Send Orders for Reprints to reprints@benthamscience.ae

Protein & Peptide Letters, **2016**, *23*, **967-975**

RESEARCH ARTICLE

The Nearest-Neighbor Effect on Random-Coil NMR Chemical Shifts Demonstrated Using a Low-Complexity Amino-Acid Sequence

Tsai-Chen Chen[1,¶], Chih-Lun Hsiao[1,¶], Shing-Jong Huang[3] and Jie-rong Huang[1,2,*]

[1]*Institute of Biochemistry and Molecular Biology,* [2]*Institute of Biomedical Informatics, National Yang-Ming University, No. 155 Section 2 Li-nong Street, Taipei, Taiwan;* [3]*Instrumentation Center, National Taiwan University. No. 1, Sec. 4, Roosevelt Road, Taipei, Taiwan*

ARTICLE HISTORY

Received: July 19, 2016
Revised: September 12, 2016
Accepted: September 12, 2016

DOI: 10.2174/0929866523666160920100045

Abstract: In NMR experiments, the chemical shift is typically the first parameter measured and is a source of structural information for biomolecules. Indeed, secondary chemical shifts, the difference between the measured chemical shifts and those expected for a randomly oriented sequence of peptides (the "random coil"), are correlated with the secondary structure of proteins; secondary shift analysis is thereby a standard approach in structural biology. For intrinsically disordered or denatured proteins furthermore, secondary chemical shifts reveal the propensity of particular segments to form different secondary structures. However, because the atoms in unfolded proteins all have very similar chemical environments, the chemical shifts measured for a certain atom type vary less than in globular proteins. Since chemical shifts can be measured precisely, the secondary chemical shifts calculated for an unfolded system depend mainly on the particular random coil chemical shift database chosen as a point of reference. Certain databases correct the random coil shift for a given residue based on its neighbors in the amino acid sequence. However, these corrections are typically derived from the analysis of model peptides; there have been relatively few direct and systematic studies of the effect of neighboring residues for specific amino acid sequences in disordered proteins. For the study reported here, we used the intrinsically disordered C-terminal domain of TDP-43, which has a highly repetitive amino-acid sequence, as a model system. We assigned the chemical shifts of this protein at low pH in urea. Our results demonstrate that the identity of the nearest neighbors is decisive in determining the value of the chemical shift for atoms in a random coil arrangement. Based on these observations, we also outline a possible approach to construct a random-coil library of chemical shifts that comprises all possible arrangement of tripeptides from a manageable number of polypeptides.

Keywords: Intrinsically disordered proteins, secondary chemical shift, random coil chemical shift, TDP-43.

1. INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy is one of the most powerful tools to study protein structure and dynamics [1, 2]. The chemical shift can be measured by NMR for almost all the atoms in a molecule and varies with their local environment. The chemical shift can therefore be used to probe different locations in a molecule, giving site-specific information for further structural and dynamic studies. Because the accurate prediction from first principles of chemical shifts for a known structure requires considerable computational power—complicated electrostatic environments have to be modeled around each atom [3], empirical databases are currently the most popular and precise resource for chemical shift predictions [4, 5]. The reverse problem, namely using chemical shifts by themselves to determine a protein structure, is challenging but nonetheless feasible with the assistance of knowledge-based approaches [6-8]. However, chemical shifts also provide structural insights directly, even without a sophisticated algorithm: secondary chemical shifts (the difference between the observed chemical shifts and the chemical shifts expected for the same amino-acid sequence in a random coil conformation) are the most straightforward indication of secondary structure elements, the correlation between the two being well known [9-11]. Secondary chemical shift analysis is therefore a standard approach in NMR protein structure determination, whereby

*Address correspondence to this author at the Institute of Biochemistry and Molecular Biology, National Yang-Ming University, No. 155 Section 2 Li-nong Street, Taipei, Taiwan; Tel: +886-2826-7258;
E-mails: jierongh@ym.edu.tw
¶*These authors contributed equally.*

0929-8665/16 $58.00+.00

secondary structure elements are identified immediately after chemical shift assignment.

The chemical shift also provides structural information for unfolded proteins. In these molecules however, the variability of the chemical shift is much lower because of the similarity of the atoms' chemical environment. Nevertheless, modern NMR spectrometers and multi-dimensional NMR techniques can be used to assign chemical shifts to each (NMR-sensitive) atom in the chain [12-14]. The secondary chemical shift then reveals the conformational propensity of a given segment or the presence of residual or transiently populated structure [15]. The residual structures adopted by unfolded proteins are key to understanding how proteins fold [15, 16]. Furthermore, identifying the structural propensity of intrinsically disordered proteins may provide information on their biological functions and their roles in various diseases [17, 18].

Chemical shifts can usually be measured very precisely thus the accuracy of the corresponding secondary shifts depends mainly on that of the set of random coil values used to derive them. Several such databases are available, most of which have been assembled either from systematic studies of model peptides (Ac-GGXGG-NH_2 host-guest systems, where X stands for one of the 20 amino acids) [19-21], or from the analysis of chemical-shift databases such as the Biological Magnetic Resonance Data Bank (BMRB)[22-24]. For folded proteins, the choice of the random coil database used to calculate the secondary shifts is not critical, but results may differ significantly for unfolded systems because the variance of the values is much smaller. Tools such as SSP [25] and δ2D [26] that base their secondary structure predictions on the chemical shifts of several atoms per residue aim to eliminate this database dependence. In addition, correction factors for neighboring residues have been included in several random-coil databases [19, 27, 28]. The fact that the amino-acid type of the nearest neighbors affects the chemical shifts of a given residue is well known, but since measuring this effect for all possible combination of dipeptides (400) or tripeptides (8000) is not realistic, it has never been quantified in a natural polypeptide. Here, we use the C-terminal domain of TDP-43 (TDP-$43^{266–414}$) as a model system to demonstrate the nearest-neighbor effect on the chemical shifts of amino acids in a random coil arrangement. This domain has low sequence complexity with several tripeptide repeats and is thus a good choice for a detailed analysis of the nearest-neighbor effect. We assigned the chemical shifts of this domain in 8 M urea at pH 2.5, a condition commonly used to determine random coil chemical shifts. Our analysis demonstrates that other than the identity of the amino acid itself, nearest neighbor type is indeed the main factor that governs the chemical shift of atoms in a random coil. Based on our findings, we also propose a way to construct a random-coil chemical-shift database that accounts for this effect.

2. MATERIALS AND METHOD

2.1. DNA Construct and Primer Design

The construct used in this study was derived from plasmid encoded human TDP-43 (Genomic Research Center, National Yang-Ming University, Taiwan). Genes coding for wild-type TDP-43 (residues 266–414) were inserted into a pET21a+-based vector with the restriction enzymes *Bam*HI and *Xho*I. The T7 tag on the vector was replaced with six histidines.

2.2. Protein Purification and Expression

Single colonies of transformed *E. Coli* cells (BL21) were picked up from an ampicillin agar plate and were used to inoculate 5.0 mL of lysogeny broth containing 0.1 $mg \cdot mL^{-1}$ ampicillin and were grown overnight in a shaker at 37 °C. This overnight culture was used to inoculate 500 mL of lysogeny broth or minimal M9 medium (containing $^{15}NH_4Cl$ and/or ^{13}C-glucose for isotope labeling), both with 0.1 $mg \cdot mL^{-1}$ ampicillin. The growth cultures were left at 37 °C in an incubated shaker until the OD_{600} reached ~0.6–0.7. The cells were induced with a final concentration of 1 mM isopropyl β-D-1-thiogalactopyranoside and were left shaking overnight at 25 °C. The cells were harvested by centrifugation (Beckman JA-10, 30 min at 8000 rpm and 4 °C) and resuspended into 15 mL of B-PER reagent (Thermo Scientific) with 0.1 $mg \cdot ml^{-1}$ lysozyme and 5 $unit \cdot ml^{-1}$ of deoxyribonuclease I, or lysed by sonication on ice for 4 min (15 s pulse on and 45 s pulse off) at 80 % power with a 9 mm probe (~2000 J generated in total by the sonicator). The lysate was centrifuged at 18000 rpm (Beckman JA-25.5) at 4 °C for 30 min. The precipitant was dissolved in 8 M urea in 20 mM Tris buffer at pH 8.0. This solution was filtered (0.45 μm) and loaded onto a nickel-charged immobilized metal-ion affinity chromatography column (1 ml or 5 ml pre-packed column, Bio-Rad). The column was washed with five column volumes of 20 mM Tris buffer with 8 M urea at pH 8.0 and then eluted with 15 column volumes of the same buffer containing 0–500 mM imidazole solution using fast protein liquid chromatography. The fractions containing target proteins were collected and loaded onto a cation-exchange column (1 ml or 5 ml pre-packed Unosphere S-column, Bio-Rad). After washing the column with five column volumes of 10 mM phosphate buffer containing 8 M urea at pH 8.0, the protein was eluted with 15 column volumes of the same buffer with a 0–1 M NaCl concentration gradient. The protein fractions were exchanged with a 10 mM glycine buffer at pH 2.5 with 8 M urea. A typical yield of ~6–10 mg of purified protein was gathered from each half-liter of minimal medium.

2.3. NMR Experiments and Data Analysis

All NMR experiments were performed on a Bruker AVANCE 800 MHz spectrometer with the sample at 283 K. The parameters used for the assignment experiments, namely HNCO, HN(CA)CO, HNCA, HN(CO)CA, HNCACB, CBCA(CO)NH, are listed in S1 Table. The data were processed and analyzed using NMRPipe [29] and Sparky [30]. Non-uniform sampling schemes were generated by TopSpin (Bruker BioSpin) and reconstructed using an iterative soft-thresholding algorithm [31].

For the chemical shift assignment experiments, the NMR signal-to-noise (S/N) ratio was high enough to allow the data to be acquired with non-uniformly sampling (see S1 Table for details of the settings used), which afforded optimal resolution in the indirect dimensions (^{15}N and ^{13}C) in a reason-

able amount of experimental time. For the constant-time type experiments, the resolution of the indirect dimensions (the ^{13}C dimension in particular) was optimized; for the other pulse sequences, we increased the total evolution time (the number of data points acquired), because the nuclear spin relaxation rates of ^{13}C and ^{15}N atoms are much longer in unstructured than in structured proteins. In addition, the chemical shift dispersion of carbonyl atoms is still large in unstructured proteins [32]; this property facilitates assignment.

2.4. Sequence Complexity and Protein Disorder Prediction

The complexity of the protein sequence was analyzed using the SEG algorithm [33], while the level of structural disorder was predicted using the PONDR program running the VSL2 or VL3 algorithm [34, 35], and the IUPRED web server [36].

2.5. Tripeptide Analysis

The BLAST non-redundant protein sequence database was downloaded from the National Center for Biotechnology Information (NCBI) website (ftp://ftp.ncbi.nlm.nih.gov/blast/db/). We used in-house C-shell and GNU Octave scripts to count the occurrence of all tripeptide combinations and to construct polypeptides comprising all 8000 triplets.

3. RESULTS

3.1. The Low Sequence Complexity of TDP-43$^{266–414}$ Makes it Suitable for Characterizing the Nearest-Neighbor Effect

Proteins with a low complexity sequence are good candidates for characterizing the nearest-neighbor effect on NMR chemical shifts because of the high occurrence therein of repeated sequence motifs. Fibrous proteins exemplify this property. Fibroin, for example, has high glycine, alanine, and serine contents, while glycine and alanine are also highly prevalent in collagen and elastin [37]. A recent study has also shown that many RNA binding proteins (such as TDP-43) have low-complexity regions in addition to their RNA binding motifs [38]. These regions are involved in regulating different protein or nucleotide interactions [39].

TDP-43 is a splicing factor that promotes pre-mRNA exon skipping and is involved in micro-RNA processing and mRNA transportation and translation regulation. This protein has also been associated with the survival of motor neurons and the formation of stress granules [40]. In addition, TDP-43 has been implicated in amyotrophic lateral sclerosis and frontotemporal labor degeneration, with its C-terminal part being the main component of the inclusion bodies found in patient biopsies [41]. The structure and function of its two RNA recognition motifs have been extensively studied [42, 43] and the structure of its N-terminal domain has very recently been solved [44, 45]. In contrast, the C-terminal domain is known to be intrinsically disordered [46], as illustrated in Fig. **1** with the predictions from three different algorithms (Fig. **1**). Sequence analysis demonstrates that the low complexity regions are in this intrinsically disordered domain (Fig. **1**). On this basis, we cloned the 149 C-terminal residues of this protein (residue numbers 266 to 414) to use as a model system. Table **1** lists the 28 amino-acid doublets and 14 triplets that appear more than once in this protein domain.

Table 1. Number of times each repeated motif in the amino-acid sequence of TDP-43$^{266–414}$ appears therein.

GG	AA	SS	NN	MM	SG	GS
8	4	4	3	2	9	8
GN	NQ	FG	GM	GF	SN	AS
7	6	6	4	4	4	3
NS	QG	GA	WG	AG	AF	NP
3	3	3	3	2	2	2
NM	QA	QN	QS	GW	MA	MG
2	2	2	2	2	2	2
NQG	GNN	GGF	GFG	NNQ	QNQ	GGG
3	3	3	3	2	2	2
GSN	GSG	GWG	FGN	FGS	SGN	SGS
2	2	2	2	2	2	2

3.2. Chemical Shift Assignment of TDP-43$^{266–414}$ in the Presence of 8 M Urea at pH 2.5

We dissolved TDP-43$^{266–414}$ in 8 M urea at pH 2.5, a condition commonly used to determine random-coil chemical shifts in model peptides [27, 28]. Although TDP-43$^{266–414}$ is prone to aggregation under physiological conditions [47, 48], its solubility is high at low pH in urea and concentrations up to 1 mM were readily achieved.

The chemical shifts of the carbonyl carbons are readily assigned from their sequential connectivity (Fig. **2A** and S1 Fig). Although the sequential walk through the Cα and Cβ chemical shifts is more ambiguous, their specificity in terms of amino acid types [49] can be used to resolve ambiguities in the carbonyl carbon sequence. We initially used an automatic assignment program, MARS [50], and then confirmed the results manually. The standard set of experiments (HNCO, HN(CA)CO, HNCA, HN(CO)CA, CBCA(CO)NH, and HNCACB) were sufficient in this case to assign all the backbone atoms. We also collected proton chemical shifts from NOESY-HSQC and TOCSY-HSQC experiments to confirm the sequential connectivity. The ^{15}N-^{1}H HSQC spectrum in Fig. **2B** illustrates the completed assignment. The chemical shifts have been deposited in the Biomolecular Magnetic Resonance Bank with the accession number 26816.

3.3. Neighboring residues influence random-coil chemical shifts

The highly repetitive sequence of TDP-43$^{266–414}$ reveals the effect of neighboring residues on the chemical shifts of a given amino acid. Focusing on the chemical shifts of the most populated amino acids in this protein, namely glycine

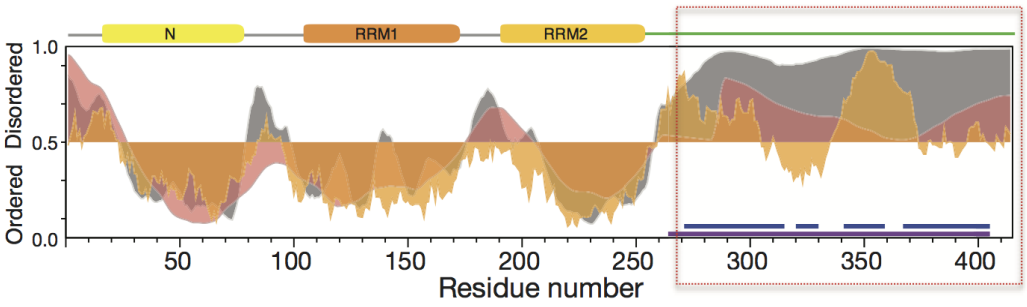

Figure 1. Structure disorder prediction and sequence complexity analysis for TDP-43. Structure disorder of the protein as predicted by PONDR VSL2 (grey), VL3 (red), and IUPRED (orange). Scores higher than 0.5 indicate that the conformation of the corresponding residue is mainly disordered. The sequence complexity was calculated using the SEG algorithm, and the two regions found to be of low complexity sequence, LC1 and LC2, are respectively indicated with a blue and a purple bars. A red dashed box has been drawn around the protein domain studied in this article. (Color version online)

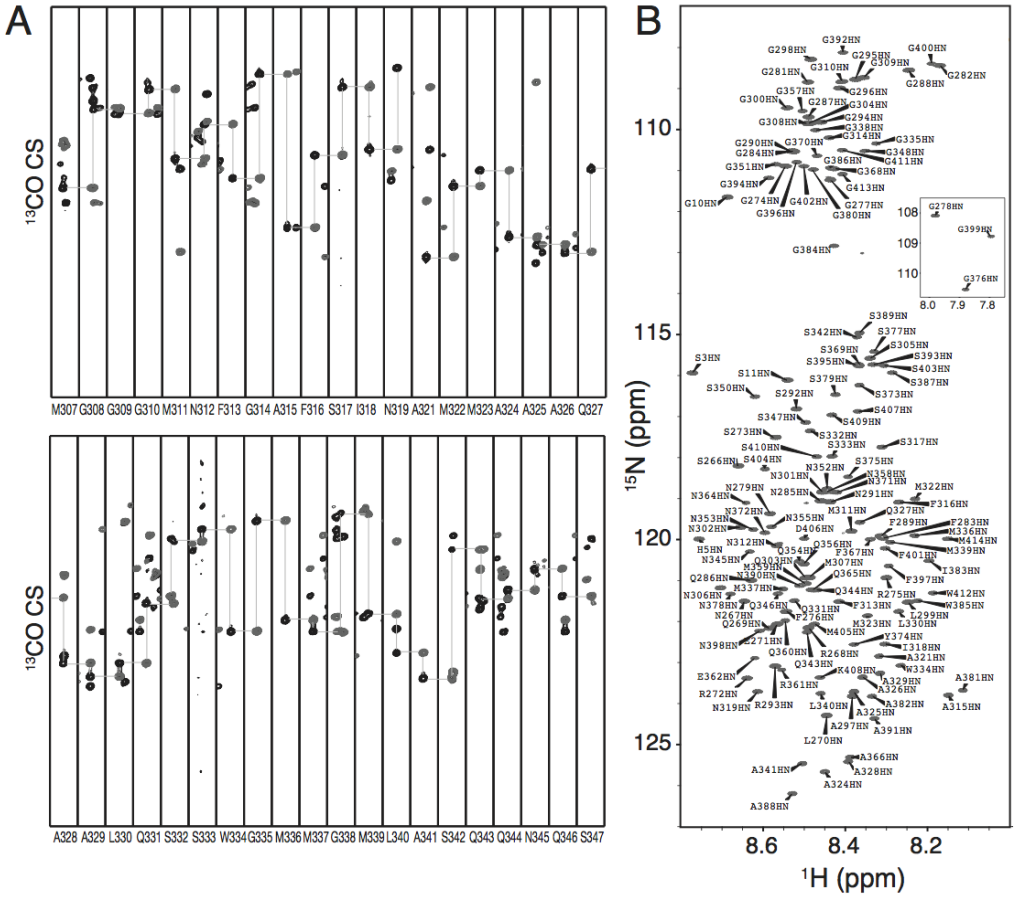

Figure 2. Chemical shift assignment of TDP-43$^{266-414}$ in 8 M urea at pH 2.5. (A) Sequential connectivity of the residues as illustrated by green lines drawn through strip plots of HNCO (gray) and HN(CA)CO (black) NMR spectra. (B) An NMR ^{15}N-^{1}H HSQC spectrum of the protein showing the assignments of all resonances.

(G), serine (S), and asparagine (N) (with 38, 24, and 20 appearances respectively), when the neighboring residues are ignored, the Cα and C' chemical shifts are spread over 0.5–1 ppm and ~1 ppm respectively (Fig. **3**, black dots). This variance is reduced when the preceding or following residue is taken into account (Fig. **3**, red, orange, and green dots). Because the following residue is covalently bound to the C' atom, C' chemical shifts from residues with identical successors should be more similar than those of residues with identical predecessors. For example, the C' chemical shifts of glycine residues preceded by another glycine have a larger variance than those of glycines followed by glycine (red dots in Fig. **3B**, top left panel). The same trend is observed for serine (Fig. **3B**, bottom left panel).

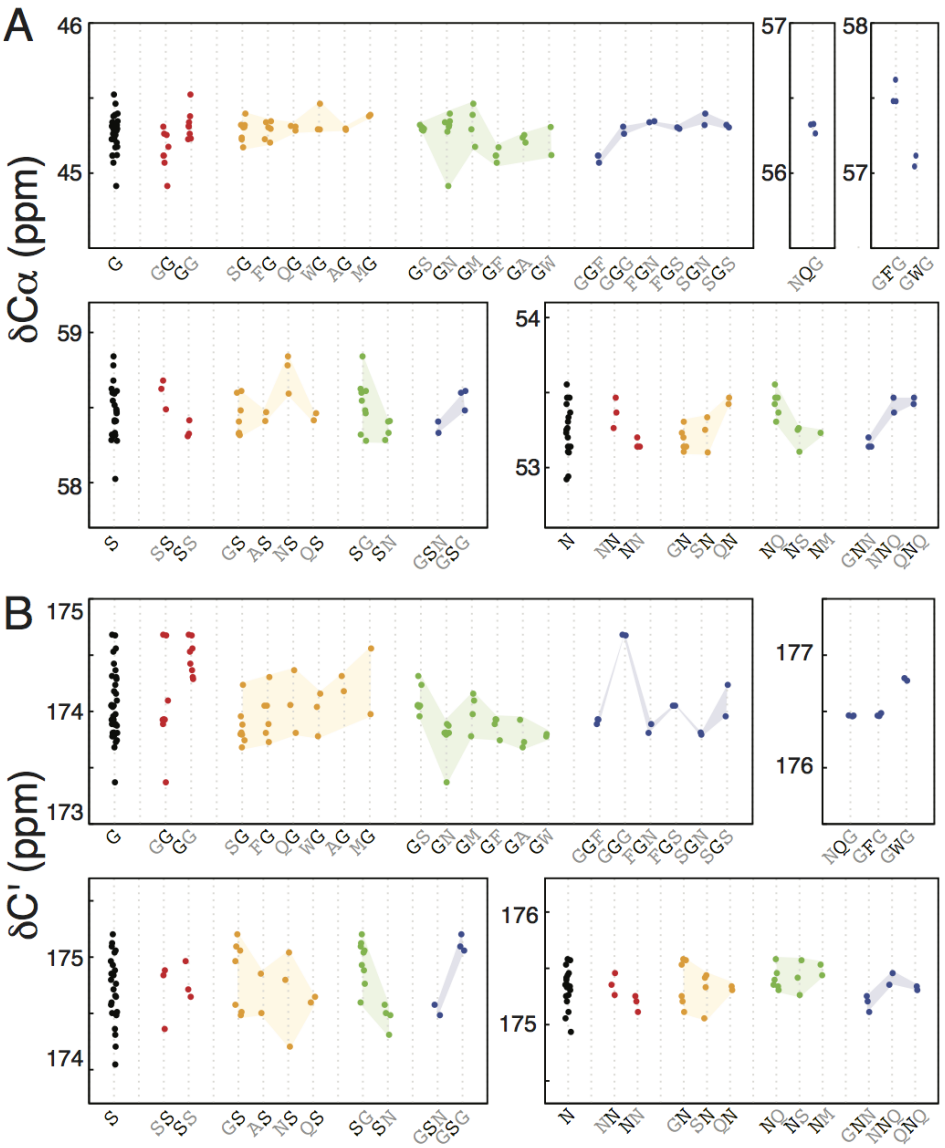

Figure 3. The nearest-neighbor effect on chemical shifts. (A) Cα and (B) C' chemical shifts of the glycine (G, top left panels), serine (S, bottom left panels), and asparagine (N, bottom right panels) residues in TDP-43$^{266–414}$ as recorded in 8 M urea at pH 2.5. The black dots show all the values for each amino acid type while colored dots are used to show the same values but grouped according to motifs that appear several times in the sequence: red dots for XX doublets (X = G, S, or N), orange dots for ZX doublets (Z ≠ G, S, or N), green dots for XZ doublets, and blue dots for ZXB triplets (B ≠ G, S, or N). A random offset has been added along the *x*-axis for the sake of clarity. (Color version online)

Qualitatively, Fig. **3B** shows that in general the chemical shifts of a given amino-acid type preceded by a given residue (yellow shaded region) vary more than do those of the same amino-acid type followed by the same residue (green shade). The trend for the Cα chemical shifts is less obvious (yellow and green shades in Fig. **3A**), probably because of the more central location of the Cα atom. More chemical shifts measured under the same conditions would be required to verify this observation quantitatively but more importantly, when both neighbors are taken into account, the chemical shift variation is substantially reduced in all cases (blue dots in Fig. **3**). Specifically, the Cα and C' chemical shifts from identical triplet motifs differ respectively by less than 0.1 ppm in all but one case (namely the C' shifts in SGS triplets). The same trend is observed for the central residues of the other three triplets (NQG, GFG, and GWG) that appear more than once in the sequence (right-most panels in Fig. **3**).

This analysis indicates that the most important factor governing the chemical shifts of a given residue in a random-coil arrangement is the nature of its neighboring residues. The chemical shifts of residues in longer identical motifs should be even more similar. Only one quintuplet (GGFGN) appears more than once in this sequence but indeed, the chemical shifts of the two central residues (F) are nearly identical (only ~0.001 ppm difference for both the Cα and C' chemical shifts).

3.4. Accounting for the sequence effect on the chemical shift may improve secondary structure predictions

We also assigned TDP-43$^{266–414}$ under physiological conditions (in 10 mM phosphate buffer at pH 6.5, BMRB access number 26728, Fig. S2) and used the chemical shifts measured in 8 M urea pH 2.5, the so-called "intrinsic chemical shifts" [51, 52], to calculate secondary chemical shifts. The increases in the Cα and C' shifts (black lines in Fig. **4**) and the decrease in the Cβ shift (data not shown) in the N-terminal part and the middle part of TDP-43$^{266–414}$ reflect the presence of α-helical components. This result is in qualitative agreement with a very recent study [53] in which the assignment was performed using data acquired in pure water at pH 4.0 (vs pH 6.5 in a phosphate buffer here).

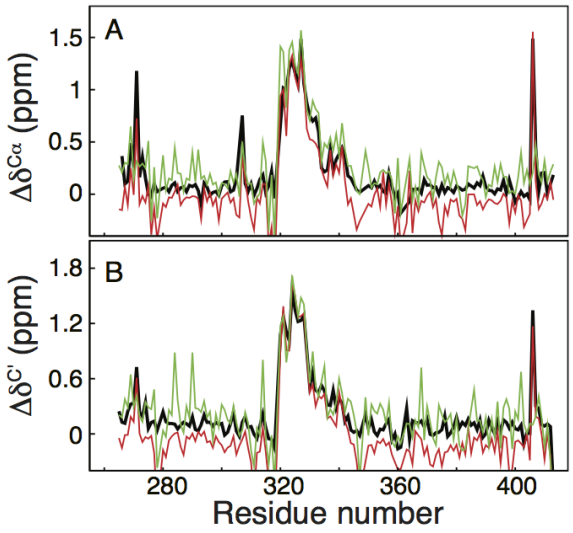

Figure 4. Secondary chemical shifts in TDP-43$^{266-414}$ (A) Cα and (B) C' secondary chemical shifts obtained by direct subtraction of the chemical shifts recorded in 8 M urea from those measured under physiological conditions (black lines), and as predicted using the random-coil database of Schwarzinger *et al.* (red lines) and of Tamiola *et al.* (green lines). (Color version online)

We also calculated the secondary chemical shifts using two different random coil databases (red and green lines in Fig. **4**). These data are noisier than those derived from the intrinsic referencing method. This observation suggests that accounting for the nearest-neighbor effect on the chemical shifts, as illustrated in Fig. **3**, can improve secondary structure predictions for disordered proteins.

4. DISCUSSION

The chemical shift is the most readily obtained parameter in biomolecular NMR studies and secondary chemical shift analysis provides useful secondary structure information for both ordered and disordered systems [15]. Since chemical shifts can be measured very precisely, the secondary chemical shift obtained depends mainly on the database of random-coil chemical shifts chosen as a basis. The differences between the different databases may be without consequence for folded proteins but become critical for unfolded proteins, for which the variations to be identified are more subtle. The effects of neighboring residues on random-coil chemical shifts have been studied previously using different approaches. Wishart et al. studied the effect of alanine and proline in Ac-GGXAGG-NH_2 and Ac-GGXPGG-NH_2 peptides (with X being one of the 20 amino acids) [19]. Schwarzinger et al. tabulated sequence-dependent correction factors from a systematic study of Ac-GGXGG-NH_2 host-guest peptides [27]. Similarly, Kjaergaard and Poulsen [54] and Prestegard et al. [28] have derived correction factors from neighboring residues in glutamine- or alanine-flanking host-guest systems. In a contrasting approach, Tamiola et al. derived nearest-neighbor correction factors from a database of chemical shifts from intrinsically disordered proteins [23].

Directly measuring the nearest-neighbor effect for all the 8000 different tripeptide combinations is not realistic however using a host-guest approach (e.g. $GGGX_1X_2X_3GGG$; with X_1, X_2, and X_3 each representing one of the 20 amino acids). To demonstrate that in a random coil, apart from the nature of the residue itself, its nearest neighbors are the main factor governing its chemical shifts, we analyzed the chemical shifts measured for a highly repetitive protein sequence, the C-terminal domain of TDP-43 (Fig. **1** and Table **1**). After assigning the chemical shifts of the backbone atoms (Fig. **2**), we systematically investigated their dependence on the nature of the preceding and succeeding amino acid (Fig. **3**). Our results show that this dependence is strong and accounts for most of the variability of the chemical shifts for a given amino acid type in a disordered structure. Furthermore, taking this effect into account provides secondary chemical shift values that are less variable (noisy) than those derived from databases are (Fig. **4**).

These results also suggest a manageable procedure to construct a random-coil chemical shift database that would account for this effect. Indeed, since two nearest neighbors have the strongest influence, rather than constructing host-guest peptides for each possible triplet combination, these 8000 motifs can be assembled into several ~100-residue-long sequences. A 102-residue polypeptide can host 100 different triplets (the last two residues in a triplet becoming the first two residues of the following triplet) so all 8000 combinations would fit into 80 of these chains. Purifying and assigning the chemical shifts of 80 different isotope-labeled 100-residue polypeptides at low pH in urea should be time-consuming but straightforward, as demonstrated in this article.

To demonstrate how these triplets could be arranged, we counted and sorted all the tripeptide motifs (~8 billion in total) that appear in the NCBI non-redundant protein sequence library and assembled them into 102-residue polypeptides. The motifs were added sequentially, avoiding repeats unless none of the unused tripeptides could fit, in which case the matching triplet with the highest natural occurrence was added again. Using this algorithm, 123 102-residue polypeptides were required for each of the 8000 triplets to be included at least once. (Note that the tripeptides that appear several times could be used to estimate errors.) In addition, since the 4000 triplets with the highest occurrence account for 80% of all protein sequences (Fig. **5**), nearly complete sequence coverage could be achieved using just 50 102-residue polypeptides.

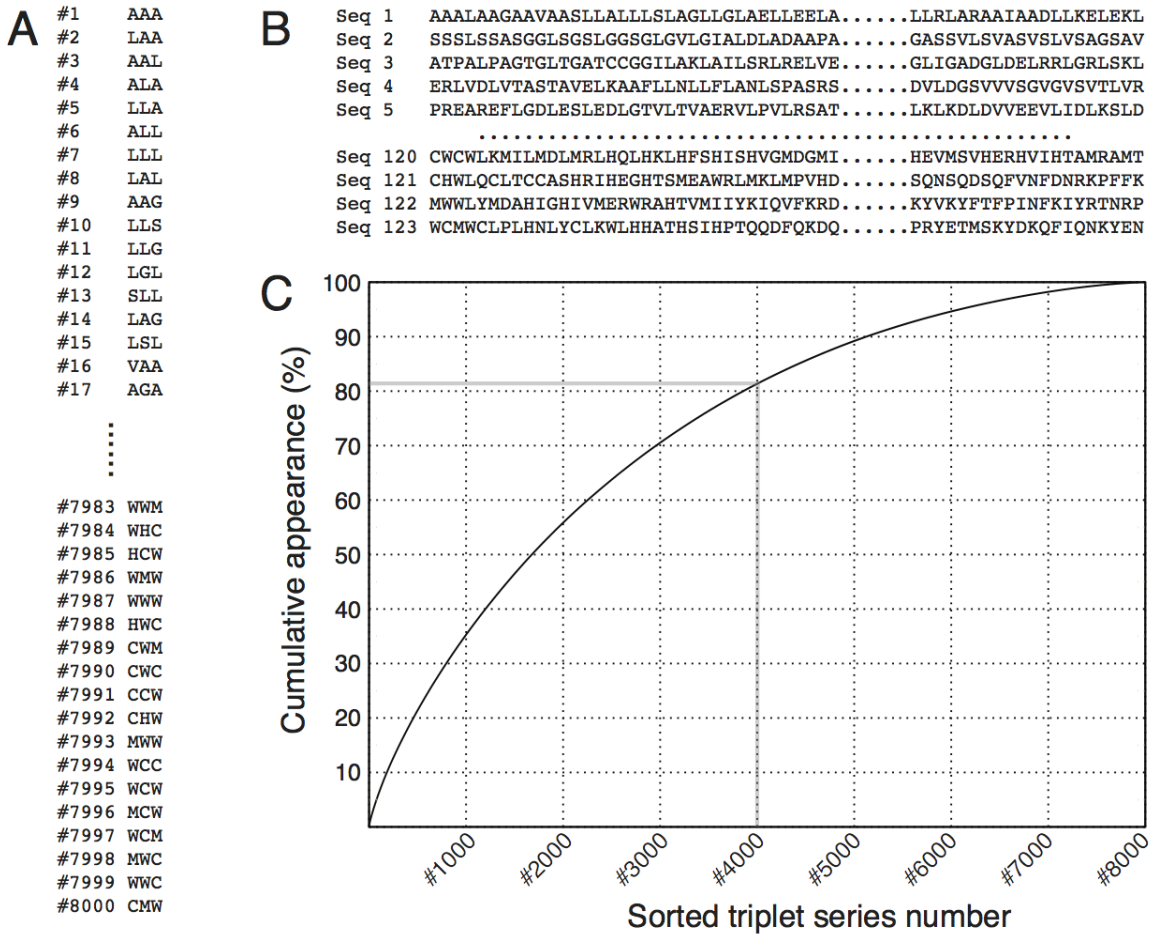

Figure 5. Assembling the minimum number of 102-residue polypeptides that include all 8000 amino-acid triplets. (A) All the tripeptides in the NCBI non-redundant protein sequence database are sorted from #1 to #8000 according to their occurrence rate. (B) The first five and last four 102-residue polypeptides assembled by sequentially adding either an unused triplet, or if none of these matched, the highest-occurring matching triplet. (C) The cumulative appearance rate of the sorted tripiptides. Percentage of all the sequences in the NCBI database covered vs the number of tripeptides (ranked by occurrence rate) included in the analysis.

Around seven million chemical shift values, covering 277 different atom types in 20 amino-acids, are deposited in the BRMB [55]. However, a relatively small number of dataset were collected from disordered proteins; only around 150 entries belong to the "Unfolded Protein" category in the BMRB, and most of them were collected in different experimental conditions. The limited number of chemical shift values and different experimental conditions impede a systematic analysis of the effect on the random-coil chemical shift from neighboring residues. Mulder and co-workers applied a delicate mathematical algorithm to derive a random coil NMR chemical shift library from manually selected (criteria such as pH from 4.5 to 7.4 and temperature from 7 °C to 31 °C) 14 intrinsically disordered proteins to overcome this challenge [23]. In contrast, we propose a "brute-force" approach to overcome the limit number of available data. According to our observation of the denatured protein with low sequence complexity, we confirm that the two nearest neighbors are determinant for the random coil chemical shift values. Therefore, we can arrange all 8000 triplets into a manageable number of polypeptide and collect NMR data in identical strong denaturing conditions. The conditions, 8 M urea and pH 2.5, are commonly used in model peptide approaches [19, 20]. These conditions ensure that the polypeptide is completely denatured, to eliminate the potential intrinsic structural propensity. In such strong acid condition, however, the chemical shifts of charged residues differ significantly from physiological conditions. Correcting factors to the pH such as Kjaergaard et al. have studied [21] may be applied to improve our random coil chemical shift library.

CONCLUSION

We have demonstrated that the amino-acid type of the two nearest neighbors is the most important factor, other than the amino-acid type of the residue itself, in determining the chemical shifts of a residue in a random coil. Our analysis suggests it should be possible to construct a random coil database that takes this effect into account, thereby improving secondary-structure predictions. An improved database of random-coil chemical shifts should also facilitate the chemical shift assignment process, especially for intrinsically disordered proteins, by providing good initial guesses.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

CONFLICT OF INTEREST

We have no conflict of interest to declare.

ACKNOWLEDGEMENTS

We thank Dr. S.G. Hyberts (Harvard University, USA) for providing the program for reconstructing NUS data, and Prof. T.H. Lin (National Yang-Ming University, Taiwan) for NMR spectrometer access. This work was supported by the MOST of Taiwan under the grants of 102-2113-M-010-003-MY2 and 104-2113-M-010-001-MY2.

NOTE ADDED IN PROOF

During the reviewing process of this manuscript, Conicella et al [56] published an article discuss the effect of ALS-related mutation of TDP-43C on the protein phase separation. In their article, they have deposited the NMR chemical assignment of TDP-43C under physiological conditions. We compared our physiological condition assignment (Fig. **4**, BMRB 26728) and theirs (BMRB 26823), and noticed that some assignments are not the same from the ^{15}N-^{1}H correlation spectrum: G310/G296, G287/G304, G288/G400, G380/G402, N390/M359, F316/Q327, A325/A326. This might be due to condition or construct difference.

REFERENCES

[1] Grzesiek, S.; Sass, H.J. From biomolecular structure to functional understanding: new NMR developments narrow the gap. *Curr. Opin. Struct. Biol*., **2009**, *19*(5), 585-95.

[2] Kay, L.E. New Views of Functionally Dynamic Proteins by Solution NMR Spectroscopy. *J. Mol. Biol*., **2016**, *428*(2 Pt A), 323-31.

[3] Case, D.a.,Chemical shifts in biomolecules. *Curr. Opin. Struct. Biol*., **2013**, 23(2), 1-5.

[4] Han, B.; Liu, Y.; Ginzinger, S.W.; Wishart, D.S. SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR*, **2011**, *50*(1), 43-57.

[5] Shen, Y.; Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR*, **2010**, *48*(1), 13-22.

[6] Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J.M.; Liu, G.; Eletsky, A.; Wu, Y.; Singarapu, K.K.; Lemak, A.; Ignatchenko, A.; Arrowsmith, C.H.; Szyperski, T.; Montelione, G.T.; Baker, D.; Bax, A. Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci. USA*, **2008**, *105*(12), 4685-90.

[7] Wishart, D.S.; Arndt, D.; Berjanskii, M.; Tang, P.; Zhou, J.; Lin, G. CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. *Nucleic Acids Res*., **2008**, *36*(Web Server issue), W496-502.

[8] Cavalli, A.; Salvatella, X.; Dobson, C.M.; Vendruscolo, M. Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci. USA*, **2007**, *104*(23), 9615-20.

[9] Mielke, S.P.; Krishnan, V.V. Characterization of protein secondary structure from NMR chemical shifts. *Prog. Nucl. Magn. Reson. Spectrosc*., **2009**, *54*(3-4), 141-165.

[10] Wishart, D.S.; Sykes, B.D. Chemical shifts as a tool for structure determination. *Methods Enzymol*., **1994**, *239*, 363-92.

[11] Spera, S.; Bax, A. Empirical correlation between protein backbone conformation and Ca and Cb 13C nuclear magnetic resonance chemical shifts. *J. Am. Chem. Soc*., **1991**, *113*, 5490-5492.

[12] Zerko, S.; Kozminski, W. Six- and seven-dimensional experiments by combination of sparse random sampling and projection spectroscopy dedicated for backbone resonance assignment of intrinsically disordered proteins. *J. Biomol. NMR*, **2015**, *63*(3), 283-90.

[13] Yoshimura, Y.; Kulminskaya, N.V.; Mulder, F.A. Easy and unambiguous sequential assignments of intrinsically disordered proteins by correlating the backbone 15N or 13C' chemical shifts of multiple contiguous residues in highly resolved 3D spectra. *J. Biomol. NMR*, **2015**, *61*(2), 109-21.

[14] Wiedemann, C.; Goradia, N.; Häfner, S.; Herbst, C.; Görlach, M.; Ohlenschläger, O.; Ramachandran, R. HN-NCA heteronuclear TOCSY-NH experiment for (1)H(N) and (15)N sequential correlations in ((13)C, (15)N) labelled intrinsically disordered proteins. *J. Biomol. NMR*, **2015**, *63*(2), 201-12.

[15] Dyson, H.J.; Wright, P.E. Unfolded proteins and protein folding studied by NMR. *Chem. Rev*., **2004**, *104*(8), 3607-22.

[16] Religa, T.L.; Markson, J.S.; Mayor, U.; Freund, S.M.; Fersht, A.R. Solution structure of a protein denatured state and folding intermediate. *Nature*, **2005**, *437*(7061), 1053-6.

[17] Wright, P.E.; Dyson, H.J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol*., **2015**, *16*(1), 18-29.

[18] Jensen, M.R.; Zweckstetter, M.; Huang, J.R.; Blackledge, M. Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. *Chem. Rev*., **2014**, *114*(13), 6632-60.

[19] Wishart, D.S.; Bigam, C.G.; Holm, A.; Hodges, R.S.; Sykes, B.D. (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*, **1995**, *5*(3), 332.

[20] Schwarzinger, S.; Kroon, G.J.; Foss, T.R.; Wright, P.E.; Dyson, H.J. Random coil chemical shifts in acidic 8 M urea: implementation of random coil shift data in NMRView. *J. Biomol. NMR*, **2000**, *18*(1), 43-8.

[21] Kjaergaard, M.; Brander, S.; Poulsen, F.M. Random coil chemical shift for intrinsically disordered proteins: effects of temperature and pH. *J. Biomol. NMR*, **2011**, *49*(2), 139-49.

[22] De Simone, A.; Cavalli, A.; Hsu, S.T.; Vranken, W.; Vendruscolo, M. Accurate random coil chemical shifts from an analysis of loop regions in native states of proteins. *J. Am. Chem. Soc*., **2009**, *131*(45), 16332-3.

[23] Tamiola, K.; Acar, B.; Mulder, F.A. Sequence-specific random coil chemical shifts of intrinsically disordered proteins. *J. Am. Chem. Soc*., **2010**, *132*(51), 18000-3.

[24] Zhang, H.; Neal, S.; Wishart, D.S. RefDB: a database of uniformly referenced protein chemical shifts. *J. Biomol. NMR*, **2003**, *25*(3), 173-95.

[25] Marsh, J.A.; Singh, V.K.; Jia, Z.; Forman-Kay, J.D. Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. *Protein Sci*., **2006**, *15*(12), 2795-804.

[26] Camilloni, C.; De Simone, A.; Vranken, W.F.; Vendruscolo, M. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry*, **2012**, *51*(11), 2224-31.

[27] Schwarzinger, S.; Kroon, G.J.; Foss, T.R.; Chung, J.; Wright, P.E.; Dyson, H.J. Sequence-dependent correction of random coil NMR chemical shifts. *J. Am. Chem. Soc*., **2001**, *123*(13), 2970-8.

[28] Prestegard, J.H.; Sahu, S.C.; Nkari, W.K.; Morris, L.C.; Live, D.; Gruta, C. Chemical shift prediction for denatured proteins. *J. Biomol. NMR*, **2013**, *55*(2), 201-9.

[29] Delaglio, F.; Grzesiek, S.; Vuister, G.W.; Zhu, G.; Pfeifer, J.; Bax, A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, **1995**, *6*(3), 277-93.

[30] Goddard, T.D.; Kneller, D.G. *Sparky 3. San Francisco, University of California*. **2005**.

[31] Hyberts, S.G.; Milbradt, A.G.; Wagner, A.B.; Arthanari, H.; Wagner, G. Application of iterative soft thresholding for fast reconstruction of NMR data non-uniformly sampled with multidimensional Poisson Gap scheduling. *J. Biomol. NMR*, **2012**, *52*(4), 315-27.

[32] Felli, I.C.; Pierattelli, R. Novel methods based on (13)C detection to study intrinsically disordered proteins. *J. Magn. Reson*., **2014**, *241*, 115-25.

[33] Wootton, J.C. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem*., **1994**, *18*(3), 269-85.

[34] Radivojac, P.; Obradović, Z.; Brown, C.J.; Dunker, A.K. Prediction of boundaries between intrinsically ordered and disordered protein regions. *Pac. Symp. Biocomput*., **2003**, 216-27.

[35] Obradovic, Z.; Peng, K.; Vucetic, S.; Radivojac, P.; Dunker, A.K. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, **2005**, *61*(Suppl 7), 176-82.

[36] Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **2005**, *21*(16), 3433-4.

[37] Otaki, J.M.; Tsutsumi, M.; Gotoh, T.; Yamamoto, H. Secondary structure characterization based on amino acid composition and availability in proteins. *J. Chem. Inf. Model*., **2010**, *50*(4), 690-700.

[38] Varadi, M.; Zsolyomi, F.; Guharoy, M.; Tompa, P. Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. *PLoS One*, **2015**, *10*(10), e0139731.

[39] Calabretta, S.; Richard, S. Emerging Roles of Disordered Sequences in RNA-Binding Proteins. *Trends Biochem. Sci*., **2015**, *40*(11), 662-72.

[40] Lee, E.B.; Lee, V.M.; Trojanowski, J.Q. Gains or losses: molecular mechanisms of TDP43-mediated neurodegeneration. *Nat. Rev. Neurosci*., **2012**, *13*(1), 38-50.

[41] Neumann, M.; Sampathu, D.M.; Kwong, L.K.; Truax, A.C.; Micsenyi, M.C.; Chou, T.T.; Bruce, J.; Schuck, T.; Grossman, M.; Clark, C.M.; McCluskey, L.F.; Miller, B.L.; Masliah, E.; Mackenzie, I.R.; Feldman, H.; Feiden, W.; Kretzschmar, H.A.; Trojanowski, J.Q.; Lee, V.M. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science*, **2006**, *314*(5796), 130-3.

[42] Lukavsky, P.J.; Daujotyte, D.; Tollervey, J.R.; Ule, J.; Stuani, C.; Buratti, E.; Baralle, F.E.; Damberger, F.F.; Allain, F.H. Molecular basis of UG-rich RNA recognition by the human splicing factor TDP-43. *Nat. Struct. Mol. Biol*., **2013**, *20*(12), 1443-9.

[43] Austin, J.A.; Wright, G.S.; Watanabe, S.; Grossmann, J.G.; Antonyuk, S.V.; Yamanaka, K.; Hasnain, S.S. Disease causing mutants of TDP-43 nucleic acid binding domains are resistant to aggregation and have increased stability and half-life. *Proc. Natl. Acad. Sci. USA*, **2014**, *111*(11), 4309-14.

[44] Mompean, M.; Romano, V.; Pantoja-Uceda, D.; Stuani, C.; Baralle, F.E.; Buratti, E.; Laurents, D.V. The TDP-43 N-terminal domain structure at high resolution. *FEBS J*., **2016**, *283*(7), 1242-60.

[45] Qin, H.; Lim, L.Z.; Wei, Y.; Song, J. TDP-43 N terminus encodes a novel ubiquitin-like fold and its unfolded form in equilibrium that can be shifted by binding to ssDNA. *Proc. Natl. Acad. Sci. USA*, **2014**, *111*(52), 18619-24.

[46] Chen, A.K.; Lin, R.Y.; Hsieh, E.Z.; Tu, P.H.; Chen, R.P.; Liao, T.Y.; Chen, W.; Wang, C.H.; Huang, J.J. Induction of amyloid fibrils by the C-terminal fragments of TDP-43 in amyotrophic lateral sclerosis. *J. Am. Chem. Soc*., **2010**, *132*(4), 1186-7.

[47] Wang, I.F.; Chang, H.Y.; Hou, S.C.; Liou, G.G.; Way, T.D.; James Shen, C.K. The self-interaction of native TDP-43 C terminus inhibits its degradation and contributes to early proteinopathies. *Nat. Commun*., **2012**, *3*, 766.

[48] Wang, Y.T.; Kuo, P.H.; Chiang, C.H.; Liang, J.R.; Chen, Y.R.; Wang, S.; Shen, J.C.; Yuan, H.S. The truncated C-terminal RNA recognition motif of TDP-43 protein plays a key role in forming proteinaceous aggregates. *J. Biol. Chem*., **2013**, *288*(13), 9049-57.

[49] Grzesiek, S.; Bax, A. Amino acid type determination in the sequential assignment procedure of uniformly 13C/15N-enriched proteins. *J. Biomol. NMR*, **1993**, *3*(2), 185-204.

[50] Jung, Y.S.; Zweckstetter, M. Mars -- robust automatic backbone assignment of proteins. *J. Biomol. NMR*, **2004**, *30*(1), 11-23.

[51] Modig, K.; Jürgensen, V.W.; Lindorff-Larsen, K.; Fieber, W.; Bohr, H.G.; Poulsen, F.M. Detection of initiation sites in protein folding of the four helix bundle ACBP by chemical shift analysis. *FEBS Lett*., **2007**, *581*(25), 4965-71.

[52] Kjaergaard, M.; Nørholm, A.B.; Hendus-Altenburger, R.; Pedersen, S.F.; Poulsen, F.M.; Kragelund, B.B. Temperature-dependent structural changes in intrinsically disordered proteins: formation of alpha-helices or loss of polyproline II? *Protein Sci*., **2010**, *19*(8), 1555-64.

[53] Lim, L.; Wei, Y.; Lu, Y.; Song, J.; ALS-Causing Mutations Significantly Perturb the Self-Assembly and Interaction with Nucleic Acid of the Intrinsically Disordered Prion-Like Domain of TDP-43. *PLoS Biol*., **2016**, *14*(1), e1002338.

[54] Kjaergaard, M.; Poulsen, F.M. Sequence correction of random coil chemical shifts: correlation between neighbor correction factors and changes in the Ramachandran distribution. *J. Biomol. NMR*, **2011**, *50*(2), 157-65.

[55] Ulrich, E.L.; Akutsu, H.; Doreleijers, J.F.; Harano, Y.; Ioannidis, Y.E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C.F.; Tolmie, D.E.; Kent Wenger, R.; Yao, H.; Markley, J.L. BioMagResBank. *Nucleic Acids Res*., **2008**, *36*(Database issue), D402-8.

[56] Conicella, A.E.; Zerze, G.H.; Mittal, J.; Fawzi, N.L. ALS Mutations Disrupt Phase Separation Mediated by alpha-Helical Structure in the TDP-43 Low-Complexity C-Terminal Domain. *Structure*, **2016**, *24*(9), 1537-49.