Journal of Biomolecular NMR, **25:** 11–23, 2003.
KLUWER/ESCOM

Data requirements for reliable chemical shift assignments in deuterated proteins

T. Kevin Hitchens, Scott A. McCallum & Gordon S. Rule*
Department of Biological Sciences, Carnegie Mellon University, 4400 Fifth Ave, Pittsburgh, PA 15213, U.S.A.

Received 20 June 2002; Accepted 18 September 2002

Key words: automated assignment, data requirement, deuterated proteins

Abstract

The information required for chemical shift assignments in large deuterated proteins was investigated using a Monte Carlo approach (Hitchens et al., 2002). In particular, the consequences of missing amide resonances on the reliability of assignments derived from C_α and CO or from C_α and C_β chemical shifts was investigated. Missing amide resonances reduce both the number of correct assignments as well as the confidence in these assignments. More significantly, a number of undetectable errors can arise when as few as 9% of the amide resonances are missing from the spectra. However, the use of information from residue specific labeling as well as local and long-range distance constraints improves the reliability and extent of assignment. It is also shown that missing residues have only a minor effect on the assignment of protein-ligand complexes using C_α and CO chemical shifts and C_α inter-residue connectivity, provided that the known chemical shifts of the unliganded protein are utilized in the assignment process.

Abbreviations: NOE, nuclear Overhauser effect.

Introduction

NMR spectroscopy has proven to be broadly applicable to the study of the structure and dynamics of proteins. It has also been extensively utilized to study protein-ligand interactions, such as in drug discovery. An inherent problem with protein NMR spectroscopy is the limit on the size of proteins that may be investigated. Within the last decade, triple resonance techniques have enabled the study of proteins with molecular weights in the 20–30 kDa range (Bax and Grzesiek, 1993). The principal advantage of triple resonance techniques over homonuclear techniques is the ability to connect adjacent residues by scalar couplings between heteronuclear spins. In addition, the increased spectral resolution conferred by the heteronuclear spin has substantially reduced spectral overlap. More recently, advances in protein deuteration (Sattler and Fesik, 1996; Kay and Gardner, 1997; Clore and Gronenborn, 1997), partial orientation of samples (Bax et al., 2001), and the beneficial effects of interference of cross-relaxation (Riek et al., 1999) have facilitated the study of significantly larger systems.

A central component of all NMR studies on proteins is the process of resonance assignment. Obviously, it is desirable to obtain both reliable and complete assignments for the system under study. In many cases, such as the study of backbone protein dynamics, it is sufficient to obtain assignments of only the amide resonances. In other cases, such as structure determination, it is necessary to obtain assignments for as many of the NMR active nuclei as possible. Resonance assignments are traditionally obtained by a four step process involving the collection of chemical shifts into spin-systems, prediction of the residue type of the spin-system from characteristic chemical shifts, sequential ordering of spin-systems using inter-residue scalar and/or inter-proton distances, and the correct placement of sequentially connected spin-systems onto the known primary sequence of the

*To whom correspondence should be addressed. E-mail: rule@andrew.cmu.edu

protein (Wüthrich, 1986). Non-traditional approaches utilize measured inter-proton distances (i.e., NOE) to construct networks of dipolar coupled spins that are then used to link spin-systems (Wand and Nelson, 1991; Bailey-Kellogg et al., 2000) or to determine the structure of the protein in the absence of assignments (see Grishaev and Llinas, 2002). However, even these non-traditional approaches ultimately base their assignments on residue-type information provided by characteristic chemical shifts associated with the coupled protons.

Since the assignment process is defined by a number of discrete steps it is highly amenable to automation. Approaches that use data from triple resonance experiments have been described in the literature and several excellent review articles exist (Zimmerman and Montelione, 1995; Moseley and Montelione, 1999). In most cases the residue type of a spin-system is predicted from C_α and C_β chemical shifts. Current automated approaches differ mainly in how they place linked spin-systems on to the primary sequence. For example, AutoAssign (Zimmerman et al., 1997) utilizes a best first approach to rapidly place connected segments onto the primary sequence. Other approaches, such as Mapper (Güntert et al., 2000) and TATApro (Atreya et al., 2000) utilize near exhaustive searches to find a global solution to the placement of connected segments on to the primary sequence.

The extent of the residue assignments that are obtained by automated methods is ultimately related to the nature, quantity, and quality of the data. In the case of smaller protonated proteins, it is common to acquire extensive sets of inter- and intra-residue H_α and H_β shifts to facilitate residue-type prediction as well as linking of sequential spin-systems. Consequently, most automated systems yield near complete assignments with these data. In the case of highly deuterated proteins it is difficult to obtain a large number of aliphatic proton shifts. Consequently, the prediction of residue-type usually depends entirely on the characteristic C_α and C_β chemical shifts. Furthermore, since protons cannot be used to establish inter-residue connectivity, other spins, such as CO and the amide nitrogen, have been utilized as a source of inter-residue connectivity.

An additional complication associated with the study of highly deuterated proteins is the reliance on the detection of signals from the amide proton. Under favorable conditions it is possible to replace all of the amide deuterons in a deuterated protein by refolding the protein in H_2O. However, in a number of systems this is not feasible. This leads to a fraction of amides that remain unobservable due to slow chemical exchange with the solvent. Earlier work on the assignments of ~50 kDa homodimeric glutathione transferases (McCallum et al., 1999; Hitchens et al., 2001) suggested that the presence of large numbers of missing amide resonances can cause significant problems in the assignment of large deuterated proteins. Here, we utilize Monte Carlo methods (Hitchens et al., 2002) to systematically investigate the extent of information required to obtain reliable chemical shift assignments when a significant fraction of amide resonances are absent from the spectra. A Monte Carlo approach was employed because it provides a natural way to sample the ensemble of possible residue assignments. Thus the best assignment solution for the given set of data can be obtained along with the ensemble of other solutions that are similar to the best solution. Information on the ensemble of potential solutions provides a unique insight into the reliability of the chemical shift assignments for any given set of data.

To investigate the relationship between the quantity and quality of experimental data and the derived assignments we have used a 259 residue protein, the N-terminal domain of enzyme I in the phosphenolypyruvate:sugar phosphotransferase system of *E. coli*, as a test system (Garrett et al., 1997). We focused on investigating the influence of four parameters on the assignment process: 1) the number of residues with observable amide resonances, 2) the type and extent of inter-residue connectivities, 3) the effect of primary and tertiary NOEs, and 4) the inclusion of residue-type information from residue specific labeling.

On the basis of these simulations we show that inter-residue connectivity provided solely by C_α and CO shifts can only be successfully used for assignment purposes if additional information from NOESY spectra and residue specific labeling is available. Inter-residue connectivities that are established using C_α and C_β shifts appear to provide sufficient information for reliable assignments *if* all of the amide resonances are observable. The number of residues that can be assigned with confidence decreases significantly if spin-systems are missing because of unobservable amide resonances. Furthermore, it may be possible to incorrectly assign a fraction of residues under these conditions. Additional inter-residue connectivity, via the carbonyl carbon significantly increases the number of residues assigned. However, near-complete

assignments can only be obtained by the inclusion of NOESY data and information from residue specific labeling or residue-specific pulse sequences. Finally, the combination of inter-residue connectivity from Cγ chemical shifts (McCallum, *et. al*, 1998) with C$_\alpha$, C$_\beta$, and CO connectivities can provide sufficient information for extensive assignments, even when a large number of amide resonances are missing from the spectrum.

The data requirements for using known assignments to aid in the assignment of another form of the protein, such as a protein-ligand complex (see McCallum et al., 2000), was also investigated. Our calculations show that prior assignments from the native form of a protein are a rich source of information for the assignment of the modified form of the protein; inter- and intra-residue C$_\alpha$ chemical shifts, along with inter-residue CO shifts appear to be sufficient to obtain reliable assignments, even when a large fraction of spin-systems are missing due to the absence of amide resonances.

Methods

Test data

The N-terminal domain of enzyme I of the phosphoenolpyruvate:sugar phosphotransferase system of E. coli (EIN) was used as a source of experimental chemical shifts (Garrett et al., 1997, BMRB accession #4106). To generate a complete data set, the small number of chemical shifts that were missing from the deposited data were generated from the known distributions of chemical shifts as provided by the BioMagResBank (Seavey et al., 1991). Data representing inter-residue crosspeaks were constructed from this chemical shift list using a uniform random distribution to generate chemical shifts for the inter-residue crosspeaks from the intra-residue chemical shifts. The random distribution was 0.4 ppm wide in the case of ^{15}N shifts, 0.4 ppm for carbon C$_\alpha$, C$_\beta$, and Cγ shifts, 0.3 ppm for CO shifts and 0.1 ppm for H$_N$ protons. To investigate the effects of the inability to back-exchange amide protons a fraction (9% or 15%) of spin-systems were removed from the data on the basis of their solvent accessibility calculated from the crystal structure (Connolly, 1993). Note that missing amide signals may also occur as a result of chemical exchange on the intermediate time-scale even if the amide deuteron can be exchanged with solvent protons. The effects of experimental signal-to-noise on the ability to observe resonance peaks was simulated by randomly deleting chemical shifts within spin-systems. If a crosspeak was deleted from one experiment then the associated crosspeak in the less sensitive NMR experiment that follows the same magnetization pathway was also deleted. For example, the deletion of an inter-residue peak in the HNCA would also result in the deletion of the peak in the HNCB experiment as well. In general, magnetization transfer pathways that utilized the carbonyl group were considered to be more efficient. Thus fewer peaks were discarded from the HN(CO)CA experiment than the HNCA experiment.

NOE cross peaks were generated from the known tertiary structure of EIN (Liao et al., 1996) using inter-proton distances of 4.5 Å. NOE crosspeaks that were generated from an extended chain are termed *primary* (1°) NOEs. NOE crosspeaks that were generated from the known three-dimensional structure of the protein are referred to as *tertiary* (3°) NOEs. The actual chemical shifts of the NOE crosspeaks were generated from uniform random distributions using the widths of the distributions for the ^{15}N-shifts and H$_N$ shifts given above. NOE crosspeaks that involved missing residues were deleted from the data set. In addition, 15% of the potential NOE crosspeaks were randomly removed, without regard to inter-proton distance, to simulate the effects of missing data due to peak overlap and spectral artifacts.

The identification of the amino-acid type of a spin-system provides valuable information to the assignment process, and the utility of such information was also investigated here. Residue type identification can be obtained in two ways, the generation of samples with residue-specific labels or residue-selective NMR experiments on uniformly labeled material. Selective isotopic labeling with ^{15}N at the amino position has been used by a large number of groups to identify the residue type of the amide peak and the technique has been reviewed by McIntosh and Dahlquist (1990). Residue specific labeling with ^{13}C at the carbonyl position can also be used to identify the residue type (see McCallum et al., 1999). In these experiments the protein is uniformly labeled with ^{15}N and the minimal growth media is supplemented with the 1-^{13}C specifically labeled amino acid (^{13}CO labeled). Use of a ^{13}C-carbonyl filtered HSQC experiment identifies amide peaks of residues that directly follow the ^{13}CO-labeled amino acids. Although this method of ^{13}C-carbonyl labeling was used for resonance assign-

ments in early NMR studies (Takahashi et al., 1991; Burk et al., 1989; Griffey et al., 1986), it has not found broad use. This is surprising since residue-type identification using ^{13}CO labeling is considerably more robust than ^{15}N residue-specific labeling because the carbonyl position is not readily altered during amino acid metabolism. In contrast, the amide group of many amino acids is readily removed in wild-type *E. coli*, necessitating the use of transaminase deficient strains (McIntosh and Dahlquist, 1990).

In this study, simulated resonance peak positions that would appear in ^{13}C-carbonyl filtered HSQC spectra from residue specific ^{13}CO labeled samples were generated by altering, using the above random distributions, the proton and nitrogen chemical shifts of the amide of the residue that follows the ^{13}C-labeled residue. Amide resonances that were identified by specific labeling were Ala, Val, Leu, Phe, Tyr, Pro. These were selected because of the low cost of these 1-^{13}C labeled amino acids. Fifteen percent of the possible peaks were removed because it may not be possible to observe all of the peaks in these samples due to low signal-to-noise or slow deuterium-hydrogen exchange rates.

A number of residue-selective NMR experiments have been described in the literature. Oschkinat and co-workers have described a suite of experiments that utilize coherence editing in conjunction with selective pulses to identify the residue-type of spin-systems (Schubert et al., 1999, 2001). Although these experiments are directed at protonated samples, they can be used to identify the amide resonances of Asn and Gln residues in deuterated proteins. Wagner and co-workers have also presented a number of triple resonance experiments for the identification of residue-type (Dötsch et al., 1996a,b; Dötsch and Wagner, 1996). Of these, the β-carbon edited HN-COCACB (Dötsch et al., 1996a) can be applied to deuterated proteins to identify residues that lack a C_γ. The HN(COCACB)CG experiment described by McCallum et al. (1998) provides both the chemical shift of the C_γ carbon as well as identifies residues that lack a C_γ carbon. Kay and co-workers have designed pulse sequences for the identification of methyl-containing residues for samples with fully protonated (Gardner et al., 1996) or partially deuterated methyl protons (Muhandiram et al., 1997). Additional carbon chemical shifts, such as those obtained from carbon TOCSY experiments (Gardner et al., 1996), provide a rich source of information for identification of residue type. However, we have not utilized information of this type in this study because of the low sensitivity of the carbon TOCSY experiments for larger deuterated proteins (see McCallum et al., 1998). Rather, we have elected to use information from specifically 1-^{13}C labeled amino acids to obtain unambiguous residue-specific assignments.

Chemical shift changes that are a result of ligand binding were simulated using data from the complex between EIN and the histidine-containing phosphocarrier protein HPr (Garrett et al., 1999). In this case, amide resonances that were absent in the un-liganded sample were also deleted in the liganded sample. However, different inter- and intra-residue crosspeaks were randomly removed for the liganded protein data set.

Chemical shift assignments

Monte Carlo methods, as implemented in the program MONTE (Hitchens et al., 2002), were used to obtain resonance assignments. The program begins by generating a random mapping of spin-systems to the primary sequence and computing an assignment score. The program then interchanges two segments consisting of one or more consecutive assignments and evaluates the score of the new mapping of the spin-systems. The new arrangement of spin-systems is kept if the score improves. If the score is worse, the new arrangement may still be accepted, depending on the 'temperature' of the system. During the assignment process, the 'temperature' of the system is gradually lowered such that newly generated spin-system arrangements that are less favorable than the previous sequence are accepted at increasingly lower frequencies. The annealing schedule that was used in this study is shown in Table 1.

The score associated with any particular mapping of spin-systems onto the primary sequence was determined by summing the contributions from inter-residue connectivity, residue type prediction from chemical shifts, predicted NOE crosspeaks, and residue specific labeling (see Hitchens et al., 2002). The contribution, or weight, of each of the above types of information to the overall score can be adjusted by the user. In this study, the relative weights of the different contributions were selected to give the correct assignment solution when connectivity and chemical shift information from C_α and CO shifts were used in conjunction with tertiary NOEs and specific labels (i.e., Table 2, row 8, left block). For consistency, these weights were used for all of the calculations, however they may not have been optimal for some combi-

Table 1. Monte Carlo annealing parameters

Segment	T-start	T-end	T-step	N-swap	γ	Swap size	NOE score	Inter-residue repulsive terms		
								CO	C_α	C_β
1	300	150	10	20 000	1	2	0.0	0	0	0
2	200	10	10	100 000	1	2	0.1	10	10	10
3	130	10	5	150 000	2	3	0.2	20	20	20
4	120	5	5	200 000	3	4	0.3	30	30	30

Segment, annealing segment for one complete cycle; T-start, starting temperature of annealing schedule; T-end, ending temperature of annealing schedule; T-step, decrement in temperature; N-swap, initial number of swaps performed at each temperature; γ, during the annealing, N-swap is increased as the temperature drops, the increment is proportional to e^{γ}, giving approximately 10^6 cycles at the lowest temperature; Swap size, maximum number of consecutive residues that can be swapped; NOE scale, weighting factor for NOE connectivity information; Repulsive terms (CO, C_α, C_β), this is the size of the repulsive term for inter-residue chemical shift matching. For example, during the last annealing segment a poorly matched inter-intra chemical shift will be given a score of -30 while identical inter- and intra-residue shifts would receive a score of $+100$. The width of the Gaussian distributions were 0.14, 0.05, 0.20, and 0.30 ppm for the matching of N_H, HN, CO, C_α and C_β shifts, respectively.

The first two annealing segments served mainly to ensure that the spin-systems were completely randomized. Correct assignments were established towards the end of the third and fourth segments. Since the scoring function changes with each segment it is necessary to increase the temperature at the beginning of each segment to insure that the system will be at equilibrium during annealing. The number of swaps were chosen such that the final assignments were invariant to changes in the number of swaps. Usually, the solution converged with a two to four fold reduction in the number of cycles, as evidenced by a lack of dependence of the final solution on the number of cycles and small changes of less than 1 part in 10^3 in the score towards the end of the segment. Five independent solutions, or cycles, were obtained for each case.

Table 2. Data requirements for assignment using C_α and CO chemical shifts

	Data					All residues present				Exchange 1 (9% missing)				Exchange 2 (15% missing)			
	C_α	CO	1°	3°	sl	Correct/ unique	(#)	Ave. ambig.	Error (#)	Correct/ unique	(#)	Ave. ambig.	Error (#)	Correct/ unique	(#)	Ave. ambig.	Error (#)
1	•			•		121	16	3.4	0	109	23	3.6	1	84	24	3.7	0
2	•			•	•	224	131	2.0	0	171	89	3.1	1	159	67	3.4	2
3	•	•				77	33	4.4	0	65	19	4.2	0	65	22	4.3	0
4	•	•	•			123	10	3.4	0	48	16	4.4	0	80	14	4.1	0
5	•	•			•	239	196	2.4	0	228	143	2.3	4	176	72	3.0	0
6	•	•	•		•	258	162	1.7	0	226	103	1.8	0	225	126	2.7	0
7	•	•		•		258	209	1.7	0	252	205	1.4	2	256	163	1.8	2
8	•	•		•	•	258	257	1.0	0	256	250	2.0	2	256	243	1.5	2

Bullets in the two left-hand columns indicate the inter-residue connectivity data was used in the assignment process. Columns marked with 1^o, 3^o, and sl refer to primary NOEs, tertiary NOEs, and information from ^{13}CO labeled samples, respectively. The remaining three blocks of the table gives statistics for the assignments when all the spin-systems are present (left block), 9% are missing (middle block) and when 15% are missing (right block). The first sub-column under Correct/Unique gives the number of correct assignments that were found in the best solution. The second sub-column gives the number of correct assignments that were found in *all* five independent runs of the program. In both cases these numbers refer to all of the residues present in the protein (258). In the case where spin-systems were missing the assignment is considered correct if a spin-system was not associated with the missing residue. The Average ambiguity (Ave. ambig.) column is the average, over all of the residues, of how many different spin-systems were assigned to a particular residue. An ambiguity of 1 means that all five of the five assignments were the same. An ambiguity of five indicates that the assignment of a residue was different in all five runs of the program. The Error column indicates the number of incorrect assignments that were undetectable in these calculation because the same incorrect assignment was found in all five runs of the program.

nations of data (see below). The score associated with inter-residue connectivities (e.g. C_α shifts from HN(CO)CA and HNCA experiments) was calculated using the following widths for the Gaussian scoring function (Hitchens et al, 2002): N, 0.14 ppm; NH, 0.05 ppm; C_α, 0.2 ppm, C_β, 0.3 ppm, C_γ, 0.1 ppm. Note that these ranges are smaller than the random variation in chemical shifts that were introduced during the initial generation of the data. A larger range of chemical shifts was chosen for the initial generation of the inter-residue crosspeaks to account for apparent changes in experimentally measured shifts that may arise from limited digital resolution or ill-defined peak positions from weak signals in experimental spectra.

In the simulations presented here a total of five independent chemical shift assignments, or cycles, were generated for each set of the data sets used in this study. An assignment was deemed correct if the correct placement of the spin-system in the primary sequence was observed in the best scored solution. An assignment was considered to be unique if the same correct placement of the spin-system was observed in all five runs of the program. An assignment was considered to be in error if it involved a unique, but incorrect, placement of the spin-system onto the primary sequence. Note that with additional cycles an incorrect assignment may be converted to an ambiguous assignment.

Results and discussion

Assignments utilizing C_α and CO connectivity information

Table 2 shows the results for assignments based solely on C_α data as well as for assignments based on combined connectivity information from primary and tertiary NOEs, ^{13}CO residue specific labeling, and CO chemical shifts and inter-residue connectivities. These calculations assumed that *all* the inter- and intra-residue C_α and CO crosspeaks are present.

As expected, inter-residue connectivities based only on C_α chemical shifts yielded few correct assignments since there is insufficient residue type information in C_α chemical shifts (not shown). The use of information from primary NOEs did not cause a significant increase in the number of assignments (not shown). As evident from the first two rows of Table 2, it is only possible to obtain a significant fraction (224/258) of correct assignments with C_α connectivity alone if all the residues are observed in the NMR spectra and the data are supplemented with information from tertiary NOEs as well as residue specific labeling (row 2, left block).

The absence of amide resonances, and their associated spin systems, reduces the number of correct assignments. For example, under the best conditions (C_α connectivity, tertiary NOEs, and residue specific labeling) only 159/258 assignments are obtained when 15% of the spin-systems are missing (row 2, right block). Of these, only 25% (67/258) are unique. In addition, two errors in the assignments are observed. Although it is likely that these erroneous assignments would become classified as ambiguous if more than five trials had been performed, some errors of this type may persist even with a large number of independent assignment trials. In summary, when amide resonances are missing it is not possible to obtain a large number of reliable assignments using solely C_α connectivity unless additional information, such as tertiary NOEs and residue specific data are also employed in the assignment process.

The addition of carbonyl inter-residue connectivities to C_α connectivities increases the overall reliability of sequentially connected spin-systems (compare rows 1 and 7 in Table 2). However, the addition of the CO chemical shifts by themselves does not cause marked improvements in assignments, largely because carbonyl chemical shifts provide little information on residue-type (see Table 2, row 3). The ambiguity in the assignment is largely due to the difficulty in uniquely placing long sequences of connected residues on to the primary sequence. This problem is illustrated by a number of long off-diagonal lines in the correlation plot shown in Figure 1A. The use of primary NOEs improves the situation, but fewer than 50% of the correct assignments are obtained in this case (Table 2, row 4). In contrast, the addition of information from residue specific labeling causes a dramatic improvement in the number of correct assignments (Figure 1B; Table 2, row 5), producing a nearly complete solution (239/258 residues) when there are no missing spin-systems. A substantial number of correct assignments (176/258) are obtained even when 15% of the spin-systems are absent. The increase in the reliability of the assignments is evident from the increased number of unique assignments, as shown by the number of black diagonal elements in Figure 1B. The combined use of primary NOE data and residue specific labeling (Table 2, row 6; Figure 1C further increases the number and reliability of the assignments.

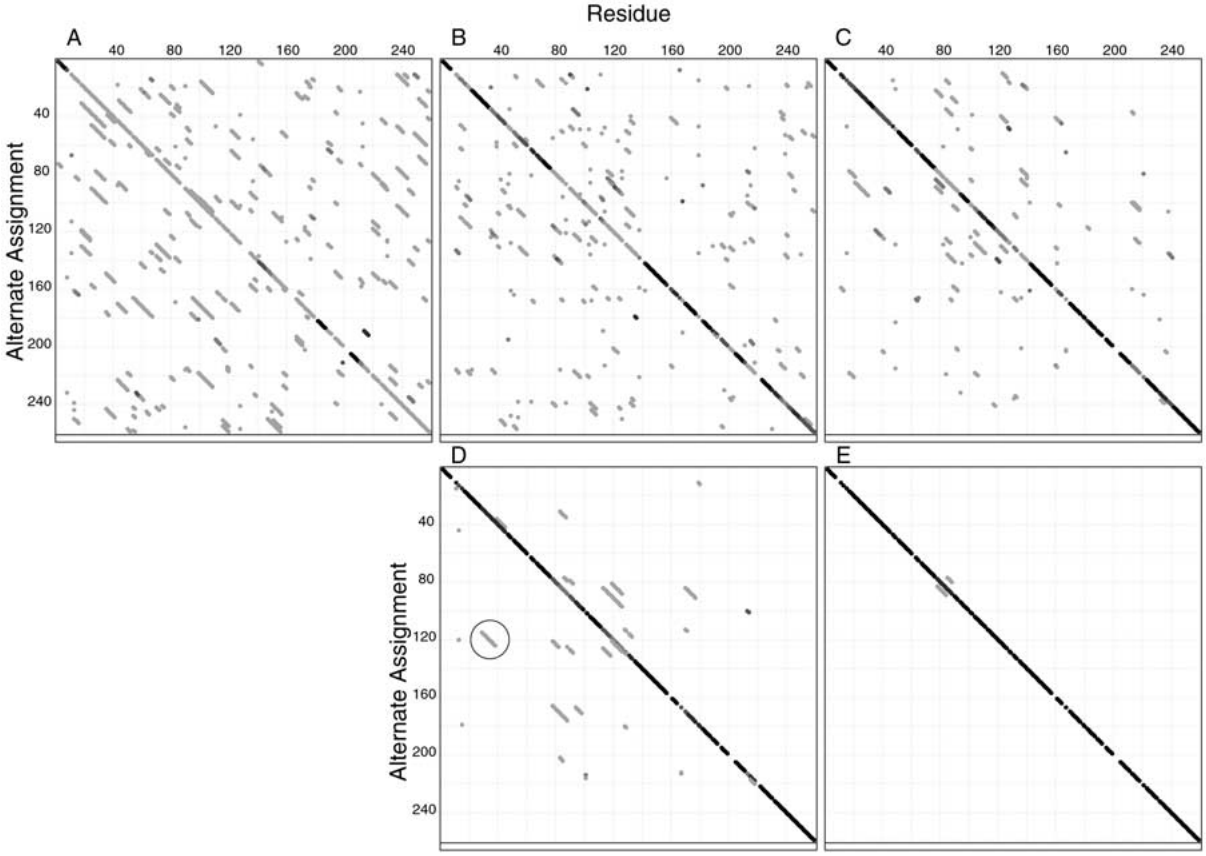

Figure 1. Uncertainty in assignments using C_α and CO data. The horizontal axis in each graph gives the residue number, the vertical axis indicates alternative assignments for a stretch of connected spin-systems. Black dots indicate the assignment is found in all five solutions. Lighter gray dots indicate that the assignment is not unique. Dots off of the diagonal indicate alternative placements for connected spin systems. For example, the dots present in the circled region in (D) indicate that the spin-systems assigned to residues 35–40 were also found to be assigned to residues 117–122 in some of the assignment solutions. All five panels correspond to the assignments obtained when 15% of the residues are missing. (A) utilized only Cα and CO data. Information from CO specific labeling was added to produce (B). Primary NOE data and CO specific labeling were combined with Cα and CO connectivities to produce the data shown in (C). (D) utilized tertiary NOE data while panel E utilized both tertiary NOE data as well as CO specific labels. Details of the assignments for (E) are shown in Figure 2.

When C_α and CO connectivity information is supplemented with tertiary NOE data (Table 2, row 7) only a few segments of connected spin-systems show alternative positions within the primary sequence, as shown in the correlation plot presented in Figure 1D. Finally, the combined presence of tertiary NOEs and residue specific ^{13}CO labeling restricts the uncertainty in the assignments to a small segment between residues 78 and 88 (Figure 1E). In this case, the proximity of the off-diagonal points to the diagonal indicates that short segments of spin-systems can be placed at alternative positions within a close segment of the primary sequence. The best assignment solution that corresponds to the data presented in Figure 1E is shown in Figure 2. This figure indicates that with the exception of the segment from 78 to 88, most of the assignments are reliable.

Although the addition of information from tertiary NOEs and residue-specific labeling caused a general improvement in the assignment solutions, this information also generated two errors that were not present when less information was used. Both of these errors involve the interchange of two residues, Phe_{11} and Lys_{13}, which are disconnected from their neighbors by intervening residues that lack observable amide resonances. In this case the C_α chemical shift information is insufficient to distinguish one residue from the other and all tertiary NOEs are satisfied. The only way to resolve this error would be to collect data from additional residue-specific samples, acquire C_β shifts,

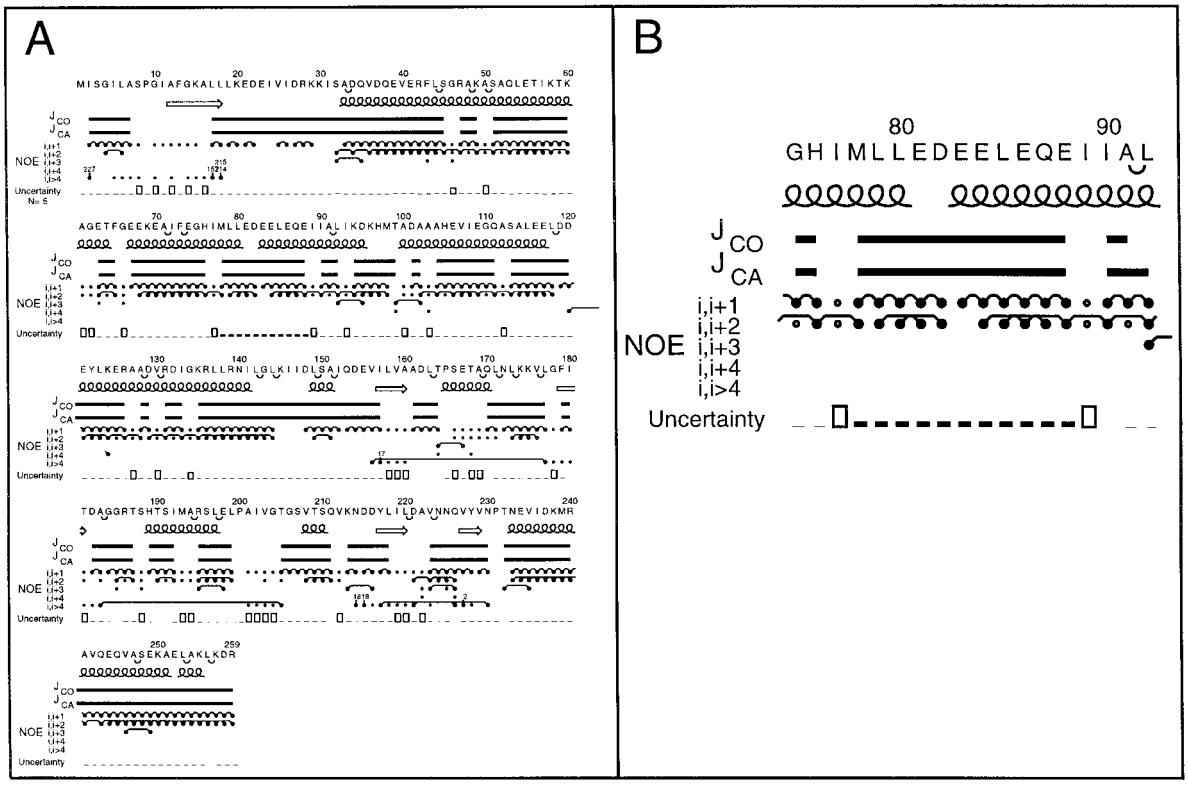

Figure 2. Summary of chemical shift assignments using C_α and CO Data. The left panel (A) shows the complete assignment solution. (B) is an enlarged view of residues 75–92. These assignments were obtained with 15% of the residues missing. Complete C_α and CO connectivity as well as tertiary NOE and CO specific labeling were used to obtain the assignments. These conditions correspond to the last block of data in the lower right-hand corner of Table 2. The half-circles shown immediately underneath the amino acid sequence indicates that this assignment agrees with the information provided from residue specific labeling. The first row beneath the primary sequence indicates the secondary structure of the protein. Inter-residue connectivities involving the CO and C_α chemical shifts are shown in the next two rows. The subsequent five rows shows NOEs, large filled circles that are joined by lines are present in the data. Numbers above large filled circles represent long range NOEs to the indicated residue. Small empty circles indicate that an NOE is possible based on the structure, but not detected in the data. The last line in each segment indicates the uncertainty in the assignment for the five (N = 5) independent runs. The height of the filled bars indicates the number of different spin-systems that were assigned to that residue. Residues 78 to 89 show some uncertainty in assignment while the rest of the residues are unambiguously assigned. In contrast to the normal output from MONTE, the empty boxes in this line indicate residues that are missing due to slow amide deuteron exchange. Normally, this symbol is reserved for proline residues.

or acquire data on partially deuterated systems (see McCallum et al., 1999).

In summary, if near complete C_α and CO connectivity information is supplemented with information from primary NOEs plus residue-specific labels it should be possible to assign the bulk of the spin systems in large deuterated proteins. If the structure of the protein is known, such that tertiary NOEs can be predicted, then C_α and CO connectivity in combination with the NOEs appears to be sufficient to arrive at a satisfactory assignment solution without the need to obtain residue specific data.

Addition of C_β chemical shifts

The assignment of deuterated proteins generally employ C_β shifts to provide additional inter-residue connectivity and to enhance the prediction of residue-type from the measured chemical shifts. In any given experiment the number of experimentally detected C_α and C_β shifts will depend on many factors, such as protein size and data quality. In the calculations discussed here, the threshold level of data required for reliable assignments was determined by varying the number of available C_α and C_β shifts. In the case where 95% of C_α and 85% of C_β inter-residue shifts are present, and all of the amide resonances are observable, complete assignments were obtained (results not shown).

Table 3. Data requirements for assignments using C_α, C_β, and CO chemical shifts

	Data							All residues present				Exchange 1 (9% missing)				Exchange 2 (15% missing)			
	C_α	C_β	C'	C_γ	1°	3°	SL	Correct/ unique	(#)	Ave. ambig.	Error (#)	Correct/ unique	(#)	Ave. ambig.	Error (#)	Correct/ unique	(#)	Ave. ambig.	Error (#)
A1	•	•	○					200	151	3.3	0	168	101	3.3	7	140	80	3.3	9
2	•	•	○		•			239	174	1.8	0	216	157	2.6	8	196	114	2.5	7
3	•	•	○		•		•	252	218	1.7	0	228	172	2.7	8	187	135	2.8	10
4	•	•	○			•		258	244	1.0	0	249	249	0.0	9	248	226	2.6	5
5	•	•	○			•	•	258	257	1.0	0	256	249	3.2	0	241	235	3.0	3
B1	•	•	•					253	212	1.6	0	235	157	2.0	4	209	158	2.8	4
2	•	•	•		•			253	215	1.6	0	243	195	2.3	5	228	177	2.3	9
3	•	•	•		•		•	258	245	1.1	0	245	238	2.5	14	239	203	2.3	7
4	•	•	•			•		258	257	1.0	0	250	250	0.0	8	248	248	0.0	10
5	•	•	•			•	•	258	257	1.0	0	258	252	3.0	0	256	250	2.0	2
C1	•	•	•	•				258	244	1.7	0	241	197	2.1	0	233	177	2.0	1
2	•	•	•	•	•			258	235	1.3	0	243	220	2.0	3	234	191	2.7	2
3	•	•	•	•	•		•	258	257	1.0	0	254	226	1.4	1	249	215	1.9	3
4	•	•	•	•		•		258	246	1.8	0	258	248	1.8	0	256	235	1.9	1
5	•	•	•	•		•	•	258	256	1.0	0	258	248	3.0	0	256	247	1.4	1

A bullet in the C_α, C_β, CO, and Cγ columns indicates that these data were used for inter-residue connectivity as well as for residue type prediction. Part A of the table shows results from C_α and C_β inter-residue connectivities. Part B adds inter-residue connectivities via the carbonyl carbon. Part C adds connectivities via the gamma carbon. A circle in the CO column indicates that the data were used for residue-type prediction only. The remaining columns in this table have the same meaning as those in Table 2. The data sets used in this table contained 90% and 95% of the intra- and inter-residue C_α shifts, 80% and 85% of the intra-and inter-residue C_β shifts, 50% and 70% of the intra- and inter-residue Cγ shifts, and 70% and 100% of the intra- and inter-residue CO shifts. In addition, only 85% of the possible inter-proton NOEs were present.

Consequently, calculations were performed below this threshold level of missing C_α and C_β resonances to assess the effect of providing additional information to the assignment process (see legend to Table 3).

The effect of missing spin-systems on the number of correct assignments was dramatic. Simply missing 9% of the spin-systems reduced the number of correct assignments from 200 to 168 (Table 3, row A1). A disturbing observation is that a number of errors in the assignments are also found. These errors arise from the placement of connected spin-systems onto alternative regions of the primary sequence because of a better matching of the C_α and C_β chemical shifts of the spin system to the characteristic chemical shifts of the incorrect residues. The addition of primary NOE data, as well as information from residue-specific labeling, increases the number of correct assignments and decreases the average ambiguity of the non-unique assignments (Table 3, rows A2 and A3). However, this additional information does not remove the assignment errors that occur when spin-systems are missing. The presence of tertiary NOEs, but no residue-specific labels, further increases the number of correct assignments (row A4). Most of these errors are resolved when residue-specific labels are used. A small number of errors (3/258) persist when 15% of the amide spin-systems are missing (Table 3, row A5, right block)

The inclusion of primary NOE data and/or residue specific ^{13}CO labeling appears to cause a small increase the number of errors, as shown by a comparison of rows A2 and A3 (right block of Table 3). A similar phenomenon also occurs with C_α, C_β, and CO connectivities (rows B2 and B3, middle section), and to a smaller extent with C_α, C_β, CO, and C_γ connectivities, as shown in the right blocks of rows C2 and C3 in Table 3. Close inspection of these solutions suggest that this increase in error is due to two factors. The weighting of information from residue specific data or the primary NOE information does enhance the stability of incorrect assignments during the Monte Carlo run, converting previously ambiguous assignments to incorrect assignments. However, the largest contribution to the number of errors appears is simply due

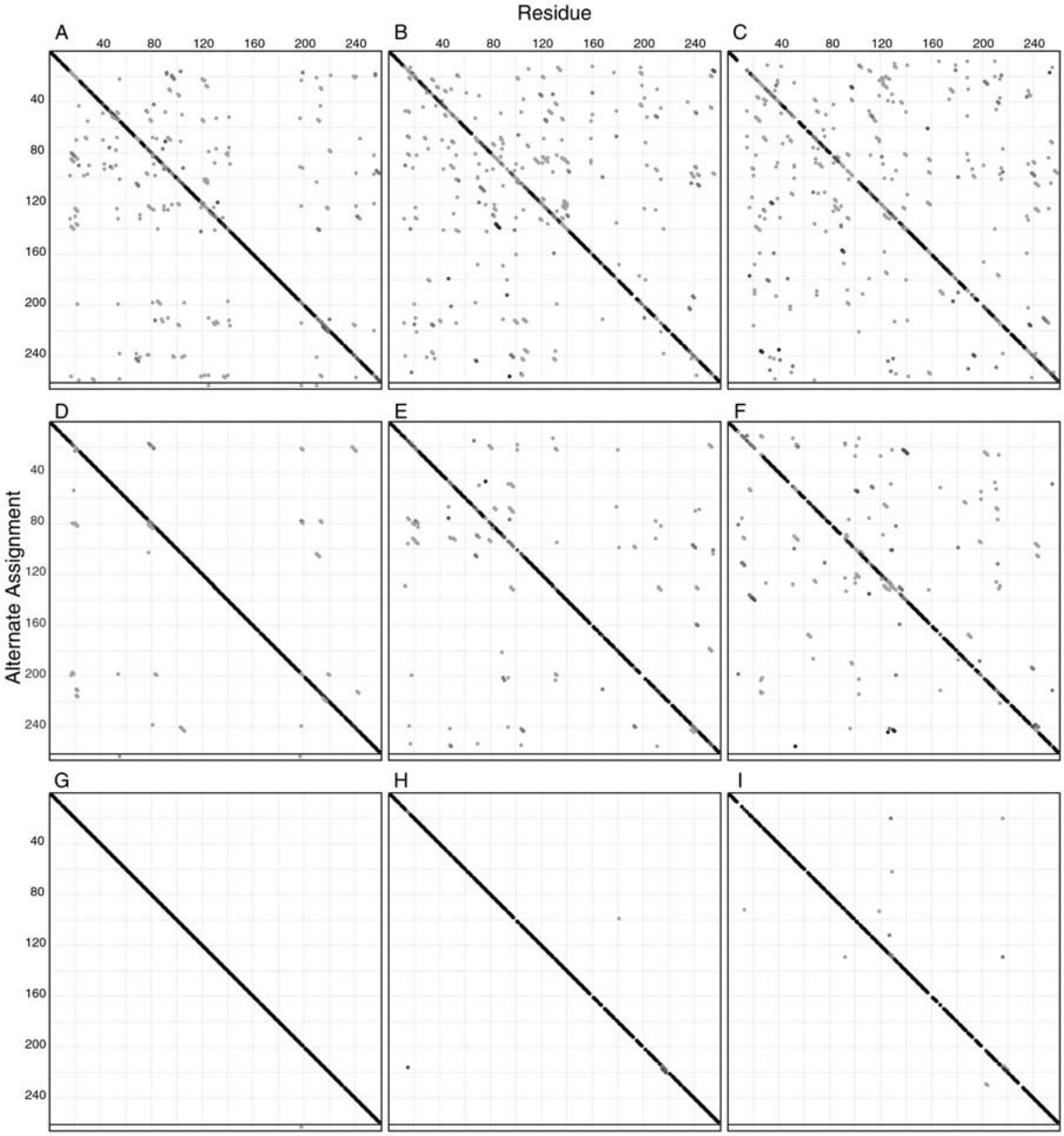

Figure 3. Effect of NOEs and specific labeling on assignments generated from C_α and C_β chemical shifts and inter-residue connectivity. See the legend for Figure 1 for an explanation of these plots. The columns in this figure represent assignments that were obtained with all residue present (left column, (A), (D), and (G)), 9% of the residues absent (middle column, (B), (E), (H)), and 15% of the residues absent (right column, (C), (F), and (I)). The rows show the effect of the inclusion of data on the uncertainty of the assignments. The top row using only C_α and C_β information and corresponds to line A1 of Table 3. The second row shows the effect of including primary NOE data and residue specific CO labeling (Table 3, line A3). The last row in this figure shows the effect of including tertiary NOE data and residue specific CO labeling. This corresponds to line A5 in Table 3.

Table 4. Use of known chemical shifts in the assignment process

Peaks matched			All residues present				Exchange 1 (9% missing)				Exchange 2 (15% missing)			
HNCA	HNCOCA	HNCO	Correct/ unique	(#)	Ave. ambig.	Error (#)	Correct/ unique	(#)	Ave. ambig.	Error (#)	Correct/ unique	(#)	Ave. ambig.	Error (#)
•	•		247	233	2.8	1	236	208	2.5	5	231	214	2.7	9
•	•	•	258	252	1.0	0	256	244	1.5	2	252	239	1.6	2

In these calculations inter-residue connectivity was obtained solely from inter- and intra-residue C_α chemical shifts. In addition, the similarity of the peak position in the liganded and unliganded spectra for the HNCA (C^i_α), HN(CO)CA ($C^{(i-1)}_\alpha$) and HNCO (CO) experiments was also used in the assignment process. The appropriate column is marked by a bullet if these data was used in the assignment process. The contribution of each potential assignment to the overall score was evaluated by comparing all three known chemical shifts (e.g. H_N, N_H,C_α) of a residue to the chemical shifts of the spin-system that was mapped to the residue during the assignment process. If all three of these shifts were within the tolerance for matching chemical shifts then the trial assignment received a high score. The score decreased as the differences in the chemical shifts between the known assignment and the trial assignment increased. Chemical shift differences that were more than 3 times the tolerance essentially received a score of zero.

to the number of independent runs that were used in these calculations. For example, multiple five-cycle calculations using C_α, C_β, and CO with primary NOE and specific labeling data give the following statistics when 9% of the amides are absent (Table 3, row B3, middle block): Correct $= 247.6 \pm 3.3$, Unique $= 239.7 \pm 1.3$, Average Ambiguity $= 2.2 \pm 0.4$, Error $= 6 \pm 5$. The number of errors can be exaggerated in these calculations, while the number of correct and unique assignments are relatively insensitive to the number of independent cycles.

The type of alternative solutions that are obtained using C_α and C_β connectivity information is summarized in correlation plots presented Figure 3. A comparison of Figure 1 and Figure 3 shows that alternative assignments occur over shorter segments of linked spin systems when C_β data is used for assignments. A large number of alternative assignments are observed in the absence of NOE or residue-specific label information (Figure 3, top panels). The addition of primary NOEs and residue-specific labels causes an overall improvement when all residues are present (Figure 3D), but a substantial number of alternative assignments occur when 15% of the amide spin-systems are missing. Inclusion of information from tertiary NOEs and residue specific ^{13}CO labels removes almost all alternative assignments, as indicated by the small number of off-diagonal points in the bottom row of Figure 3 and the large number of correct assignments (row A5).

In summary, reliable assignments can be obtained from C_α and C_β information if most of the inter-residue connectivities are present and none of the spin-systems are missing in the NMR spectrum. However, if as few as 9% of the spin-systems are missing then it may be necessary to supplement the inter- and intra residue C_α and C_β chemical shifts with tertiary NOE data as well as some residue-specific data.

Addition of CO chemical shifts

A higher fraction of correct assignments are obtained if inter-residue connectivity information from CO chemical shifts is added to the information provided by C_α and C_β shifts (see Table 3, rows B1–5). The average ambiguity drops considerably (e.g., from 3.3 to 1.6 in the case when all residues are present). However, as with the case of using solely C_α and C_β information, significantly fewer residues are assigned when spin-systems are missing. When amide signals are absent, the addition of primary NOE data offers marginal improvements to the characteristics of the assignment (row B2). Although additional spin systems are assigned, there may be more errors in the assignment (compare rows B1 to B2). When 15% amides are absent, near complete assignments can be obtained if C_α, C_β, and CO shifts are supplemented with either primary NOEs plus specific labeling (row B3), tertiary NOEs alone (row B4), or tertiary NOEs with specific labeling (row B5). Only the latter combination permits a relatively error free solution with a small number of assignment cycles.

Addition of C_γ chemical shifts

Connectivity and residue-type information can also be obtained from Cγ shifts (McCallum et al., 1998). Since the experiments that elucidate Cγ shifts are of low sensitivity we have modeled the experimental data by assuming that only 50% of the inter-residue cor-

relations are observed. As shown in Table 3 (part C) the inclusion of Cγ shifts increases the number of correct assignments. However, it is still difficult to obtain complete assignments when spin-systems are absent from the data. As in the case of C_α and C_β information, the addition of primary NOE data and the residue specific labeling improves the number of correct assignments, but may introduce some errors. The inclusion of tertiary NOE and residue specific labeling produces excellent results, providing near complete resonance assignments in all cases with a small number of errors.

Effect of incorporating known assignments

Chemical shift assignments can be facilitated by using known chemical shifts of one form of a protein (e.g. unliganded) to aid in the assignments of another form (see McCallum et al., 1999; Hitchens et al., 2001). Since the structural homology between the unliganded and liganded-protein is high, the chemical shifts in both forms will be similar and can be used to guide the assignment process. Surprisingly, very little information is required to assign protein-ligand complexes when previous assignments are available. Table 4 shows that C_α chemical shifts and inter-residue connectivities are sufficient if all of the amide residues are present in the spectrum. If a significant fraction of residues are missing, then HNCO chemical shifts may be necessary.

Conclusions

The results presented here show that care needs to be taken when limited inter-residue connectivity data and residue type identification are available for the assignment of large deuterated proteins. In particular, problems can arise if even a small number of spin-systems are missing from the data. Alternative assignment solutions become prevalent and a number of errors in the assignments may occur. The influence of missing spin-systems on assignments will depend on the nature of the available data and the method that is used to obtain the assignments. In particular, it is clear that a substantial number of errors will occur when assignment schemes are used that are based largely on a small number of inter-residue connectivities (Bhavesh et al., 2001) .

Although synthetic data sets were used in the work presented here, the observed trends have been noted in the assignment of several large dimeric proteins (McCallum et al., 1999; Hitchens et al., 2001). Clearly, the actual extent of experimental data required for assignments will depend on the specific properties of the protein under study. Nonetheless, the results presented here should serve as a useful guide as to the type and quality of data required for reliable chemical shift assignments.

Acknowledgements

This work was supported by NIH (GM61253) and the Eberly Chair in Structural Biology (GSR), and by a National Institutes of Health NRSA fellowship to T.K.H. We thank J.A. Lukin for his continued contributions to the Monte Carlo assignment program.

References

Atreya, H.S., Sahu, S.C., Chary, K.V.R. and Govil, G. (2000) *J. Biomol. NMR*. **17**, 125–136.

Bailey-Kellogg, C., Widge, A., Kelley, J.J., Berardi, M.J., Bushweller, J.H. and Donald, B.R. (2000) *J. Comput. Biol*, **7**, 537–58.

Bax, A. and Grzesiek, S. (1993) *Accounts Chem. Res.*, **26**, 131–138.

Bax, A.., Kontaxis G. and Tjandra, N. (2001) *Methods Enzymol.*, **339**, 127–174.

Bhavesh, N.S., Panchal, S.C. and Hosur, R.V. (2001) *Biochemistry*, **40**, 14728–14735.

Burk, S.C., Papastavros, M.Z., McCormick, F. and Redfield, A.G.(1989) *Proc. Natl. Acad. Sci. USA*, **86**, 817–820.

Clore, G.M. and Gronenborn, A.M. (1997) *Nat. Struct. Biol. NMR Suppl.*, 849–853.

Connelly, M.L. (1993) *Science*, **221**, 709–713.

Dötsch, V. and Wagner, G. (1996) *J. Mag. Res.*, **B111**, 310–313.

Dötsch, V., Matsuo, H. and Wagner, G. (1996a) *J. Mag. Res.*, **B112**, 95–100.

Dötsch, V., Oswald, R.E. and Wagner, G. (1996b) *J. Mag. Res.*, **B110**, 107–111.

Gardner, K.H., Konrat, R., Rosen, M.K. and Kay, L.E. (1996) *J. Biomol. NMR*, **8**, 351–356.

Garrett, D.S., Seok, Y.J., Liao, D.I., Peterkofsky, A., Gronenborn, A.M. and Clore. G.M. (1997) *Biochemistry*, **36**, 2517–2530.

Garrett, D.S., Seok, Y.J., Peterkofsky, A., Gronenborn, A.M. and Clore, G.M. (1999) *Nat. Struct. Biol.*, **6**, 166–173.

Griffey, R.H., Redfield, A.G., Mcintosh, L.P., Oas, T.G. and Dahlquist, F.W. (1996) *J. Am. Chem. Soc.*, **108**, 6816–6817.

Grishaev, A. and Llinas, M. (2002) *Proc. Natl. Acad. Sci. USA*, **99**, 6707–6712.

Güntert, P., Salzmann, M., Braun, D. and Wüthrich, K. (2000) *J. Biomol. NMR*, **18**, 129–137.

Hitchens, T.K., Lukin, J.A., Zhan, Y., McCallum, S.A. and Rule, G.S. (2002) *J. Biomol NMR*, **25**, 1–9.

Hitchens, T.K., Mannervik, B. and Rule, G.S. (2001) *Biochemistry*, **40**, 11660–11669.

Kay, L.E. and Gardner, K.H. (1997) *Curr. Opin. Struct. Biol.*, **7**, 722–731.

Liao, D.I., Silverton, E., Seok, Y.J., Lee, B.R., Peterkofsky, A. and Davies, D.R. (1996) *Structure*, **4**, 861–872.

McCallum, S.A., Hitchens, T.K. and Rule, G.S. (1998) *J. Magn. Reson.*, **134**, 350–354.

McCallum, S.A., Hitchens, T.K. and Rule, G.S. (1999) *J. Mol. Biol.* **285**, 2119–2132.

McCallum, S.A., Hitchens, T.K., Torborg, C. and Rule, G.S. (2000) *Biochemistry*, **285**, 2119–2132.

McIntosh, L.P. and Dahlquist, F.W. (1990) *Quart. Rev. Biophys.*, **23**, 1–38.

Moseley, H.N.B. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.

Muhandiram, D.R., Johnson, P.E., Yang, D., Zhang, O., McIntosh, L.P. and Kay, L.E. (1997) *J. Biomol. NMR*, **10**, 283–288.

Riek, R., Wider, G., Pervushin, K. and Wüthrich, K. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 4918–4923.

Sattler, M. and Fesik, S.W. (1996) *Structure*, **4**, 1245–1249.

Schubert, M., Oschkinat, H. and Schmieder, P. (2001) *J. Magn. Reson.*, **148**, 61–72.

Schubert, M., Smalla, M., Schmieder, P. and Oschkinat, H (1999) *J. Magn. Reson.*, **141**, 34–43.

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217–236.

Takahashi, H., Odaka, A., Kawaminami, S., Matsunaga, C., Kato, K., Shimada, I. and Arata, Y. (1991) *Biochemistry*, **30**, 6611–6619.

Wand, A.J. and Nelson, S.J. (1991) *Biophys. J.* **59**, 1101–1112.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley and Sons, New York, NY.

Zimmerman, D.E. and Montelione, G.T. (1995) *Curr. Opin. Struct. Biol.*, **5**, 664–673.

Zimmerman, D.E., Kulikowski, C.A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C., Powers, R. and Montelione, G.T. (1997) *J. Mol. Biol.* **20**, 592–610.