30 ms were used. Acquired with 400 points and 200 ppm spectral widths in the indirect dimensions using 88 scans

backbone or side chain ^{13}C and ^{15}N chemical shifts. Figure 1 shows an excerpt of an ssNMR 2D ^{13}C–^{13}C Dipolar Assisted Rotational Resonance (DARR) (Takegoshi et al. 2001) spectrum of Ubiquitin, containing cross peaks for carbons close in space, shown together with the simulated spectrum with the picked peak positions indicated. It illustrates that with increasing line width, the signals merge, hence the number of observable cross peaks decreases. The simulated spectrum with a Gaussian line width of 0.3 ppm, corresponding to a full width at half maximum height of 0.707 ppm, resembles the experimental spectrum reasonably well. To be more specific, 79.2 % of the peaks in the simulated spectrum (SS) can be refound in the observed spectrum (OS) whereas 79.1 % of the peaks in the OS, were present as well in the SS. Matches were allowed within 0.6 ppm, the most frequent difference in the peak positions is 0.0745 ppm (highest column in histogram) (see Figure S2a in Supporting Material). The main difference between observed and simulated spectra is in the peak intensities, where the experimental data has a

complex dependency on the local structure and dynamics of the proteins. The correlation coefficient for the peak height for the matched peaks is 0.647 (see Figure S2b in Supporting Material). Another difference in the appearance is that noise is observed in the experimental spectrum; whereas the noise is implicitly accounted for in the simulated spectra [see Eqs. (8)–(10)]. We emphasize that our aim with VirtualSpectrum is not to produce peak lists that are completely identical to observed data. This will never be the case, since we deliberately include randomized errors in our implementation both for the intensity (Eq. 5) and for the peak positions (Eq. 9 and through the parameter s_l described in the "Methods"). These errors were introduced to emulate, in a stochastic manner, the complex features in real spectra with deviations from a simple single-distance dependence (Eq. 4) for the intensity and distortions in peak positions due to noise.

It is evident that increasingly fewer peaks are observed, due to signal overlap, as the line width increases. For the analysis presented here, we define a peak at position, $\mathbf{c}_i$, with local density ρ_i (Eq. 7a, 7b) to be *resolved* if for all other signals ρ_j:

$$\rho_j(\mathbf{c_i}) < 0.1\rho_i(\mathbf{c_i}) = 0.1h_i \tag{11}$$

VirtualSpectrum was applied to simulate peak lists for various common 2D and 3D ssNMR correlation spectra for different proteins while incrementing the simulated resonance line widths. (see Fig. 2). We compare the overlap in the spectra by analyzing the fraction of observed picked peaks, which are resolved, f_{res}, for a certain line width:

$$f_{res}(\Gamma) = N_{peak}(\Gamma)/N_{peak}(\Gamma_0) \tag{12}$$

where $N_{peak}(\Gamma)$ is the number of peaks in the peak list produced with Gaussian line width, Γ, and Γ_0 is the smallest line width analyzed. This fraction was plotted as a function of the Gaussian line width, ranging from 0.1 to 1.0 ppm, shown in Fig. 2 top. Note that a Gaussian line width, Γ, corresponds to a full width at half maximum height of 2.355Γ ppm. Several trends appear in accordance with expectations. Firstly, overlap is more pronounced in the 2D spectrum compared to the 3D spectra as expected. For example, for GB1, the fraction of observed peaks for a Gaussian line width of 0.3 ppm is $f_{\text{res}}(0.3) = 0.67$ and 0.93 for 2D DARR and 3D NCACO, respectively. Secondly, it is notable that most peaks are resolved among the 3D spectra in the NCACX and, in general, fewest in the NCACO/CONCA. This is consistent with the larger dispersion found for Cα compared to C′ and the large dispersion found in the side chain resonances ("CX") present in NCACX/NCOCX spectra. This suggests that C′ resonances might be a bottleneck for ssNMR resonance assignments. Finally, it is seen here also that larger proteins have more overlapping signals as a smaller fraction of peaks is resolved with increasing protein sizes through GB1 < Ubiquitin < HEWL. For example for a Gaussian line width of 0.3 ppm for the NCACO, the fraction of observed peaks are $f_{\text{res}}(0.35) = 0.91$, 0.68, and 0.66, for GB1, Ubiquitin, and HEWL, respectively. Our analysis can also be applied to make simple queries, e.g. if one has a protein of the size of GB1 and wish to be able to acquire an NCACX spectrum with at least 90 % peaks resolved then, based on the curve in Fig. 2, this would only be possible if a sample can be prepared with a Gaussian line width of maximum 0.3 ppm.

We have used the peak lists generated by VirtualSpectrum to analyze the difficulty of grouping peaks, from ssNMR spectra, into proper residue spin systems. The position of each peak is subject to distortions in the position both due to noise present in the spectrum, as modeled here by Eq. 9, but also owing to overlap with other signals. This leads to a variation in the peak position of the same nuclei within the groups of peaks belonging to the same residue. Here this variation was quantified by grouping all ssNMR peaks found in the 3D peak lists generated by VirtualSpectrum (the experiments analyzed in Fig. 2) according to the residue number of the N backbone atom. If a peak was a result of merging of more than one signal, the ^{15}N chemical shift of the peak was included in the analysis for all the corresponding residues. The standard deviation of all peak coordinates with ^{15}N shifts within the same residue peak group (spin system) was calculated and the average, $\langle\sigma\rangle$, of this standard deviation was calculated among all such peak groups for the protein. This quantity can be considered as the typical alignment error when trying to connect peaks to generate spin systems.

This alignment error is shown for the three different proteins, as a function of the resonance line width (Fig. 3). Unsurprisingly, the alignment error increased with the resonance (Gaussian) line width, Γ. The smallest resonance line widths, usually attainable in ssNMR, $0.2 < \Gamma < 0.5$ ppm, for GB1 and Ubiquitin, the alignment error is smaller than Γ, meaning that the alignment error is dominated by errors in the peak position, arising from noise in the spectra. HEWL is a larger protein; with more overlapping signals for relatively small line widths (see Fig. 2). In contrast to the trend for GB1 and Ubiquitin discussed above, for HEWL (and for GB1 and Ubiquitin for larger line widths) the average alignment error becomes approximately equal to the Gaussian line width. This indicates that the alignment error is predominantly caused by distortions in peak positions, due to overlap between signals in the spectra. Our findings here are based on the assumption that Eq. 9 can describe the error in the peak position. We argue that this is a sound description of the error although systematic noise such as baseline distortions are not included in the model and, the constants used in the

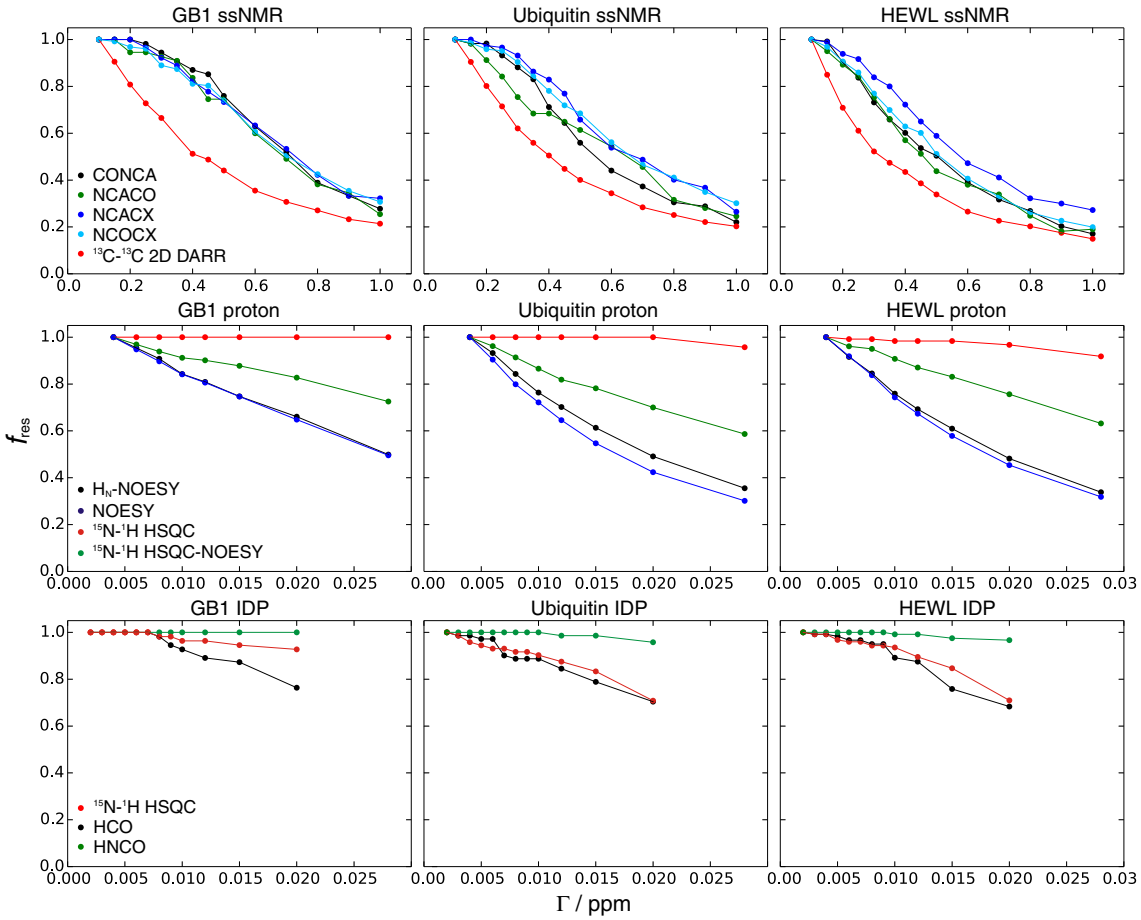

Fig. 2 Signal overlap in NMR peak lists derived by VirtualSpectrum. *Line plot* of the fraction of observed picked resolved peaks, f_{res} (Eq. 12), as a function of the Gaussian line width, Γ (Eq. 7a, 7b), for ^{1}H (liquid state NMR) and ^{13}C (ssNMR) in the direct dimension, shown for GB1 (*left*), Ubiquitin (*middle*), and Hen Egg White Lysozyme (HEWL, *right*). Different experiments are distinguished with different colors (see annotations). Data is shown for solid state NMR (labeled "ssNMR", *top*), ^{1}H and ^{15}N-edited liquid state NMR (proton, *middle*), and using chemical shift data (see text) for unstructured/intrinsically disordered proteins (IDP, *bottom*). For the indirect dimension, a line width, $\Gamma_{ind} = k_{ind}\Gamma$, proportional to the line width in the direct dimension was used; for the ssNMR data $k_{ind} = 2.0$ was used for both ^{13}C and ^{15}N. For the liquid state NMR experiments we used: $k_{ind} = 1.5$ for NOESY and H_N-NOESY, $k_{ind} = 2.0$ and 10.0 for ^{1}H and ^{15}N, respectively, in ^{15}N-HSQC-NOESY. $k_{ind} = 6.0$ was used for ^{15}N in both ^{15}N-HSQC and HNCO whereas $k_{ind} = 2.0$ was used for ^{13}C in both HCO and HNCO. These line width were used in all spectra of the same type described both in the text and in other figures. The protein structures used to generate the through-space transfer intensities were from pdb ids, 2igd, 1ubq and 1vdq for GB1, Ubiquitin and HEWL, respectively. The used chemical shifts were from BMRB ids; for ssNMR: 17810, 7111 and 4831(backbone) + 4563(side chain) for GB1, Ubiquitin and HEWL, respectively, for proton NMR: 7280, 5387 and 4831(C′/N) + 4562(H_N) for GB1, Ubiquitin and HEWL, respectively, and for unstructured proteins: 16627, 16626 and 18365 for GB1, Ubiquitin and HEWL, respectively

equation are heuristic parameters, which can be set by the user. This analysis underscores the difficulty in generating correct residue spin systems from ssNMR peaks, in particular for larger proteins, since there would be a risk of incorrectly aligning peaks not belonging to the same residue. Hence, this would lead to the formation of wrong spin systems, which are supposed to form the basis for the resonance assignments.

We have analyzed the effect of increasing line width, and of using different NMR experiments, on the ssNMR resonance assignment of GB1 using VirtualSpectrum and our software, GAMES_ASSIGN (for automatic assignment of resonances), in a parallel study (Nielsen et al. 2014). In that study we show that resonance assignments of GB1 are reliable, using the 2D and 3D experiments described above, for a Gaussian line width up to 0.5 ppm. Beyond this line width the assignment errors increase rapidly. We also demonstrate, by expanding the data set to include other simulated spectra, that the success of the assignment can be improved by including more experiments. Our analysis also showed that experiments, such as N(CO)CACX and CAN(CO)CX involving Cα were the most successful, but

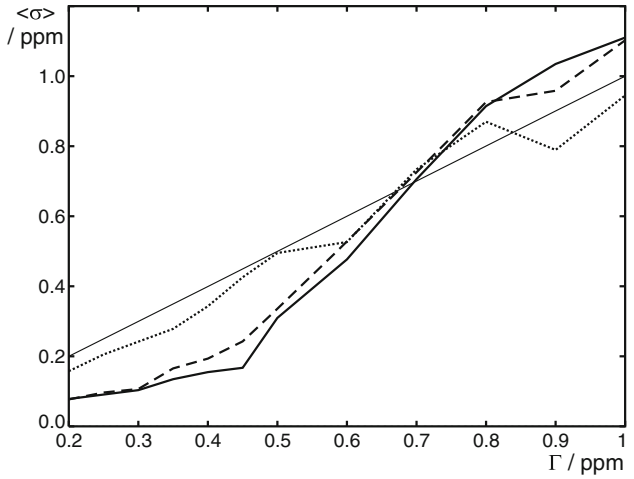

Fig. 3 Average alignment error (see text) when comparing the ^{15}N chemical shift for the same residue in spectra of NCACX, NCOCX, NCACO and CONCA (see Fig. 2), shown as a function of the Gaussian line width, Γ (Eq. 7a, 7b), for GB1 (*full line*), Ubiquitin (*dashed line*), and HEWL (*dotted line*). The line y = x is shown as a thin line for reference

4D experiments were even more successful. This illustrates how VirtualSpectrum is useful when evaluating and comparing the impact of using different experiments, without the need to acquire experimental data.

Liquid state NMR with natural abundance

^{1}H homonuclear liquid state NMR has been used for decades to study relatively small proteins (Wüthrich 1986; Williamson et al. 1985; Driscoll et al. 1989; Simorre et al. 1991; Wittekind et al. 1992; Bouaziz et al. 1992; Breg et al. 1995). Nowadays, most proteins studied by NMR utilize ^{15}N and ^{13}C isotope labeling through bacteria expression or peptide synthesis (Kigawa et al. 1999; Marley et al. 2001; Kainosho et al. 2006; Goto and Kay 2000) to overcome problems related to overlapping signals and facilitate resonance assignments. In cases where such labeling techniques would be problematic, or if one would prefer a less expensive solution, it is often possible, though more challenging, to study the protein using the classical natural abundance approach. For this approach there is a limitation in protein size both due to increasing signal overlap and increasing line width due to relaxation.

NOESY is the fundamental NMR experiment for deriving the structure of a protein with natural abundance NMR (Jeener et al. 1979; Kumar et al. 1980). The NOESY spectrum of HEWL was simulated using VirtualSpectrum. The overall appearance of the simulated spectrum is rather similar to the observed (see Fig. 4), i.e. the signal to noise ratio in the spectra and the resolution in the spectra are comparable. There are differences in the peak positions between the spectra, we argue that these are primarily due to small differences between the chemical shift used for simulating the spectra (bmrb ID 4563) and the actual chemical shift of our sample, under the exact conditions used here for the experimental NOESY spectrum. Using the same quantitative comparison as described for the DARR spectrum, here 60.4 % and 61.5 % of the peaks can be found within 0.05 ppm, when searching for the simulated peaks in the observed spectrum and vice versa, respectively. Again, as with the ssNMR through-space transfer spectrum, DARR, the most notable difference is in the intensities of peaks. In this case the correlation coefficient is 0.273, when comparing data points in the spectra as described above for DARR. However, the correlation coefficient increases to 0.469 when not comparing the data points at the exact same coordinate but rather comparing pairs of matched peak heights at the individual peak maxima (data not shown). VirtualSpectrum uses a simplistic phenomenological approach to derive the intensities (Eq. 4) for through-space transfers of magnetization, neglecting third spin interactions, such as spin-diffusion and, hence, differences in the intensities are to be expected. It is possible to use signal positions and intensities simulated by other tools as input for VirtualSpectrum, but this option was not demonstrated here.

The amount of overlap between signals was analyzed quantitatively, measuring f_{res} (Eq. 11), by simulating spectra using VirtualSpectrum with different (Gaussian) line widths ranging between 0.004 and 0.028 ppm. Fewer peaks are observed in the NOESY spectrum with larger line widths and the amount of overlap increases with the size of the protein (Fig. 2), e.g. f_{res} decreases to 0.50 for GB1 and ca. 0.3 for Ubiquitin and HEWL. In this analysis, degenerate assignments of non-identical methylene protons were not considered as overlap (but regarded as one single signal), whereas distinct assignments for methylene protons were normally treated with the possibility of overlap between signals. The overlapping signals can be partly resolved by using ^{15}N isotope labeling and acquiring a 3D ^{15}N–^{1}H-HSQC-NOESY spectrum. This spectrum was simulated with VirtualSpectrum resolving of ca. half of the signals, which overlap in the 2D NOESY. This relationship also holds, approximately, when comparing to the H_N only part of the 2D NOESY, which can be considered as a projection of the 3D spectrum. The 2D ^{15}N–^{1}H HSQC spectrum is another 2D-projection of the 3D spectrum. In the HSQC almost all peaks are still resolved with increasing line width and, counter-intuitively, more peaks overlap in the 3D spectrum. This is because the signals that still overlap in the 3D spectra are mostly those with a common HSQC peak, between two protons of the same methylene group (data not shown).

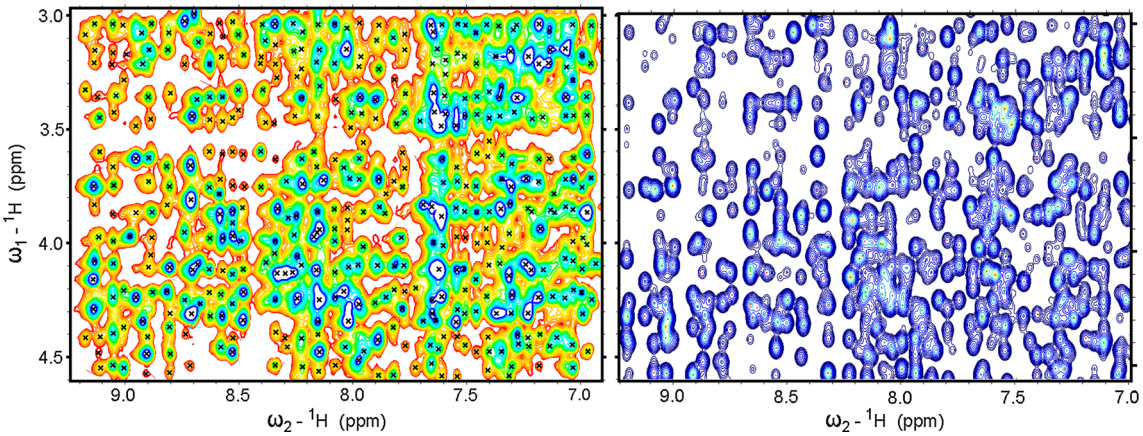

Fig. 4 Comparison of experimental NOESY spectrum (*left*) and NOESY spectrum of HEWL simulated by VirtualSpectrum (*right*). A proton line width of 0.012 ppm and 0.018 ppm in the direct and indirect dimensions, respectively, was used. For all other parameters, see "Methods" and legend to Fig. 2. (*left*) Experimental 600 MHz watergate-NOESY spectrum, showing picked peaks used for quantitative comparison with *black crosses*, acquired with a 500 ms mixing time using 400 × 562 points and a spectral with of 9,000 Hz acquired on a 2 mM sample of HEWL (EC 3.2.1.17; from Sigma) in 90/10 % v/v H2O/D2O, pH 3.5; T = 35 C

Intrinsically disordered proteins

Traditionally it was believed that protein function was dependent on the unique folded tertiary structure, determined by the amino acid sequence (Anfinsen 1973; Bryngelson et al. 1995). However, recent progress in protein analysis has revealed that this is not always true. Some proteins are intrinsically disordered (IDPs) having no well-defined tertiary structure (Wright and Dyson 1999; Dolgikh et al. 1981; Bychkova et al. 1988; Dunker et al. 2001; Dyson and Wright 2005; Tompa 2002; Uversky 2002) or only fold in complex with targets (Dyson and Wright 2002). Liquid state NMR is one of the preferred experimental techniques to study IDPs, since it allows measuring site-specific time-averaged chemical shifts and relaxation parameters providing information on the dynamics and secondary structure propensities (Bertoncini et al. 2005; Jensen et al. 2009; Meier et al. 2008; Mittag and Forman-Kay 2007; Shojania and O'Neil 2006; Sugase et al. 2007). However, due to the scarcity of regular secondary structure, the NMR spectra suffer from poor resonance dispersion leading to severe overlap between signals. The nuclei with the highest dispersion relative to the line width are backbone ^{15}N and ^{13}C′, and therefor often NMR experiments used for studying IDPs are based on these nuclei.

VirtualSpectrum was applied to simulate 2D HCO and ^{1}H–^{15}N HSQC correlation spectra. Simulated and observed spectra, for the intrinsically disordered protein, hNL3-Cyt, are practically superimposable as seen in Fig. 5 (except for minor forms, probably due to slow cis/trans isomerism, present in the experimental data). It should also be noted that the peak shapes for the observed and simulated spectra are almost identical validating the use of a Gaussian lineshape. In some cases a Lorentzian shape or multiplet pattern may be more appropriate, which is also possible to use in VirtualSpectrum.

We have used published chemical shifts for GB1, Ubiquitin and HEWL in a denatured (unstructured) state (Vajpai et al. 2010; Sziegat et al. 2012) as input for VirtualSpectrum to simulate NMR correlation spectra and compare it to the structured state. It is clear that there is much more overlap in the HSQC spectra for the unstructured compared to the structured state. For Ubiquitin, in a simulated 2D ^{1}H–^{15}N HSQC using a Gaussian line width of 0.02 ppm, all signals were resolved for the structured state, whereas only 83 % were observed for the unstructured state (Fig. 2). The dispersion power, i.e. the ability to resolve 2D resonance correlations, for ^{15}N and ^{13}C′ was analyzed quantitatively by using VirtualSpectrum by simulating peak lists for 2D HCO and ^{1}H–^{15}N HSQC and 3D HNCO with increasing resonance line width (see Fig. 2 bottom). In our hands, for the three proteins analyzed, ^{15}N has a slightly better dispersion power than ^{13}C′, as the fraction of observed peaks decreases a little more in the 2D HCO compared to the ^{1}H–^{15}N HSQC spectra when increasing the line widths. By comparison, the 3D HNCO spectrum is much more resolved with almost all signals observed separately for all proteins at the line widths tested. E.g. the minimum fraction of observed peaks, f_{res}, is 0.958 at proton Gaussian line width, $\Gamma = 0.02$ ppm, for 3D HNCO of Ubiquitin compared to $f_{res} = 0.70$ and 0.71 for HCO and HSQC, respectively, at the same line width.

Effect of protein size and secondary structure on spectral overlap

We have studied the overlap in NMR spectra for only three relative small protein structures, with a global fold

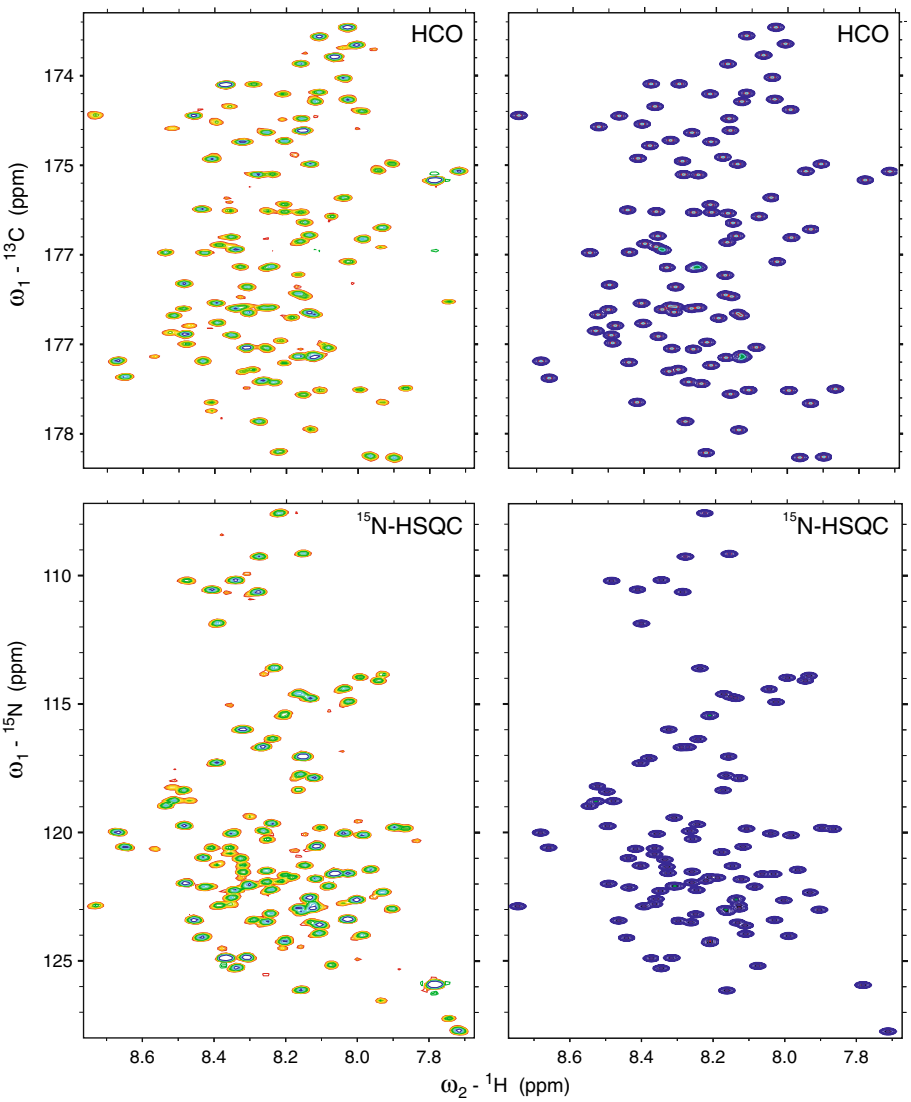

Fig. 5 Experimental spectra (*left*) and spectra simulated by VirtualSpectrum for the intrinsically disordered protein, hNL3-Cyt (Wood et al. 2012; Paz et al. 2008) with chemical shifts from BMRB id 17289. The very low intensity peaks in the experimental spectra (*red* only contours) are due to minor forms. A Gaussian line width of 0.006 ppm in the direct (^{1}H) dimension and 0.012 and 0.036 ppm in the indirect ^{13}C and ^{15}N dimension, respectively. All other parameters for VirtualSpectrum are as described in "Methods"

containing both helixes and beta sheets. We wish to extend the range of these simulations, to more systematically address the effect of protein size and secondary structure on the overlap in NMR spectra. Therefore the ^{15}N–^{1}H 2D HSQC spectra were simulated for a set of invented proteins of different sizes and secondary structures analyzing the fraction of observed peaks (see Fig. 6). The spectra were simulated with a Gaussian line width of 0.015 ppm and 0.09 ppm for ^{1}H and ^{15}N, respectively, and the other parameters were set as described in the legend to Fig. 1 and in the "Methods". The sequences for the invented proteins were generated randomly using statistics on protein sequence, secondary structure and chemical shifts from a library of 681 protein chains (Nielsen et al. 2012). The secondary structure along the sequence was built using occurrences of 32.9, 24.1 and 43.0 % for helix, beta sheet and coil structures, respectively (based on statistics in the library). For sequences with mixed structure, the first 32.9 % residues were set to be helix, the next 24.1 % beta-sheet and the final 43 % coil. For the predominantly helix sequences, the same numbers were used—except that all beta-sheet residues were replaced with helix residues, for the predominantly beta sheet sequences the opposite replacement was done and for the unstructured sequences only coil was used. The sequences were constructed randomly by using the conditional probabilities, for each amino acid for a given secondary structure, from the library. The chemical shifts were drawn from a normal distribution using average and standard deviation for the secondary structure and amino acid specific chemical shifts in the library. The chemical shifts for the unstructured sequences were taken from a library of neighbor corrected sequence-specific random coil chemical shifts of intrinsically disordered proteins (Tamiola et al. 2010) using the

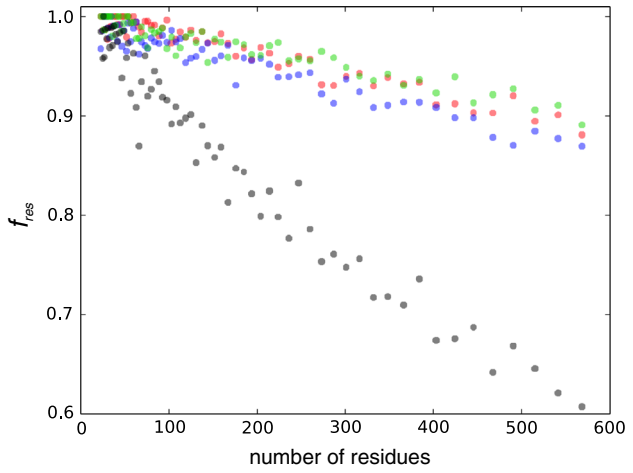

Fig. 6 Fraction of observed peaks in simulated 2D ^{15}N–^{1}H HSQC spectra for invented proteins. The spectra were simulated as described in the main text. The fraction of observed peaks, f_{res}, calculated as the number of observed peaks divided by the number of residues—1—number of prolines in the sequence, is shown as *red*, *green*, and *blue disks* for, mixed secondary structure, predominantly beta sheet and predominantly helix, respectively. The *black disks* show f_{res} in the HSQCs for sequences with no secondary structure. For all displayed data points, three repetitions of the simulation were performed and the averages of f_{res} from the three simulations are shown here

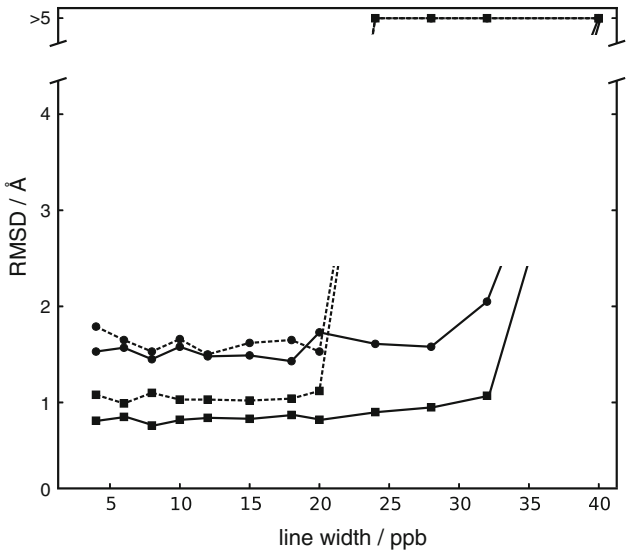

Fig. 7 The impact of line width, signal/noise and noise peaks on protein structure accuracy and precision studied by Cyana and VirtualSpectrum. The diagram shows data for Cyana structure calculations of the 129 amino acids protein HEWL based on 3D ^{15}N-HSQC-NOESY and ^{13}C-HSQC-NOESY spectra simulated by VirtualSpectrum (see also Fig. 2 middle row, right column). *Each symbol* represents a Cyana structure calculation with automatic peak assignments and distance calibration using default Cyana parameters (Herrmann et al. 2002). The outcome of each structure calculation is shown as a function of Γ_H in the direct proton dimension. The accuracy, measured as the heavy atom RMSD against the reference structure (pdb ID = 1vdq), is shown with *filled circles*. The precision, measured as the heavy atom average heavy RMSD to the mean structure within the 20 structure ensemble, is shown with *filled boxes*. *Solid connecting lines* indicate values for a fraction of added noise peaks, f_N (see "Methods" and Eq. 10) of 2 % the total number of peaks and a signal to noise, SN ($\Rightarrow h_{min} = 1/SN$, see "Methods") of 75 whereas *broken lines* indicate values with $f_N = 10$ % and SN = 75. Values above 5 Å are shown truncated to 5 Å, see also Table S1 in the Supporting Material for more simulations

reported rms values of 0.64 ppm and 0.14 ppm for ^{15}N and ^{1}H H_N, respectively, as the standard deviation in the random number generation.

It was found that for mixed secondary structure sequence, the fraction of observed peaks decreases approximate linearly to the number of residues studied (24 to 568), with a slope of ca. 0.0002. For a protein of ca. 200 residues $f_{res} = 0.97$ on average. Note that this means that up to 6 % of the peaks overlap for such a protein (if each overlap in pairs of two). Overlapping peaks in ^{15}N–^{1}H 2D HSQC spectra also implies that the corresponding spin systems will have degenerate, "^{15}N–^{1}H roots", and therefore spin system generation through ^{15}N–^{1}H alignments, is potentially problematic. Furthermore, it is seen that protein sequences with predominantly helix residues (57 % of the residues) have slightly fewer resolved peaks compared to predominantly beta sheet sequences and mixed secondary structure sequences with a similar or larger number of residues (Fig. 6). This is consistent with the larger chemical shifts dispersion found in beta-sheets compared to helices. Spectra were also simulated for sequences of unstructured proteins, with chemical shifts taken from a library of unstructured proteins (Tamiola et al. 2010). It is seen here that more peaks overlap in this case with only 82 % peaks observed on average for a structured protein with ca. 200 residues (Fig. 6). This illustrates the difficulty of assigning unstructured proteins of medium size with conventional spectra, prompting for the use of spectra with more dimensions, e.g. 4D spectra, for assignments.

Impact of spectral resolution and sensitivity on protein structure calculation

Using good quality experimental data is very important for the accuracy of structure determination. This was quantified here by the use of Cyana structure calculations (Herrmann et al. 2002), based on data simulated using VirtualSpectrum. The structure was calculated for the 129 amino acid protein HEWL using simulated data for 3D ^{15}N-HSQC-NOESY and ^{13}C-HSQC-NOESY. The NOESY data were simulated using default Cyana parameters and different input settings: increasing line widths and with differences in sensitivity, as modeled by the signal/noise ratio and number of added noise peaks (see "Methods" and legends to Fig. 7). In addition, peak assignment tolerances were set to 2.0 times the Gaussian line width, Γ, in each dimension. For the indirect proton dimension, a line width of 2 * Γ_H was used whereas for the indirect ^{13}C/^{15}N

dimensions a line width of 10 * Γ_H was used in both cases (for more parameters see legend to Fig. 7). Cyana performs automatic structure calculations using the un-assigned simulated peak lists, and is tolerant to noise peaks, which are filtered out during the structure calculation. Increasing the line width leads to more spectral overlap (fewer peaks observed, see Fig. 2 middle row, right column), but also leads to the need for larger peak assignment tolerances during the automatic peak assignment procedure. Consequently, on average, Cyana had more assignment possibilities for each peak for larger line widths. During the iterative structure refinement and peak assignments algorithm in Cyana, peak assignments that are not consistent with the candidate structure are removed. Hence during the final cycle, Cyana had unique assignments for most of the peaks leading to a well-established structure (see below), but with an increasingly slower convergence for the data set for larger line widths and larger tolerances (data not shown).

The accuracy and precision for the structure calculations are shown in Fig. 7 as a function of the Gaussian line width, Γ. It is seen that the accuracy, measured as the RMSD deviation from the reference structure, remains approximately constant on a converged value for line widths up to 20 ppb in the direct proton dimension (all other line widths were proportional to this line width). For data with good sensitivity, modeled here by a signal/noise ratio, SN = $1/h_{min}$ (see "Methods"), of SN = 75 and adding only a fraction of noise peaks $f_N = 2$ % (Eq. 10), the converged accuracy RMSD levels are ca. 1.6 Å for small line widths up to 28 ppb (where 37 % fewer peaks are observed compared to the spectrum with smallest line width), and increases to 2.05 Å for $\Gamma = 32$ ppb, then diverges for $\Gamma = 40$ ppb (RMSD = 19.9 Å, see Table S1). Conversely, for data of lesser sensitivity; SN = 25, $f_N = 10$ %, the RSMD levels out at ca. 1.65 Å for small line widths and starts suddenly to diverge for $\Gamma \geq 24$ ppb. It appears that decreasing the data quality does not lead to a gradual decrease in the accuracy of the structure, but rather that the poorer data is remedied by the Cyana algorithm, until a threshold is reached for the data imperfections beyond which the structure calculation diverges. For the precision of the structure the scenario is slightly different with ensemble heavy atom RMSDs against the average structure of ca. 0.8 and 1.1 Å for small line widths (Fig. 7) for the good sensitivity and less sensitive data (see above), respectively. The precision RMSD increases gradually for the good sensitivity data from a line width between 20 and 32 ppb (see Fig. 7 and Table S1). Furthermore (see Table S1), noise peaks seem to be better tolerated in structure calculations than low SN (at least for the ranges studied here), with converged structures for $f_N = 10$ and 20 % (SN = 75 and $\Gamma = 4$, 12 and 28 ppb) whereas low SN = 10 produces diverged structures in all cases ($f_N = 2$ % and $\Gamma = 4$, 12 and 28 ppb) and SN = 25 a diverged structure for the largest simulated line width $\Gamma = 28$ ppb.

Conclusion

We have presented a tool, VirtualSpectrum, for generating assigned peak lists for various multi-dimensional NMR experiments. VirtualSpectrum can serve to produce artificial data to test NMR procedures in cases where experimental data is insufficient or problematic to obtain. VirtualSpectrum is fast, flexible, robust and produces peak lists, which can be tuned to match experimental quality, and overall appearance, of most multi-dimensional NMR experiments. VirtualSpectrum uses an analytic expression to represent the spectrum, and the peak positions are derived by numerical routines that identify local maxima, seamlessly handling overlap between signals. Our analysis of a few case studies shows, as expected, that the amount of overlapping peaks increases with the resonance line width and proteins size, and that more overlap is present in spectra for proteins in an unstructured compared to a structured state. In addition proteins with predominantly helical secondary structure have more overlap compared to mixed secondary structure and with predominantly beta-sheet structure. Furthermore, we demonstrated the applicability of VirtualSpectrum showing that (1) spin system generation, for relatively large line widths, is prone to errors in peak alignments approximately equal to the Gaussian line width, (2) backbone ^{15}N is equally good or slightly better than $^{13}C'$ to resolve peaks in 2D spectra for intrinsically disordered proteins, and (3) for solid state NMR side chain carbons are better than Cα, which is again better than C$'$ to resolve peaks in 3D spectra. We expect that VirtualSpectrum will find widespread applications in the future for the generation and analysis of NMR data, and in particular, will be used to evaluate the performance of software for structure determination or resonance assignments. Here VirtualSpectrum was applied along with Cyana to test the influence of spectral resolution and sensitivity on the accuracy of structure calculation revealing that structure calculations with Cyana can tolerate overlap in the spectra with 37 % fewer peaks observed for good sensitivity data, whereas calculations diverge with less overlap for data corresponding to lesser sensitivity.

Acknowledgments We acknowledge Frans A. A. Mulder for kindly providing the 2D projections of the HNCO spectrum for hNL3-Cyt and the Watergate-NOESY for HEWL. We thank M.Sc. Julie S. Nielsen for careful proofreading of the manuscript.

References

Advanced Chemistry Development, Inc. (ACD/Labs), NMR predictors (2007) Toronto, ON, Canada

Allard P, Helgstrand M, Hard T (1997) A method for simulation of NOESY, ROESY, and off-resonance ROESY spectra. J Magn Reson 129(1):19–29

Anfinsen CB (1973) Principles that govern folding of protein chains. Science 181(4096):223–230

Bartels C, Guntert P, Billeter M, Wuthrich K (1997) GARANT—a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. J Comput Chem 18(1):139–149

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucl Acids Res 28(1):235–242

Bertoncini CW, Jung YS, Fernandez CO, Hoyer W, Griesinger C, Jovin TM, Zweckstetter M (2005) Release of long-range tertiary interactions potentiates aggregation of natively unstructured alpha-syncline. Proc Natl Acad Sci USA 102(5):1430–1435

Binev Y, Aires-De-Sousa J (2004) Structure-based predictions of H-1 NMR chemical shifts using feed-forward neural networks. J Chem Inf Comput Sci 44(3):940–945

Blake CCF, Koenig DF, Mair GA, North ACT, Phillips DC, Sarma VR (1965) Structure of hen egg-white lysozyme—a 3-dimensional Fourier synthesis at 2Å resolution. Nature 206(4986):757–761

Boelens R, Koning TMG, Kaptein R (1988) Determination of biomolecular structures from proton–proton NOE's using a relaxation matrix approach. J Mol Struct 173:299–311

Bouaziz S, Vanheijenoort C, Guittet E, Lallemand JY (1992) Application of homonuclear 3-dimensional NMR-spectroscopy to the study of a protein in solution. J Chim Phys Phys Chim Biol 89(2):147–156

Bouvignies G, Meier S, Grzesiek S, Blackledge M (2006) Ultrahigh-resolution backbone structure of perdeuterated protein GB1 using residual dipolar couplings from two alignment media. Angew Chem Intl Edit 45(48):8166–8169

Breg JN, Sarda L, Cozzone PJ, Rugani N, Boelens R, Kaptein R (1995) Solution structure of porcine pancreatic procolipase as determined from H-1 homonuclear 2-dimensional and 3-dimensional NMR. Eur J Biochem 227(3):663–672

Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein-folding—a synthesis. Proteins Struct Funct Gen 21(3):167–195

Bychkova VE, Pain RH, Ptitsyn OB (1988) The molten globule state is involved in the translocation of proteins across membranes. FEBS Lett 238(2):231–234

Castellani F, van Rossum B, Diehl A, Schubert M, Rehbein K, Oschkinat H (2002) Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. Nature 420(6911):98–102

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPIPE—a multidimensional spectral processing system based on unix pipes. J Biomol NMR 6(3):277–293

Dolgikh DA, Gilmanshin RI, Brazhnikov EV, Bychkova VE, Semisotnov GV, Venyaminov SY, Ptitsyn OB (1981) Alpha-lactalbumin—compact state with fluctuating tertiary structure. FEBS Lett 136(2):311–315

Donne DG, Gozansky EK, Gorenstein DG (1995) Exact vs approximate methods for simulation of 3D NOE–NOE spectra. J Magn Reson, Ser B 106(2):156–163

Driscoll PC, Gronenborn AM, Beress L, Clore GM (1989) Determination of the 3-dimensional solution structure of the antihypertensive and antiviral protein BDS-I from the sea-anemone anemonia-sulcata—a study using nuclear magnetic-resonance and hybrid distance geometry-dynamical simulated annealing. Biochemistry 28(5):2188–2198

Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CR, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang CH, Kissinger CR, Bailey RW, Griswold MD, Chiu M, Garner EC, Obradovic Z (2001) Intrinsically disordered protein. J Mol Graph Model 19(1):26–59

Dyson HJ, Wright PE (2002) Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol 12(1):54–60

Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6(3):197–208

Fossi M, Castellani T, Nilges M, Oschkinat H, van Rossum BJ (2005) SOLARIA: a protocol for automated cross-peak assignment and structure calculation for solid-state magic-angle spinning NMR spectroscopy. Angew Chem Intl Edit 44(38):6151–6154

Gallagher T, Alexander P, Bryan P, Gilliland GL (1994) 2 crystal-structures of the B1 immunoglobulin-binding domain of streptococcal protein-G and comparison with NMR. Biochemistry 33(15):4721–4729

Goddard TD, D. G. Kneller, SPARKY 3 SPARKY3. University of California, San Francisco

Golotvin SS, Vodopianov E, Pol R, Lefebvre BA, Williams AJ, Rutkowske RD, Spitzer TD (2007) Automated structure verification based on a combination of 1D H-1 NMR and 2D H-1-C-13 HSQC spectra. Magn Reson Chem 45(10):803–813

Goto NK, Kay LE (2000) New developments in isotope labeling strategies for protein solution NMR spectroscopy. Curr Opin Struct Biol 10(5):585–592

Gradmann S, Ader C, Heinrich I, Nand D, Dittmann M, Cukkemane A, van Dijk M, Bonvin A, Engelhard M, Baldus M (2012) Rapid prediction of multi-dimensional NMR data sets. J Biomol NMR 54(4):377–387

Gronwald W, Kalbitzer HR (2004) Automated structure determination of proteins by NMR spectroscopy. Prog Nucl Magn Res Spectrosc 44(1–2):33–96

Guerry P, Herrmann T (2011) Advances in automated NMR protein structure determination. Q Rev Biophys 44(3):257–309

Herrmann T, Guntert P, Wuthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol 319(1):209–227

Hu K-N, Qiang W, Tycko R (2011) A general Monte Carlo/simulated annealing algorithm for resonance assignment in NMR of uniformly labeled biopolymers. J Biomol NMR 50(3):267–276

Igumenova TI, McDermott AE, Zilm KW, Martin RW, Paulson EK, Wand AJ (2004) Assignments of carbon NMR resonances for microcrystalline ubiquitin. J Am Chem Soc 126(21):6720–6727

Jeener J, Meier BH, Bachmann P, Ernst RR (1979) Investigation of exchange processes by 2-dimensional NMR-spectroscopy. J Chem Phys 71(11):4546–4553

Jensen MR, Markwick PRL, Meier S, Griesinger C, Zweckstetter M, Grzesiek S, Bernado P, Blackledge M (2009) Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. Structure 17(9):1169–1185

Jung YS, Zweckstetter M (2004) Mars—robust automatic backbone assignment of proteins. J Biomol NMR 30(1):11–23

Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Ono AM, Guntert P (2006) Optimal isotope labelling for NMR protein structure determinations. Nature 440(7080):52–57

Keepers JW, James TK (1984) A theoretical study of distance determinations from NMR. Two-dimensional nuclear overhauser effect spectra. J Magn Reson 57:404–426

Kigawa T, Yabuki T, Yoshida Y, Tsutsui M, Ito Y, Shibata T, Yokoyama S (1999) Cell-free production and stable-isotope

labeling of milligram quantities of proteins. FEBS Lett 442(1):15–19
Kumar A, Ernst RR, Wuthrich K (1980) A two-dimensional nuclear overhauser enhancement (2d noe) experiment for the elucidation of complete proton–proton cross-relaxation networks in biological macromolecules. Biochem Biophys Res Commun 95(1):1–6
Lange A, Becker S, Seidel K, Giller K, Pongs O, Baldus M (2005) A concept for rapid protein-structure determination by solid-state NMR spectroscopy. Angew Chem Intl Edit 44(14):2089–2092
Linge JP, O'Donoghue SI, Nilges M (2001) Automated assignment of ambiguous nuclear overhauser effects with ARIA. Nucl Magn Reson Bio Macromol Pt B 339:71–90
Linge JP, Habeck M, Rieping W, Nilges M (2003) ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics 19(2):315–316
Loquet A, Bardiaux B, Gardiennet C, Blanchet C, Baldus M, Nilges M, Malliavin T, Bockmann A (2008) 3D structure determination of the Crh protein from highly ambiguous solid-state NMR restraints. J Am Chem Soc 130(11):3579–3589
Loquet A, Gardiennet C, Bockmann A (2010) Protein 3D structure determination by high-resolution solid-state NMR. C R Chim 13(4):423–430
Macura S, Ernst RR (1980) Elucidation of cross relaxation in liquids by two-dimensional NMR-spectroscopy. Mol Phys 41(1):95–117
Malmodin D, Papavoine CHM, Billeter M (2003) Fully automated sequence-specific resonance assignments of heteronuclear protein spectra. J Biomol NMR 27(1):69–79
Manolikas T, Herrmann T, Meier BH (2008) Protein structure determination from C-13 spin-diffusion solid-state NMR spectroscopy. J Am Chem Soc 130(12):3959–3966
Marley J, Lu M, Bracken C (2001) A method for efficient isotopic labeling of recombinant proteins. J Biomol NMR 20(1):71–75
Matsuki Y, Akutsu H, Fujiwara T (2007) Spectral fitting for signal assignment and structural analysis of uniformly C-13-labeled solid proteins by simulated annealing based on chemical shifts and spin dynamics. J Biomol NMR 38(4):325–339
Meier S, Blackledge M, Grzesiek S (2008) Conformational distributions of unfolded polypeptides from novel NMR techniques. J Chem Phys 128(5):052204
Mittag T, Forman-Kay JD (2007) Atomic-level characterization of disordered protein ensembles. Curr Opin Struct Biol 17(1):3–14
Moseley HNB, Montelione GT (1999) Automated analysis of NMR assignments and structures for proteins. Curr Opin Struct Biol 9(5):635–642
Moseley HNB, Monleon D, Montelione GT (2001) Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. Nucl Magn Reson Biol Macromol Pt B 339:91–108
Moseley HNB, Sperling LJ, Rienstra CM (2010) Automated protein resonance assignments of magic angle spinning solid-state NMR spectra of beta 1 immunoglobulin binding domain of protein G (GB1). J Biomol NMR 48(3):123–128
Naito A, Kawamura I (2007) Solid-state NMR as a method to reveal structure and membrane-interaction of amyloidogenic proteins and peptides. Biochim Biophys Biomembr 1768(8):1900–1912
Nielsen JT, Eghbalnia HR, Nielsen NC (2012) Chemical shift prediction for structure calculation and quality assessment using an optimally parameterized force field. Prog Nucl Magn Reson Spectrosc 60:1–28
Nielsen JT, Kulminskaya NV, Bjerring M, Nielsen NC (2014) Automated robust and accurate assignment of protein resonances for solid state NMR. J Biomol NMR 59(2):119–134
Nilges M (1995) Calculation of protein structures with ambiguous distance restraints - automated assignment of ambiguous noe crosspeaks and disulfide connectivities. J Mol Biol 245(5):645–660
Nilges M, Macias MJ, Odonoghue SI, Oschkinat H (1997) Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. J Mol Biol 269(3):408–422
Nocedal J, Wright SJ (2000) Numerical optimization. Springer, Berlin
Oliphant TE (2007) Python for scientific computing. Comput Sci Eng 9(3):10–20
Paz A, Zeev-Ben-Mordehai T, Lundqvist M, Sherman E, Mylonas E, Weiner L, Haran G, Svergun DI, Mulder FAA, Sussman JL, Silman I (2008) Biophysical characterization of the unstructured cytoplasmic domain of the human neuronal adhesion protein neuroligin 3. Biophys J 95(4):1928–1944
Redfield C, Dobson CM (1988) Sequential H-1-NMR assignments and secondary structure of hen egg-white lysozyme in solution. Biochemistry 27(1):122–136
Rieping W, Habeck M, Bardiaux B, Bernard A, Malliavin TE, Nilges M (2007) ARIA2: automated NOE assignment and data integration in NMR structure calculation. Bioinformatics 23(3):381–382
Schneider R, Odronitz F, Hammesfahr B, Hellkamp M, Kollmar M (2013) Peakr: simulating solid-state NMR spectra of proteins. Bioinformatics 29(9):1134–1140
Shojania S, O'Neil JD (2006) HIV-1 Tat is a natively unfolded protein—the solution conformation and dynamics of reduced HIV-1 Tat-(1-72) by NMR spectroscopy. J Biol Chem 281(13):8347–8356
Simorre JP, Caille A, Marion D, Marion D, Ptak M (1991) 2-dimensional and 3-dimensional H-1-NMR studies of a wheat phospholipid transfer protein—sequential resonance assignments and secondary structure. Biochemistry 30(49):11600–11608
Stevens TJ, Fogh RH, Boucher W, Higman VA, Eisenmenger F, Bardiaux B, van Rossum B-J, Oschkinat H, Laue ED (2011) A software framework for analysing solid-state MAS NMR data. J Biomol NMR 51(4):437–447
Sugase K, Dyson HJ, Wright PE (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature 447(7147):1021–1025
Sziegat F, Silvers R, Hahnke M, Jensen MR, Blackledge M, Wirmer-Bartoschek J, Schwalbe H (2012) Disentangling the coil: modulation of conformational and dynamic properties by site-directed mutation in the non-native state of hen egg white lysozyme. Biochemistry 51(16):3361–3372
Takegoshi K, Nakamura S, Terao T (2001) C-13-H-1 dipolar-assisted rotational resonance in magic-angle spinning NMR. Chem Phys Lett 344(5–6):631–637
Tamiola K, Acar B, Mulder FAA (2010) Sequence-specific random coil chemical shifts of intrinsically disordered proteins. JACS 132(51):18000–18003
Thiriot DS, Nevzorov AA, Zagyanskiy L, Wu CH, Opella SJ (2004) Structure of the coat protein in Pf1 bacteriophage determined by solid-state NMR Spectroscopy. J Mol Biol 341(3):869–879
Tompa P (2002) Intrinsically unstructured proteins. Trends Biochem Sci 27(10):527–533
Tycko R (2006) Molecular structure of amyloid fibrils: insights from solid-state NMR. Q Rev Biophys 39(1):1–55
Tycko R, Hu K-N (2010) A Monte Carlo/simulated annealing algorithm for sequential resonance assignment in solid state NMR of uniformly labeled proteins with magic-angle spinning. J Magn Reson 205(2):304–314
Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao H, Markley JL (2008) BioMagResBank. Nucl Acids Res 36:D402–D408
Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. Protein Sci 11(4):739–756
Vajpai N, Gentner M, Huang JR, Blackledge M, Grzesiek S (2010) Side-chain Chi(1) conformations in urea-denatured ubiquitin and

protein G from (3)J coupling constants and residual dipolar couplings. J Am Chem Soc 132(9):3196–3203

Vijaykumar S, Bugg CE, Cook WJ (1987) Structure of ubiquitin refined at 1.8 Å resolution. J Mol Biol 194(3):531–544

Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinas P, Ulrich EL, Markley JL, Ionides J, Laue ED (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. Prot Struct Funct Bioinf 59(4):687–696

Williamson MP, Havel TF, Wuthrich K (1985) Solution conformation of proteinase inhibitor-iia from bull seminal plasma by h-1 nuclear magnetic-resonance and distance geometry. J Mol Biol 182(2):295–315

Wittekind M, Rajagopal P, Branchini BR, Reizer J, Saier MH, Klevit RE (1992) Solution structure of the phosphocarrier protein HPR from bacillus-subtilis by 2-dimensional NMR-spectroscopy. Protein Sci 1(10):1363–1376

Wood K, Paz A, Dijkstra K, Scheek RM, Otten R, Silman I, Sussman JL, Mulder FAA (2012) Backbone and side chain NMR assignments for the intrinsically disordered cytoplasmic domain of human neuroligin-3. Biomol NMR Assign 6(1):15–18

Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293(2):321–331

Wüthrich K (1986) NMR of proteins and nucleic acids. Wiley, New York

Zech SG, Wand AJ, McDermott AE (2005) Protein structure determination by high-resolution solid-state NMR spectroscopy: application to microcrystalline ubiquitin. J Am Chem Soc 127(24):8618–8626

Zhu LM, Reid BR (1995) An improved noesy simulation program for partially relaxed spectra—birder. J Magn Reson, Ser B 106(3):227–235

Zimmerman DE, Kulikowski CA, Huang YP, Feng WQ, Tashiro M, Shimotakahara S, Chien CY, Powers R, Montelione GT (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. J Mol Biol 269(4):592–610