Journal of Magnetic Resonance 205 (2010) 304–314

Contents lists available at ScienceDirect

Journal of Magnetic Resonance

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

A Monte Carlo/simulated annealing algorithm for sequential resonance assignment in solid state NMR of uniformly labeled proteins with magic-angle spinning

Robert Tycko *, Kan-Nian Hu

Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-0520, United States

ARTICLE INFO

Article history:
Received 24 March 2010
Revised 19 May 2010
Available online 25 May 2010

Keywords:
Automated assignment
Chemical shifts
Amyloid
Prion
HET-s
Stochastic recoupling

ABSTRACT

We describe a computational approach to sequential resonance assignment in solid state NMR studies of uniformly ^{15}N,^{13}C-labeled proteins with magic-angle spinning. As input, the algorithm uses only the protein sequence and lists of $^{15}N/^{13}C_{\alpha}$ crosspeaks from 2D NCACX and NCOCX spectra that include possible residue-type assignments of each crosspeak. Assignment of crosspeaks to specific residues is carried out by a Monte Carlo/simulated annealing algorithm, implemented in the program MC_ASSIGN1. The algorithm tolerates substantial ambiguity in residue-type assignments and coexistence of visible and invisible segments in the protein sequence. We use MC_ASSIGN1 and our own 2D spectra to replicate and extend the sequential assignments for uniformly-labeled HET-s(218–289) fibrils previously determined manually by Siemer et al. (J. Biomol. NMR, 34 (2006) 75–87) from a more extensive set of 2D and 3D spectra. Accurate assignments by MC_ASSIGN1 do not require data that are of exceptionally high quality. Use of MC_ASSIGN1 (and its extensions to other types of 2D and 3D data) is likely to alleviate many of the difficulties and uncertainties associated with manual resonance assignments in solid state NMR studies of uniformly labeled proteins, where spectral resolution and signal-to-noise are often sub-optimal.

Published by Elsevier Inc.

1. Introduction

In solid state nuclear magnetic resonance (NMR) studies of uniformly ^{15}N,^{13}C-labeled proteins with magic-angle spinning (MAS), assignment of the observed ^{15}N and ^{13}C resonances to specific residues is a prerequisite for the determination of molecular structure or characterization of molecular dynamics. Resonance assignment typically proceeds in a sequential manner, by connecting ^{13}C signals of residue k with ^{13}C signals of residue $k+1$ through the ^{15}N signals of backbone amide sites, often in two-dimensional (2D) or three-dimensional (3D) NCACX and NCOCX spectra [1–15]. Manual sequential assignment from 2D NCACX and NCOCX spectra is easy when most ^{15}N chemical shifts are unique and well resolved and when most ^{13}C chemical shifts can be assigned to unique residue types. When overlap and degeneracy of ^{15}N resonances is severe and residue-type assignments are ambiguous, manual assignment becomes difficult because of the many possible candidates for $k/k+1$ residue pairs that must be explored and either proven or disproven. 3D spectroscopy helps, but still may not readily yield unique assignments. The situation becomes more complicated when only certain segments of the protein sequence contribute to the solid state NMR spectra [16–23], due to variations in rigidity and structural order, and when the identity of these segments is unknown. In the end, resonance assignment is a tedious and potentially error-prone process except when the solid state NMR data are of extremely high quality.

In this paper, we describe an alternative approach to resonance assignment in solid state MAS NMR of uniformly labeled proteins, in which residue-specific assignments are generated in an automated manner from lists of crosspeaks in 2D NCACX and NCOCX spectra by a Monte Carlo/simulated annealing (MC/SA) computational algorithm. Using our 2D spectra of uniformly-labeled HET-s(218–289) fibrils, for which resonance assignments from a manual analysis of 2D and 3D spectra have been reported previously by Siemer et al. [17], we show that the MC/SA algorithm leads to complete and correct assignments even when there is ambiguity in the residue-type assignments of many of the 2D crosspeaks. Unlike manual assignment procedures, the MC/SA algorithm provides a complete and objective picture of the information content of the solid state NMR data, allowing either full or partial assignments to be extracted from data that are not necessarily ideal.

* Corresponding author. Address: National Institutes of Health, Building 5, Room 112, Bethesda, MD 20892-0520, United States. Fax: +1 301 496 0825.
E-mail address: robertty@mail.nih.gov (R. Tycko).

1090-7807/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jmr.2010.05.013

2. Methods

2.1. Sample preparation

Uniformly ^{15}N,^{13}C-labeled HET-s(218–289) (sequence MKID AIVGRNSAKD IRTEERARVQ LGNVVTAAAL HGGIRISDQT TNSVET-VVGK GESRVLIGNE YGGKGFWDN HHHHHH, representing residues 218–289 of the *Podospora anserina* HET-s protein with an additional N-terminal Met residue and a C-terminal hexa-His tag) was expressed and purified as previously described [17,21]. Fibrils were prepared by incubation at 0.4 mM protein concentration in Tris buffer at pH 8 and 4 °C for 14 days. Fibril formation was confirmed by transmission electron microscopy (see Fig. S1 of supplementary information). Fibrils were pelleted, lyophilized, and packed in a MAS rotor (1.8 mm outer diameter, 10.5 μl volume), then rehydrated in the rotor by addition of 5 μl of water. The sample contained approximately 5 mg of HET-s(218–289).

2.2. NMR measurements

Spectra were obtained with three-channel MAS probes constructed by the group of Dr. Ago Samoson (National Institute of Chemical Physics and Biophysics, Tallinn, Estonia). 2D NCACX and NCOCX spectra were acquired with a Varian Infinity spectrometer at 17.6 T (188.0 MHz ^{13}C NMR frequency) and 17.0 kHz MAS, using 5.0 ms cross-polarization for ^{15}N–^{13}C polarization transfer to either C_α (NCACX) or CO (NCOCX) sites after the t_1 period, followed by 2.82 ms finite-pulse radio-frequency-driven recoupling (fpRFDR) [24,25] for ^{13}C–^{13}C polarization transfer before the t_2 period. ^{13}C π pulses in the fpRFDR periods were 20.0 μs at 45 ppm carrier frequency (NCACX) or 10.0 μs at 105 ppm carrier frequency (NCOCX). Two-pulse phase-modulated (TPPM) proton decoupling [26] at 110 kHz was applied during t_1 and t_2. Continuous-wave decoupling at 110 kHz was applied during ^{15}N–^{13}C and ^{13}C–^{13}C polarization transfer periods. Maximum t_1 and t_2 periods were 9.10 ms and 7.68 ms, respectively. Total measurement times were 22 h for each 2D spectrum, with 1.0 s recycle delays.

The 2D ^{13}C–^{13}C (CC) spectrum was acquired with a Varian InfinityPlus spectrometer at 14.1 T (150.7 MHz ^{13}C NMR frequency) and 40.0 kHz MAS, using a novel zero-quantum stochastic dipolar recoupling (ZQ-SDR) pulse sequence for longitudinal ^{13}C–^{13}C polarization transfers between t_1 and t_2. The ZQ-SDR sequence consisted of four fpRFDR blocks with 7.5 μs ^{13}C π pulses, 32 rotor periods in each block, separated by randomly-chosen delays that ranged from 0 to 3 rotor periods in length. The delays were determined by a random number generator within the pulse program. This ZQ-SDR sequence is related conceptually to the previously-described double-quantum SDR technique [27,28], and similarly leads to ^{13}C-^{13}C polarization transfers that are unaffected by quantum mechanical interference between non-commuting pairwise dipole–dipole couplings (*i.e.*, "dipolar truncation"). TPPM decoupling at 125 kHz was applied during the t_1 and t_2 periods. No decoupling was applied during the ZQ-SDR period. Maximum t_1 and t_2 periods were 7.97 ms and 15.36 ms, respectively. Total measurement time was 17 h, with 1.5 s recycle delays.

2D NMR spectra were processed with NMRPipe software [29]. Crosspeaks were picked manually using Sparky software (available at http://www.cgl.ucsf.edu/home/sparky/). Residue-type assignments of C_α chemical shifts were determined manually from crosspeak patterns (principally in the CC spectrum, but also from sets of ^{13}C resonances with a common ^{15}N chemical shift in the NCACX and NCOCX spectra). ^{13}C and ^{15}N chemical shifts are referred to tetramethylsilane and liquid ammonia, respectively, consistent with the work of Siemer et al. [17].

2.3. MC/SA algorithm

Residue-specific assignments were determined from lists of $^{15}N/^{13}C_\alpha$ crosspeaks in the 2D NCACX and NCOCX spectra with the Fortran95 computer program MC_ASSIGN1 (see supplementary information), which implements the algorithm in Fig. 1. The crosspeak lists include the ^{15}N and ^{13}C chemical shifts, the uncertainties in these shifts, the maximum degeneracies (in case more than one residue may contribute to a given crosspeak), and the possible residue-type assignments. Crosspeak lists are prepared by manual analyses of the 2D spectra. Residue-type assignments are typically determined from patterns of sidechain ^{13}C chemical shifts in 2D CC spectra, supplemented by comparisons with the NCACX and NCOCX spectra. Each $^{15}N/^{13}C_\alpha$ crosspeak can have multiple residue-type assignments. MC_ASSIGN1 attempts to assign one NCACX crosspeak and one NCOCX crosspeak to each residue in the protein sequence (or leave certain residues without assigned crosspeaks, called a "null assignment", if the number of residues exceeds the number of crosspeaks) in such a way that the number of "good connections" (N_g) is maximized, the numbers of "bad connections" (N_b) and "edges" (N_e) are minimized, and the number of unused crosspeaks (N_u) is minimized. N_g is the number of residues with non-null NCACX and non-null NCOCX assignments for which the two $^{13}C_\alpha$ shifts agree to within the allowed uncertainty, plus the number of $k/k+1$ pairs for which the ^{15}N shift in the NCOCX assignment of residue k and the NCACX assignment of residue $k+1$ agree to within the allowed uncertainty. N_b is the total number of these $^{13}C_\alpha$ shift pairs and ^{15}N shift pairs that do not agree to within the allowed uncertainty. N_e is the number of residues that have a null NCACX assignment and a non-null NCOCX assignment or vice versa, plus the number of $k/k+1$ pairs for which residue k has a null NCOCX assignment and residue $k+1$ has a non-null NCACX assignment or vice versa. N_u is the number of crosspeaks in the NCACX and NCOCX lists that have been not been assigned to any residues. With these definitions, any assignment candidate has a score S, defined by

$$S \equiv w_1 N_g - w_2 (N_b + \frac{1}{4} N_e) - w_3 N_u \quad (1)$$

where w_1, w_2, and w_3 are weighting factors that are incremented gradually during execution of the algorithm. The factor of 1/4 in Eq. (1) is somewhat arbitrary, motivated by the idea that N_e should be minimized (*i.e.*, the lengths of fully assigned segments should be maximized if possible) but need not be zero, while N_b should be zero in the final, correct assignment. If the uncertainties in ^{15}N or ^{13}C chemical shifts in two spectra are ε_a and ε_b, a connection is considered good if the absolute value of the chemical shift difference is less than $\sqrt{\varepsilon_a^2 + \varepsilon_b^2}$.

Starting with null assignments for all residues and with w_1, w_2, and w_3 set to their minimum values, MC_ASSIGN1 chooses a residue at random and randomly changes its current NCACX assignment to another assignment (*i.e.*, another NCACX crosspeak) that is allowed for the given residue type. If the current assignment is not null, the assignment can be changed to null with 40% probability or to another allowed assignment (if at least one exists) with 60% probability. If a given crosspeak has degeneracy n_{max}, then the same crosspeak can be assigned to as many as n_{max} residues. The change in score ΔS resulting from this random change in NCACX assignment of a single randomly-chosen residue is then calculated. The quantity $\exp(\Delta S)$ is then compared to a random number x from the interval (0, 1). If $\exp(\Delta S) \geqslant x$, the new assignment is accepted. If $\exp(\Delta S) < x$, the new assignment is rejected and the old assignment is restored.

In the same manner, MC_ASSIGN1 then attempts to change the NCOCX assignment of another randomly-chosen residue. One