J Biomol NMR (2013) 55:201–209
DOI 10.1007/s10858-012-9702-x

ARTICLE

Chemical shift prediction for denatured proteins

James H. Prestegard · Sarata C. Sahu · Wendy K. Nkari · Laura C. Morris · David Live · Christian Gruta

Received: 11 October 2012 / Accepted: 23 December 2012 / Published online: 8 January 2013

Abstract While chemical shift prediction has played an important role in aspects of protein NMR that include identification of secondary structure, generation of torsion angle constraints for structure determination, and assignment of resonances in spectra of intrinsically disordered proteins, interest has arisen more recently in using it in alternate assignment strategies for crosspeaks in ^{1}H–^{15}N HSQC spectra of sparsely labeled proteins. One such approach involves correlation of crosspeaks in the spectrum of the native protein with those observed in the spectrum of the denatured protein, followed by assignment of the peaks in the latter spectrum. As in the case of disordered proteins, predicted chemical shifts can aid in these assignments. Some previously developed empirical formulas for chemical shift prediction have depended on basis data sets of 20 pentapeptides. In each case the central residue was varied among the 20 amino common acids, with the flanking residues held constant throughout the given series. However, previous choices of solvent conditions and flanking residues make the parameters in these formulas less than ideal for general application to denatured proteins. Here, we report ^{1}H and ^{15}N shifts for a set of alanine based pentapeptides under the low pH urea denaturing conditions that are more appropriate for sparse label assignments. New parameters have been derived and a Perl script was created to facilitate comparison with other parameter sets. A small, but significant, improvement in shift predictions for denatured ubiquitin is demonstrated.

J. H. Prestegard (✉) · S. C. Sahu · W. K. Nkari · L. C. Morris · D. Live · C. Gruta
Athens, GA, USA
e-mail: jpresteg@ccrc.uga.edu

Keywords Sparse labeling · Disordered proteins · Denatured proteins · NMR · Resonance assignments

Introduction

While we intend to focus here on the use of chemical shift prediction in the assignment of resonances in sparsely labeled proteins, chemical shift prediction has a long history in the NMR structural biology community, beginning with the identification of secondary structure in folded proteins (Wishart et al. 1995a). These early efforts have evolved into improved methods for secondary structure analysis (Camilloni et al. 2012), and they have become important contributors to restraints for torsion angles in protein structure determination (Shen et al. 2009). Chemical shift is also one source of structural information that is accessible in larger systems (Takeuchi et al. 2007), and it is becoming an increasingly important parameter in the facilitation of computational prediction of large protein structures using sparse data sets (Lange et al. 2012; Raman et al. 2010). To make use of chemical shifts in secondary or tertiary structure determination, good reference shifts from random coil or disordered regions of proteins are essential. This contributed some of the initial motivation for tabulating shifts and using the data to derive sequence dependent correction factors to random coil chemical shifts (Wang and Jardetzky 2002a; Wishart et al. 1995a). These same factors can prove useful in new assignment strategies.

One area where assignment by sequence dependent shift prediction has proved useful is in studies of intrinsically disordered proteins. It is now recognized that intrinsically disordered regions in proteins play important roles in biological function (Dyson and Wright 2005; Peti et al. 2001; Rezaei-Ghaleh et al. 2012). NMR is one of the few ways of

monitoring these regions and this has heightened interest in making assignments for these regions (Camilloni et al. 2012; Marsh et al. 2006). In disordered regions chemical shifts are dictated largely by amino acid type and the nearest neighbors in the protein sequence. This has led to several attempts to produce chemical shift prediction tools based on assignments of a series of synthetic peptides (Schwarzinger et al. 2001; Kjaergaard and Poulsen 2011), along with tools based on available data for coil regions of proteins in the PDB (Wang and Jardetzky 2002b), or data on denatured or intrinsically disordered proteins that have been deposited in the BMRB (Camilloni et al. 2012; De Simone et al. 2009; Tamiola et al. 2010). Here we present a further refinement of an approach based on a series of synthetic peptides, selected to better represent conditions relevant to the use of chemical shift prediction as an assignment tool for sparsely labeled large proteins.

With the increasing awareness of the importance and prevalence of glycosylation in eukaryotic proteins, in particular, and large proteins in general, we have been trying to address the NMR challenges these present. Considering that eukaryotic cell lines need to be utilized for expressing properly glycosylated proteins, and that these cell lines prefer supplementation with isotopically labeled amino acids, labeling with ^{13}C and ^{15}N can be prohibitively expensive, especially if one insists on uniform isotopic labeling (Dutta et al. 2012; Gossert et al. 2011). Furthermore, eukaryotic hosts do not tolerate the perdeuteration that is essential for maintaining resolution in uniformly labeled large proteins. Using single or small sets of amino acids as a source of sparse isotope labels is attractive in ameliorating some of the expense and resolution requirements imposed with uniform labeling. Implementation of single amino acid labeling does, however, require abandoning commonly used triple resonance assignment strategies.

We therefore have pursued development of an alternative approach to assignment (Nkari and Prestegard 2009; Feng et al. 2007), the most recent version of which requires assignment of the ^{1}H–^{15}N HSQC spectrum of the denatured protein. The published version of the recent strategy relies on correlating crosspeaks in ^{1}H–^{15}N HSQC spectra of native proteins to the crosspeaks in the spectra of the same protein in the denatured state, with assignment of the denatured spectrum ultimately achieved by correlating denatured spectrum crosspeaks with those in spectra of digested, separated, and mass spectrometrically identified peptides (Nkari and Prestegard 2009). It is clear that the latter step might be eliminated if robust, sequence based, chemical shift prediction tools for the denatured protein existed. This is the goal of the present effort.

As our systems are denatured by dissolution in 8 M urea, pH 2.5, the assignment tool introduced by Schwarzinger et al. (2001) is particularly appealing. This tool is based on data from spectra of a series of Ac-Gly-Gly-Xxx-Gly-Gly-NH_2 host–guest peptides dissolved in the same denaturing buffer we use, where Xxx is one of the 20 common amino acids inserted in an all Gly host peptide. It is impractical to synthesize pentapeptides, or even tripeptides representing all possible amino-acid sequences. Hence, the Gly-Gly host sequence was taken to be representative of any flanking region in a disordered protein, and the effects on glycine residues adjacent to (or two removed from) Xxx, on changing from Xxx = Gly to Xxx = another amino acid, are assumed to be the same as they would be, had the observed glycine been any amino acid. This allows parameters in a chemical shift prediction formula of the following form to be determined from perturbations to glycine chemical shifts in a limited set of peptides:

$$\delta_i = \delta_{ref}(Xxx) + \delta_{i-2}(Yyy) + \delta_{i-1}(Yyy) + \delta_{i+1}(Yyy) + \delta_{i+2}(Yyy) \quad (1)$$

Here $\delta_{ref}(Xxx)$ is the reference chemical shift for the amino acid of interest and the remaining $\delta(Yyy)$ terms are correction factors for amino acids found in the protein sequence 1 and 2 positions removed toward the N and C termini, respectively.

There have been some recent attempts to avoid the synthesis requirements of the host–guest strategy and its underlying assumptions, and to extend sequence specific parameterization by mining the growing number of deposited chemical shifts for denatured proteins and proteins with intrinsically disordered regions (Camilloni et al. 2012; Tamiola et al. 2010). A database approach does offer much promise as depositions increase in the future, and there is an inherent advantage in eliminating host peptide bias in guest reference shifts. However, our interest in a particular set of denaturing conditions, and the currently limited number of depositions under these conditions, make a procedure based on the host–guest strategy worthy of further exploration.

Hence, we accept the general approach of Schwarzinger et al., but are concerned that glycine may not be the best choice for a representative amino acid because of the unique aspects of the distribution of allowed torsional conformers (represented in a Ramachandran map) for glycine in contrast to those for many other amino acids. This limitation has also been recognized by others. Recently (Kjaergaard and Poulsen 2011) synthesized a series of peptides using glutamine as a host. However, their interest was primarily in intrinsically disordered proteins, and denaturants were not used in their studies. Here we present a set of data, and a prediction tool, based on Ac–Ala–Ala–Xxx–Ala–Ala–NH_2 peptides studied under

the denaturing conditions used in our sparse label assignment strategy. Our choice of alanine as a host amino acid was made without the benefit of arguments presented in the Kjaergaard paper and may have some limitations. Alanine is known to have a preference for a significant portion of polyproline II (PPII) structure in host–guest systems based on glycine and studied in aqueous solution (estimated at 82 % PPII), but all amino acids except histidine induce 50 % or more propensity for this structure (Shi et al. 2006). The presence of urea is also known to affect propensities for various structures (Bennion and Daggett 2003), although recent studies suggest that the % of PPII, for at least a glycine host–guest system, does not change significantly (Li et al. 2011). It is nevertheless clear that recently generated Ramachandran maps support the choice of alanine over glycine (Ting et al. 2010). We present comparisons of predictions made with our tool to those made with other tools for denatured ubiquitin, a protein representative of our interest in assignment of denatured proteins, to document possible advantages of prediction parameters based on an alternate host amino acid.

Materials and methods

Synthesis of the peptides was carried out with the solid-phase methodology based on α-amino group Fluorenylmethyloxycarbonyl (Fmoc) protection. Fmoc protected amino acid building blocks with the conventional sidechain functional group protections compatible with this chemistry were used as needed. The Fmoc protected amino acids were obtained from Novabiochem, as was the Rink amide resin support. Peptides were assembled on a CEM Liberty microwave-assisted automated peptide synthesizer using standard protocols in the software. Syntheses were run at 0.1 mmol scale. 4-Methyl piperidine, 20 % in dimethylformamide (DMF) was used for Fmoc deprotection, in two cycles, the first for 30 s with microwave heating, and the second for 5 min. Couplings, at fivefold excess of Fmoc amino acid to α-amino sites on the resin, were carried out for 5 min with microwave heating to 75 °C and facilitated by 1-hydroxybenzotriazole (HOBt)/O-Benzotriazole-*N*,*N*,*N*′,*N*′-tetramethyl-uronium-hexafluoro-phosphate (HBTU) (Novabiochem) and with diisopropylethylamine (DIEA) (Sigma-Aldrich) as the base. Nitrogen gas bubbling was used for agitation of the resin during deprotection and coupling steps. The peptide resin as recovered from the synthesizer had the N-terminal Fmoc group already removed. The peptide resins were then manually treated with 20 % acetic anhydride in DMF to acetylate the amino terminus. Release of the peptides from the resin and removal of the chain protecting groups was achieved by treatment with trifluoroacetic acid (TFA)/triisopropylsilane (TIPS)/water 95/2.5/2.5, or 88/5/5/2 TFA/phenol/Water/TIPS. The cleavage solution in which the peptide was dissolved was recovered from the resin by filtration with several additional washes with TFA. The filtrates containing the Ac–Ala–Ala–Xxx–Ala–Ala–NH_2 peptides were partially evaporated under vacuum on a rotary evaporator and the remaining solution either added to anhydrous diethyl ether at 0 °C to form a peptide precipitate or diluted with 20 % acetic acid and extracted with chloroform and ether. In the case of precipitation, the precipitate was collected after centrifugation, dissolved in water, and lyophilized. When extraction was employed, the aqueous phase containing the peptide was lyophilized. Further purification was done with C-18 reverse phase HPLC using a gradient of acetonitrile in water with 0.1 % TFA. Solvents and reagents other than those already indicated were reagent grade from Sigma-Aldrich, except for the acetonitrile which was HPLC grade. 10 mgs or more of each of the peptides were obtained and used for the NMR studies. Except for the Ac–Ala–Ala–Ala–Ala–Ala–NH_2 reference peptide, for which resolution of NMR resonances proved to be a problem, no enrichment in ^{15}N was required. For the Ac–Ala–Ala–Ala–Ala–Ala–NH_2 peptide the central residue was ^{15}N enriched (Sigma-Aldrich Isotope laboratories).

Peptides were dissolved in a 90 % H_2O, 10 % D_2O, 8 M urea, 1 mM DSS, solution adjusted to pH 2.5. Concentrations ranged widely depending on yields and solubilities of peptides, but averaged ~20 mM. HSQC spectra were collected on a Varian 800 MHz spectrometer system using a standard sequence (gNHSQC) from the Varian BioPack. Data were typically acquired with 20–30 indirect points and 2,048 direct points with a recycle time of 2 s over a period of 10–15 min for a 20 mM sample. Data were processed using a 90° shifted sinbell function for weighting and zero filling to 256 points in the indirect dimension. Referencing was initially calculated using the frequency of the HDO signal, but later corrected to values using the DSS methyl proton resonance, following the method of Wishart et al. (1995b). 2D ^1H–^1H ROESY, and ^1H–^1H TOCSY spectra were acquired for the purpose of assigning ^1H–^{15}N HSQC crosspeaks. For typical 20 mM samples these were acquired with 128–256 indirect points and 2,048 direct points with a recycle time of 2 s over a period of 3–8 h, using mixing times of 65 ms for TOCSY and 300 ms for ROESY respectively.

Results

Spectra for peptides were assigned using the distinct NOE connection of an amide resonance to the acetyl methyl for residue 1 and NOE connection of the amide resonance of residue 2 to the alpha and/or beta protons of residue 1. The central amino acid, Xxx, usually displayed a distinct alpha

Table 1 Amide ^{1}H and ^{15}N chemical shifts for AlaAlaXxxAlaAla peptides in 8 M urea, pH 2.5

Residue	Observed chemical shifts									
	$^{1}H^{N}$	A1-$^{1}H^{N}$	A2-$^{1}H^{N}$	A4-$^{1}H^{N}$	A5-$^{1}H^{N}$	^{15}N	A1-^{15}N	A2-^{15}N	A4-^{15}N	A5-^{15}N
Ala	8.29	8.22	8.4	8.25	8.25	123.2	129.5	123.1	123.1	123.5
Arg	8.32	8.19	8.49	8.39	8.29	120.4	129.1	123	125.5	123.9
Asn	8.36	8.22	8.41	8.26	8.31	117.4	129.5	123.1	124.4	122.8
Asp	8.38	8.21	8.38	8.23	8.17	117.4	129.1	122.7	124.4	123.1
Cys	8.33	8.22	8.41	8.45	8.25	117.4	129.6	122.5	126.4	123.3
Gln	8.32	8.22	8.38	8.33	8.26	119.2	129.4	122.8	125.1	123.6
Glu	8.27	8.21	8.38	8.32	8.24	119	129.6	122.8	124.1	123.6
Gly	8.32	8.23	8.41	8.18	8.24	107.8	129.7	123	123.7	123.2
His	8.51	8.23	8.41	8.45	8.38	117.3	129.3	122.7	124.8	124.2
Ile	8.1	8.2	8.38	8.37	8.23	120.4	129.9	123.4	126.2	124.1
Leu	8.17	8.2	8.35	8.31	8.21	121.4	129.5	122.7	124.9	123.7
Lys	8.22	8.16	8.31	8.3	8.25	120.4	129.5	122.7	125.1	123.6
Met	8.33	8.22	8.39	8.34	8.27	119.4	129.3	122.7	125.1	123.7
Phe	8.09	8.2	8.32	8.2	8.14	118.8	129.5	122.5	125.5	123.8
Pro	–	8.18	8.42	8.34	8.28	–	129.3	124.3	124.3	123.6
Ser	8.27	8.23	8.44	8.37	8.2	114.1	130.4	123.1	126	123.4
Thr	8.06	8.21	8.44	8.3	8.22	112.3	129.4	122.9	125.8	123.5
Trp	7.95	8.16	8.31	7.92	7.96	119.2	129.3	122.7	125.4	123.3
Tyr	8.04	8.18	8.32	8.16	8.12	118.9	129.4	122.6	125.7	123.8
Val	8.1	8.23	8.41	8.42	8.28	119.1	129.4	123.2	127.6	124.1
RMSD	0.143	0.022	0.047	0.121	0.086	3.477	0.285	0.404	1.023	0.351

resonance which could be connected to its own amide and side chain resonances in the TOCSY spectrum, but not to one of the alanine methyl groups. The amide proton resonance was also in many cases connected to an HSQC peak with a unique nitrogen chemical shift. The residue 4 alanine amide could be connected to the alpha and/or beta resonances of residue Xxx and the residue 5 alanine could be assigned by default as well as by NOE to the NH_2 protons of the C-terminal carboxamide group. Once several peptides were assigned, the relatively invariant HSQC peak positions for the 1 and 5 positions were used to facilitate the assignment process. In the case of the $(Ala)_5$ peptide, the uniquely enriched central position was used to identify its crosspeak and facilitate connections to the 2 position.

Chemical shifts for ^{1}H and ^{15}N resonances of all peptides studied are given in Table 1. Note the relative invariance of the amide resonances for position A1 (root mean square deviation (RMSD) of ± 0.022 ppm in ^{1}H and ± 0.29 ppm in ^{15}N). The largest variations are at the A4 position (RMSDs are 0.121 and 1.02 ppm for ^{1}H and ^{15}N respectively). This is not surprising as the amide for this alanine is closest to the side chain of the Xxx residue. What is surprising is that there are some significant deviations in the amide proton resonances of the A5 residue. A few are larger than the RMSD of the A4 values (mostly aromatic residues, -0.088 for Phe, -0.108 for Tyr, 0.153 for His, and -0.268 for Trp). While some prediction methods have focused on just the nearest neighbors, this observation suggests that addition of a term from the A5 variations could be important. These anomalies have been noted previously (Schwarzinger et al. 2001) and have been attributed to the tendency of aromatic sidechains to bend under the backbone and allow the aromatic ring to interact with the amide hydrogen of the $i + 2$ (A5) residue (Kemmink and Creighton 1993). The possibility of preserving other long range conformational preferences, even in a denaturing solvent also does exist (Bennion and Daggett 2003; Shi et al. 2006). Although the postulated aromatic interaction is not with the terminal amide (the peptides used terminate with an NH_2 group, not a carboxylic group), it is possible that these interactions are further promoted by the proximity of the carboxy terminus and the lack of steric limitations from an $i + 3$ (A6) sidechain. If this were the origin of the effect, use of a correction term from the A5 data would be inappropriate. In our case, inclusion of the term from the A5 data shows a slight improvement in predictions.

Conversion of position specific differences between chemical shifts for Xxx = Ala and Xxx = other amino acids to correction factors for prediction of chemical shift

Table 2 Comparison of ^{15}N chemical shift correction factors

	Schwarzinger i − 1	Tamiola i − 1	Kjaergaard i − 1	This Work i − 1	Schwarzinger i + 1	Tamiola i + 1	Kjaergaard i + 1	This Work i + 1
Ala	0	0	0	0	0	0	0	0
Arg	2.19	2.55	2.41	2.4	0.19	−0.06	0.03	−0.1
Asn	1.44	1.34	1	1.3	0.07	−0.22	−0.4	0
Asp	1.43	1.32	0.84	1.3	0.13	−0.37	−0.17	−0.4
Cys	3.64	3.74	3.35	3.3	0.07	0.79	0.01	−0.6
Gln	2.19	2.34	2.25	2.0	0.19	−0.08	−0.14	−0.3
Glu	2.08	1.87	1.82	1	0.13	−0.10	−0.17	−0.3
Gly	0.57	0.64	0.16	0.6	0.33	0.00	−0.18	−0.1
His	2.25	2.40	2.22	1.7	−0.22	−0.22	−0.46	−0.4
Ile	5.44	4.84	5.42	3.1	0.19	−0.16	0.11	0.3
Leu	1.62	1.47	1.61	1.8	0.19	−0.39	−0.3	−0.4
Lys	2.14	1.95	2.4	2	0.13	0.07	−0.07	−0.4
Met	2.14	2.05	2.2	2	0.13	−0.14	−0.18	−0.4
Phe	3.35	2.88	2.44	2.4	−0.16	−0.48	−0.59	−0.6
Pro	1.44	1.21	1.22	1.2	0.01	1.08	1.29	1.2
Ser	3.12	2.94	2.46	2.9	0.3	−0.05	0.03	0
Thr	3.35	3.32	3.13	2.7	0.3	0.08	0.09	−0.2
Trp	3.76	1.43	2.3	2.0	0.07	−0.27	−0.37	−0.3
Tyr	3.58	3.37	2.64	2.6	−0.1	−0.38	−0.54	−0.5
Val	4.91	5.15	5.08	4.5	0.19	−0.04	0.18	0.1
	i − 2	i − 2	i − 2	i − 2	i + 2	i + 2	i + 2	i + 2
Ala	0		0	0	0		0	0
Arg	0.09		0.39	0.4	0.06		−0.04	−0.4
Asn	−0.02		−0.55	−0.7	−0.06		−0.03	0
Asp	−0.14		−0.79	−0.4	0		0.04	−0.4
Cys	0.15		0.14	−0.2	0.06		−0.05	0.1
Gln	0.09		0.21	0.1	0.06		0	0.1
Glu	0.03		0.07	0.1	0.06		0.02	0.1
Gly	0.15		−0.24	−0.3	0.12		−0.02	0.2
His	0.32		0.16	0.7	0		−0.07	−0.2
Ile	0.15		0.86	0.6	−0.06		−0.07	0.4
Leu	0.09		0.09	0.2	0.06		0.04	0
Lys	0.09		0.38	0.1	0.06		−0.04	0
Met	0.09		0.23	0.2	0.06		0.06	−0.2
Phe	−0.31		0.12	0.3	−0.06		0.21	0
Pro	−0.02		0.13	0.1	−0.06		−0.09	−0.2
Ser	−0.02		−0.41	−0.1	0.06		0.01	0.9
Thr	0.03		0.22	0	0.06		−0.02	−0.1
Trp	−0.49		−0.55	−0.2	0.12		−0.07	−0.2
Tyr	−0.37		0.08	0.3	−0.12		0.2	−0.1
Val	0.09		0.84	0.6	−0.12		−0.08	−0.1

requires association of i + 2, i + 1, i − 1, and i − 2 sequence positions with changes in shifts of A1, A2, A4, and A5 respectively. The correction factors for the twenty peptides we studied are given in Tables 2 and 3 along with comparisons to correction factors from other studies. In the latter cases published factors had to be adjusted for the fact that the corrections are in each case relative to a different substitution; in our case Xxx for Ala, in other cases Xxx for

Table 3 Comparison of ^{1}H chemical shift correction factors

	Schwarzinger i − 1	Tamiola i − 1	Kjaergaard i − 1	This Work i − 1	Schwarzinger i + 1	Tamiola i + 1	Kjaergaard i + 1	This Work i + 1
Ala	0	0	0	0	0	0	0	0
Arg	0.08	0.16	0.12	0.14	0.03	0.02	0.03	0.09
Asn	0.06	0.04	0.1	0.01	0.02	0.03	0.07	0.01
Asp	0.07	−0.01	−0.05	−0.02	0.02	0.05	0.05	−0.02
Cys	0.13	0.21	0.15	0.2	0.03	0.43	0.07	0.01
Gln	0.08	0.14	0.1	0.08	0.03	0.04	0.04	−0.01
Glu	0.08	0.09	0.08	0.07	0.02	0.04	0.03	−0.02
Gly	−0.07	−0.09	−0.11	−0.07	0.05	0.06	0.07	0.01
His	0.13	0.05	−0.02	0.2	0.01	−0.08	0.03	0.01
Ile	0.1	0.16	0.13	0.12	−0.01	−0.02	0.01	−0.02
Leu	0.07	0.02	0.01	0.06	0.02	−0.03	0.04	−0.05
Lys	0.07	0.09	0.02	0.05	0.02	0.00	0.02	−0.09
Met	0.08	0.12	0.06	0.09	0.03	−0.01	−0.12	−0.01
Phe	0.03	−0.01	−0.17	−0.05	−0.07	−0.08	0.01	−0.08
Pro	0.12	0.20	0.18	0.09	−0.13	0.03	0.03	0.02
Ser	0.09	0.14	0.08	0.12	0.02	0.05	−0.01	0.04
Thr	0.07	0.12	0.08	0.05	0.05	0.10	0.07	0.04
Trp	−0.03	−0.64	−0.43	−0.33	−0.08	0.09	0.02	−0.09
Tyr	0.02	−0.14	−0.21	−0.09	−0.06	−0.04	0.02	−0.08
Val	0.1	0.17	0.14	0.17	0	0.00	0.01	0.01
	i − 2	i − 2	i − 2	i − 2	i + 2	i + 2	i + 2	i + 2
Ala	0		0	0	0		0	0
Arg	0.04		0.08	0.04	0.01		0	−0.03
Asn	0.03		−0.03	0.06	0		0	0
Asp	0		−0.03	−0.08	−0.01		0	−0.01
Cys	0.03		0.02	0	0.01		−0.01	0
Gln	0.04		0.07	0.01	0		0	0
Glu	0.03		0.04	−0.01	0		0	−0.01
Gly	0.1		0.07	−0.01	0.01		0.01	0.01
His	0.1		0.09	0.13	0		−0.01	0.01
Ile	0.01		0.08	−0.02	0		−0.03	−0.02
Leu	0.02		0	−0.04	0.01		0	−0.02
Lys	0.04		0.07	0	0.01		0	−0.06
Met	0.04		0.05	0.02	0.01		−0.01	0
Phe	−0.27		−0.05	−0.11	−0.02		−0.01	−0.02
Pro	−0.02		0.04	0.03	−0.03		−0.02	−0.11
Ser	0.02		−0.06	−0.05	0.01		0.01	0.01
Thr	0.04		0.03	−0.03	0.11		0	−0.01
Trp	−0.52		−0.29	−0.29	−0.07		−0.04	−0.06
Tyr	−0.32		−0.08	−0.13	−0.03		−0.01	−0.04
Val	0.02		0.09	0.03	0		−0.03	0.01

Gln (Kjaergaard and Poulsen 2011) or Xxx for Gly (Schwarzinger et al. 2001; Tamiola et al. 2010). This was done by subtracting the Ala for Gln or Ala for Gly correction factor from each published correction factor.

There are some consistencies in the various sets, such as the i − 1 factors (A4 differences) being the most significant, and i + 2 factors (A1 differences) being very small. Also, the same i − 2 factors (aromatic amino acids) which

are large in our set are frequently large in other sets. However, there are significant variations in the magnitudes of corrections; this is perhaps not surprising as the reference amino acids vary (Ala vs. Gly vs. Gln) and some correction factors are based on urea denaturation and others on intrinsically disordered regions. There is then reason to think that for our applications an alanine reference amino acid may be better than a glycine reference amino acid and that correction factors derived from urea denatured data would work best.

A prediction tool was written in both the C programing language and as a Perl script using the correction factors derived from our urea denatured data on alanine pentapeptides and reference chemical shifts taken from each central amino acid in those peptides. In the Perl script, the user is able to choose a correction table for any one of the sets discussed using the published parameters and reference amino acid shifts. One can argue that average random coil shifts would provide a more appropriate set of reference shifts for applications to intrinsically disordered proteins and we have provided this as an option in our own correction tables. We have also allowed an option of using just $i \pm 1$ corrections or including $i \pm 2$ as well. Input to the program is simply a file with the amino acid sequence and a correction factor table, both selected by entry of file names on the command line calling the program. The output is a predicted set of chemical shifts for each amino acid in the sequence. Referencing is a principle source of variation in application to any protein. We have chosen to reference proton shifts to DSS in 8 M urea, pH 2.5. Nitrogen shifts are calculated indirectly with the ^{1}H resonance of DSS referenced to zero ppm. The tool is available from our website at http://tesla.ccrc.uga.edu/software/.

Discussion

A convenient test of our chemical shift prediction tool is provided by data from the literature: a set of ubiquitin crosspeaks for a ^{1}H–^{15}N HSQC spectrum taken in 8 M urea and assigned using conventional double and triple resonance strategies (Peti et al. 2001). Figure 1 is a superposition of experimental and predicted crosspeaks in which referencing has been adjusted to minimize deviations between the two sets. The agreement is reasonable in the sense that the distribution of crosspeaks is similar. Some clustering in horizontal bands likely reflects the more substantial contribution of the central amino acid type to ^{15}N shifts as compared to amino acid type contributions to ^{1}H shifts. This may also contribute to the greater accuracy of shift prediction for ^{15}N shifts compared to ^{1}H shifts, for which respective RMSDs of 0.99 and 0.09 ppm translate to percentages of chemical shift ranges of 5 and 12 %. In Fig. 1, where lines connect predicted and experimental shifts, it is clear that only a few of the outliers could be assigned with any confidence using just chemical shift prediction. Ellipses are drawn at ± 1 standard deviation for the few cases where this can be done.

For our anticipated application to sparsely labeled proteins the percentage of unambiguous assignments should be higher. Average percentages for RMSDs relative to range increase when grouping predictions by central amino acid type (27 and 25 %). But if only a single type of amino acid were labeled, and distribution among types were uniform, one would expect on average only 4 crosspeaks for an 80 amino acid protein. Spreading the four peaks randomly over the range of chemical shifts, one would expect about an 80 % chance of all four being separated by at least one RMSD and of the ambiguous cases, more than a 90 % chance that only one pair would have an ambiguous assignment. For larger proteins assignments will be less complete, but there are other sources of information that can remove ambiguities, for example, a simple NOESY-HSQC spectrum of the denatured protein of interest. NOE peaks in a denatured protein typically connect a given ^{15}N–^{1}H crosspeak to the alpha and beta proton shifts of the $i - 1$ residue and allow assignment of the $i - 1$ residue to a class of amino acids (long chain, alanine, single methylene etc.) (Peti et al. 2001). This would further reduce ambiguities.

Comparison of the quality of prediction using our set of correction factors to the quality using previous sets is useful. One comparison is presented in Fig. 2 in which we show correlations between experimental and predicted

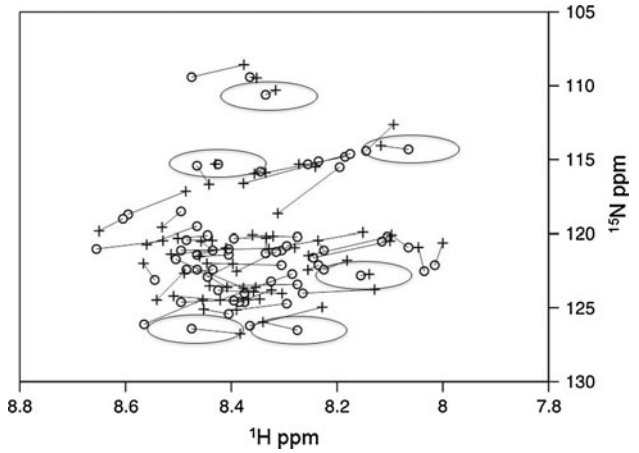

Fig. 1 Simulated ^{1}H–^{15}N HSQC spectrum comparing experimental (+) and predicted (O) crosspeaks for ubiquitin in 8 M in urea. Experimental data are from Peti et al. (2001). Only residues 6–74 are included to minimize end effects. Glycine 47 was eliminated as an extreme outlier in all predictions. *Ellipses* are drawn for two cases where definitive assignments can be made. *Ellipse axes* are two standard deviations in each dimension

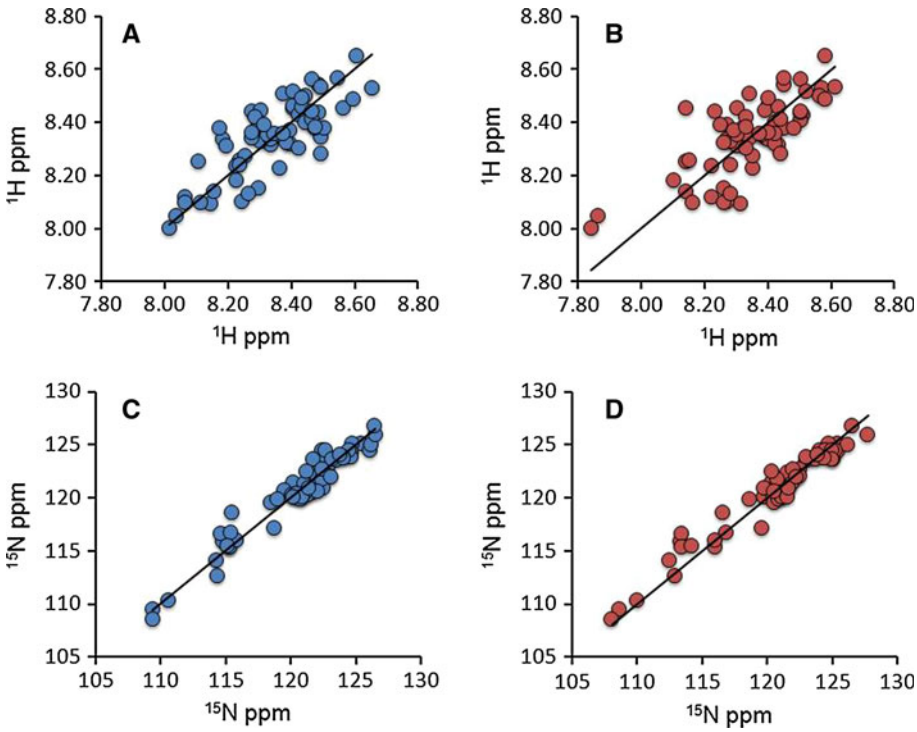

Fig. 2 Comparison of ^{1}H and ^{15}N experimental chemical shifts to predicted chemical shifts using parameter sets from this work (**a**, **c** respectively) and using parameters from Schwarzinger et al. (**b**, **d** respectively) (Schwarzinger et al. 2001)

^{15}N and ^{1}H amide shifts for our set and the Schwarzinger set which is also based on low pH urea denatured pentapeptides, but using glycine rather than alanine for flanking residues. The predictions in both cases are fairly good for ^{15}N shifts (R^2 values of 0.94 and 0.93 for our set and the Schwarzinger set respectively). The largest variations are for ^{1}H shifts. Here our predictions show a small, but significant advantage (R^2 values of 0.59 and 0.45 ppm respectively). That correlations are not as good for amide ^{1}H chemical shifts using either set is not surprising. ^{1}H chemical shifts have been notoriously difficult to predict, likely because of substantial solvent and hydrogen bonding effects and the importance of long range effects in ^{1}H chemical shifts. That our predictions are better suggests that the choice of Ala as a representative amino acid is better than Gly, especially as both the Schwarzinger set and our set use identical denaturing conditions.

Comparisons to other sets are not plotted, but the trend is the same. ^{15}N R^2 values for the Tamiola and Kjaergaard sets are 0.94 and 0.96 respectively, both very good. ^{1}H R^2 values for the Tamiola and Kjaergaard sets are 0.33 and 0.47 respectively. The fact that the Kjaergaard set does fairly well may reflect the importance of choosing a good representative amino acid (their choice was Gln). The fact that neither set does quite as well as ours on amide proton predictions for urea denatured ubiquitin may simply reflect that these sets were not based exclusively on urea denatured peptides or proteins and were primarily intended for application to intrinsically disordered regions of proteins.

In summary, we have produced a piece of prediction software especially tailored to predicting HSQC spectra of denatured proteins. It shows a small, but significant, improvement over existing software for at least an application to a urea denatured test set. Direct applications are anticipated for emerging assignment strategies that correlate native HSQC spectra of sparsely labeled proteins with spectra of their denatured counterparts and rely on the ease of assignment of the denatured spectra. Additional applications may be found in the study of intrinsically disordered proteins where it joins a list of similarly based prediction software.

Acknowledgments This work was supported by grants from the NIH to the Resource for Integrated Glycotechnology at the University of Georgia (P41 RR005351-23 and P41 GM103390-23), to JHP for research support (R01 GM061268-09), from the NIH to J. L. Urbauer from the Shared Instrumentation Program for upgrading an NMR console (S10RR027097-01A1) and to DL from the Shared Instrumentation Program for the purchase of the peptide synthesizer (1S10 RR027155-01).

References

Bennion BJ, Daggett V (2003) The molecular basis for the chemical denaturation of proteins by urea. Proc Natl Acad Sci USA 100:5142–5147

Camilloni C, De Simone A, Vranken WF, Vendruscolo M (2012) Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. Biochemistry 51:2224–2231

De Simone A, Cavalli A, Hsu STD, Vranken W, Vendruscolo M (2009) Accurate random coil chemical shifts from an analysis of

loop regions in native states of proteins. J Am Chem Soc 131:16332

Dutta A, Saxena K, Schwalbe H, Klein-Seetharaman J (2012) Isotope labeling in mammalian cells. In: Shekhtman A, Burz DS (eds) Protein NMR techniques. Methods in molecular biology, vol 831, 3 edn. Humana Press, Totowa, NJ, pp 55–69

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

Feng L, Lee HS, Prestegard JH (2007) NMR resonance assignments for sparsely ^{15}N labeled proteins. J Biomol NMR 38:213–219

Gossert AD, Hinniger A, Gutmann S, Jahnke W, Strauss A, Fernandez C (2011) A simple protocol for amino acid type selective isotope labeling in insect cells with improved yields and high reproducibility. J Biomol NMR 51:449–456

Kemmink J, Creighton TE (1993) Local conformations of peptides representing the entire sequence of bovine pancreatic trypsin-inhibitor and their roles in folding. J Mol Biol 234:861–878

Kjaergaard M, Poulsen FM (2011) Sequence correction of random coil chemical shifts: correlation between neighbor correction factors and changes in the Ramachandran distribution. J Biomol NMR 50:157–165

Lange OF, Rossi P, Sgourakis NG, Song YF, Lee HW, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, Baker D (2012) Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proc Natl Acad Sci USA 109:10873–10878

Li WF, Qin M, Tie ZX, Wang W (2011) Effects of solvents on the intrinsic propensity of peptide backbone conformations. Phys Rev E 84:041933

Marsh JA, Singh VK, Jia Z, Forman-Kay JD (2006) Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. Protein Sci 15:2795–2804

Nkari WK, Prestegard JH (2009) NMR resonance assignments of sparsely labeled proteins: amide proton exchange correlations in native and denatured states. J Am Chem Soc 131:5344–5349

Peti W, Smith LJ, Redfield C, Schwalbe H (2001) Chemical shifts in denatured proteins: resonance assignments for denatured ubiquitin and comparisons with other denatured proteins. J Biomol NMR 19:153–165

Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu GH, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D (2010) NMR structure determination for larger proteins using backbone-only data. Science 327:1014–1018

Rezaei-Ghaleh N, Blackledge M, Zweckstetter M (2012) Intrinsically disordered proteins: from sequence and conformational properties toward drug discovery. ChemBioChem 13:930–950

Schwarzinger S, Kroon GJ, Foss TR, Chung J, Wright PE, Dyson HJ (2001) Sequence-dependent correction of random coil NMR chemical shifts. J Am Chem Soc 123:2970–2978

Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223

Shi ZS, Chen K, Liu ZG, Kallenbach NR (2006) Conformation of the backbone in unfolded proteins. Chem Rev 106:1877–1897

Takeuchi K, Ng E, Malia TJ, Wagner G (2007) 1-^{13}C amino acid selective labeling in a ^{2}H^{15}N background for NMR studies of large proteins. J Biomol NMR 38:89–98

Tamiola K, Acar B, Mulder FA (2010) Sequence-specific random coil chemical shifts of intrinsically disordered proteins. J Am Chem Soc 132:18000–18003

Ting D, Wang GL, Shapovalov M, Mitra R, Jordan MI, Dunbrack RL (2010) Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. PLoS Comput Biol 6:e1000763

Wang Y, Jardetzky O (2002a) Investigation of the neighboring residue effects on protein chemical shifts. J Am Chem Soc 124:14075–14084

Wang Y, Jardetzky O (2002b) Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci 11:852–861

Wishart DS, Bigam CG, Holm A, Hodges RS, Sykes BD (1995a) ^{1}H, ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. J Biomol NMR 5:67–81

Wishart DS, Bigam CG, Yao J, Abildgaard F, Dyson HJ, Oldfield E, Markley JL, Sykes BD (1995b) H-1, C-13 AND N-15 chemical-shift referencing in biomolecular NMR. J Biomol NMR 6:135–140