TDP-43 N terminus encodes a novel ubiquitin-like fold and its unfolded form in equilibrium that can be shifted by binding to ssDNA

Haina Qin[a,1], Liang-Zhong Lim[a,1], Yuanyuan Wei[b,1], and Jianxing Song[a,b,2]

[a]Department of Biological Sciences, Faculty of Science, and [b]National University of Singapore Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore 119260

Edited by David Baker, University of Washington, Seattle, WA, and approved November 19, 2014 (received for review July 23, 2014)

Transactivation response element (TAR) DNA-binding protein 43 (TDP-43) is the principal component of ubiquitinated inclusions characteristic of most forms of amyotrophic lateral sclerosis (ALS) and frontotemporal dementia-frontotemporal lobar degeneration with TDP-43–positive inclusions (FTLD-TDP), as well as an increasing spectrum of other neurodegenerative diseases. Previous structural and functional studies on TDP-43 have been mostly focused on its recognized domains. Very recently, however, its extreme N terminus was identified to be a double-edged sword indispensable for both physiology and proteinopathy, but thus far its structure remains unknown due to the severe aggregation. Here as facilitated by our previous discovery that protein aggregation can be significantly minimized by reducing salt concentrations, by circular dichroism and NMR spectroscopy we revealed that the TDP-43 N terminus encodes a well-folded structure in concentration-dependent equilibrium with its unfolded form. Despite previous failure in detecting any sequence homology to ubiquitin, the folded state was determined to adopt a novel ubiquitin-like fold by the CS-Rosetta program with NMR chemical shifts and 78 unambiguous long-range nuclear Overhauser effect (NOE) constraints. Remarkably, this ubiquitin-like fold could bind ssDNA, and the binding shifted the conformational equilibrium toward reducing the unfolded population. To the best of our knowledge, the TDP-43 N terminus represents the first ubiquitin-like fold capable of directly binding nucleic acid. Our results provide a molecular mechanism rationalizing the functional dichotomy of TDP-43 and might also shed light on the formation and dynamics of cellular ribonucleoprotein granules, which have been recently linked to ALS pathogenesis. As a consequence, one therapeutic strategy for TDP-43–causing diseases might be to stabilize its ubiquitin-like fold by ssDNA or designed molecules.

amyotrophic lateral sclerosis | FTLD-TDP | TDP-43 | NMR spectroscopy | ubiquitin-like fold

In 2006, transactivation response element (TAR) DNA-binding protein 43 (TDP-43) was identified as the major constituent of the proteinaceous inclusions that are characteristic of most forms of amyotrophic lateral sclerosis (ALS) and the most common pathological subtype of frontotemporal dementia-frontotemporal lobar degeneration with TDP-43–positive inclusions (FTLD-TDP) (1, 2). Since then, numerous studies have confirmed that TDP43 protein is mechanistically linked to neurodegeneration (3, 4). TDP43 is a 414-residue protein that has been previously recognized to be composed of a nuclear localization signal (NLS), two RNA recognition motifs (RRM1 and RRM2) hosting a nuclear export signal (NES), and C-terminal glycine-rich prion-like domain (Fig. 1*A*). The NLS and NES regulate the shuttling of TDP-43 between the nucleus and the cytoplasm (5), whereas the RRM1 and RRM2 are responsible for binding to nucleic acids including single- or double-stranded DNA/RNA (5–8). The prion-like domain mediates protein–protein interactions between TDP-43 and other hnRNP members (9), which also hosts most known ALS-associated TDP-43 mutations. TDP43 is an aggregation-prone protein (1–4, 10–14), and its abnormal aggregation has been found in ∼97% ALS and ∼45% frontotemporal dementia (FTD) patients. Additionally, TDP-43 immunoreactive inclusions have also been observed in an increasing spectrum of other neurodegenerative disorders, which include ALS/parkinsonism–dementia complex of Guam, Alzheimer's disease (AD), dementia with Lewy bodies (DLB), Pick's disease, argyrophilic grain disease, and corticobasal degeneration (3, 14).

Previous investigations have been mostly focused on the recognized domains of TDP-43. As a consequence, the functional roles of the extreme N terminus remain largely uncharacterized. Only very recently has it been identified that the extreme N terminus (residues 1–75) functions as a double-edged sword: it not only regulates normal TDP-43 functions, but also drives neurodegeneration in TDP-43 proteinopathies (15). More specifically, the deletion of the N terminus, or even the first nine residues, is sufficient to abolish the TDP-43–regulated RNA splicing, as well as the aggregation of the full-length TDP-43, which is required for impaired neurite outgrowth (15). In this regard, the elucidation of the structure of the TDP-43 N terminus would provide valuable insights into the molecular mechanism underlying this intriguing dichotomy. Unfortunately, despite intense efforts, high-quality NMR data could not be acquired for the TDP-43 N terminus due to its severe aggregation (16).

Here as facilitated by our discovery in 2005 that protein aggregation can be significantly minimized by reducing salt

Significance

Transactivation response element (TAR) DNA-binding protein 43 (TDP-43) inclusion is a histological hallmark of FTLD-TDP and amyotrophic lateral sclerosis. Its N terminus was just revealed as a double-edged sword indispensable for both physiology and proteinopathy, but its structure remains unknown due to aggregation. Here we revealed (*i*) the TDP-43 N terminus encodes a well-folded structure in equilibrium with its unfolded form; (*ii*) despite previous failure in detecting sequence homology to ubiquitin, the folded state assumes a novel ubiquitin-like fold; and (*iii*) this ubiquitin-like fold could bind ssDNA, thus representing the first capable of directly binding nucleic acid. Taken together, our results provide a molecular mechanism rationalizing the functional dichotomy of TDP-43 and further imply one therapeutic strategy for TDP-43–causing diseases.

Author contributions: J.S. designed research; H.Q., L.-Z.L., Y.W., and J.S. performed research; H.Q., L.-Z.L., Y.W., and J.S. analyzed data; and J.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]H.Q., L.-Z.L., and Y.W. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: dbssjx@nus.edu.sg.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1413994112/-/DCSupplemental.

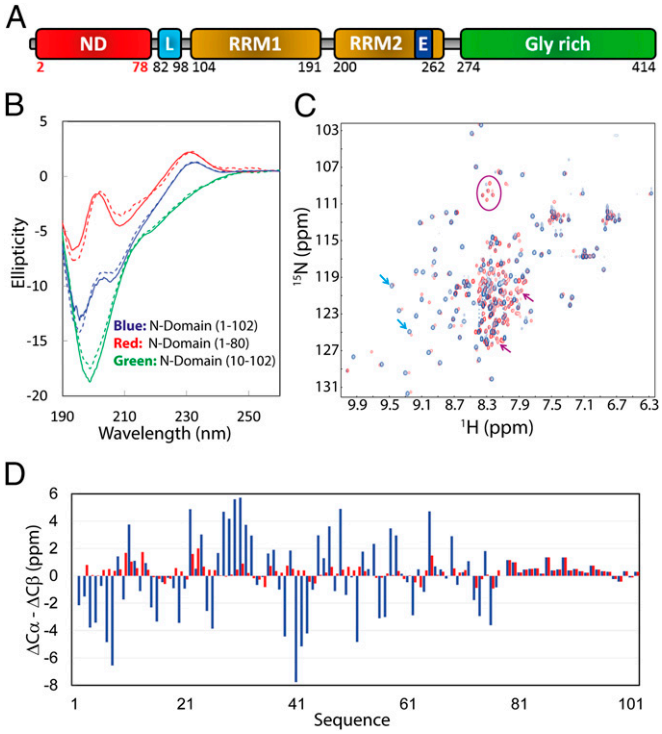

Fig. 1. Characterization of two coexisting conformations adopted by the TDP-43 N-domain. (*A*) Domain organization of the 414-residue TDP-43 protein, which is composed of the N-domain determined in the present study, nuclear localization signal (L), two RNA recognition motifs (RRM1 and RRM2) hosting a nuclear export signal (E), and C-terminal glycine-rich domain. (*B*) Far-UV CD spectra at protein concentrations of 15 μM of the N-domain (1–102) in Milli-Q water at pH 4.0 (blue line) and in 1 mM phosphate buffer at pH 7.5 (blue dotted line); N-domain (1–80) in Milli-Q water at pH 4.0 (red line) and in 1 mM phosphate buffer at pH 7.5 (red dotted line); and N-domain (10–102) in Milli-Q water at pH 4.0 (green line) and in 1 mM phosphate buffer at pH 7.5 (green dotted line). (*C*) Superimposition of the 2D NMR ^{1}H-^{15}N HSQC spectra of the N-domain (1–102) in Milli-Q water at pH 4.0 at a at protein concentration of 40 μM (blue) and 1 mM (red). Cyan arrows are used to indicate the peaks from the well-folded form whose relative intensities are much higher at 40 μM (blue) than those at 1 mM (red). Purple arrows and oval are used to indicate the peaks from the unfolded form whose relative intensities are much lower at 40 μM (blue) than those at 1 mM (red). (*D*) Residue specific (ΔCα–ΔCβ) chemical shifts of the N-domain (1–102) in the folded (blue) and unfolded (red) forms.

concentrations in aqueous solution (17, 18), we successfully conducted extensive biophysical studies of differentially truncated TDP-43 N-domains including determination of its 3D structure. The study revealed that the TDP-43 extreme N terminus encodes two coexisting conformations in equilibrium: one well-folded and another highly disordered. Unexpectedly, despite low sequence homology, the well-folded state has been determined to adopt a novel ubiquitin-like fold, which represents, to our knowledge, the first capable of directly binding ssDNA. Remarkably, the binding to ssDNA shifted the conformational equilibrium toward reducing the unfolded population. Taken together, the present results provide a mechanism rationalizing the functional dichotomy of TDP-43 and might also shed light on the formation and dynamics of cellular ribonucleoprotein (RNP) granules, which have been recently linked to ALS pathogenesis. As a consequence, one therapeutic strategy for TDP-43–causing diseases might be to stabilize its ubiquitin-like fold by ssDNA or designed molecules.

Results

The TDP-43 N Terminus Encodes Two Coexisting Conformations in Equilibrium. As shown in Fig. 1*B*, similar to the previous report (16), the N-domain (1–102) has very unusual far-UV circular dichroism (CD) spectra with the maximal negative signal at 196 nm, an additional negative signal at 206 nm, and a positive signal at 232 nm but no positive signal below 200 nm, which are very similar in both salt-minimized aqueous solution (Milli-Q water, pH 4.0) and 1 mM phosphate buffer (pH 7.5). Most importantly, in Milli-Q water, we were able to acquire high-quality heteronuclear single quantum correlation (HSQC) spectra with protein concentrations up to 1 mM. Comparison of its HSQC spectra at different protein concentrations immediately revealed that the TDP-43 N-domain encodes two coexisting conformations in equilibrium: the well-folded form with large HSQC spectral dispersions and the predominantly unfolded one with narrow dispersions (Fig. 1*C*). Interestingly, the equilibrium of two conformations appears to be concentration dependent: at a protein concentration of 40 μM, the intensity of HSQC peaks from the folded state is much higher than that from the unfolded one. By contrast, at 1 mM, the unfolded state has much higher intensity for HSQC peaks (Fig. 1*C*). The coexistence of the folded and unfolded states under native conditions has not been frequently observed as a structured protein is believed to require a sufficient stability from its unfolded state to implement biological functions. In addition to a SH3 domain (19), the TDP-43 N terminus thus provides another example showing that a native sequence can encode a folded and an unfolded state in slow exchange equilibrium, which thus bears critical implications in understanding the physiological and pathological functions of TDP-43.

By analyzing 3D NMR spectra including HN(CO)CACB and CCC(CO)NH, as well as HSQC-total correlation spectroscopy (TOCSY) and HSQC-nuclear overhauser effect spectroscopy (NOESY), we successfully achieved the sequential assignments for both states (Fig. S1). For the folded state, except for Met1 and Pro78, Cα and Cβ chemical shifts of all residues have been assigned, whereas for the unfolded state, Ser2 was additionally unassigned. The results revealed that only residues 1–78 have two set of HSQC peaks, suggesting that residues 1–78 undergo the exchange between two conformations, whereas residues 79–102 adopt only one conformation. Fig. 1*D* presents the (ΔCα–ΔCβ) chemical shifts of both states of the N-domain (1–102), which are an indicator of the secondary structures in both folded and disordered proteins (20). For the unfolded form, all 102 residues have very small (ΔCα–ΔCβ) deviations, indicating that this form is predominantly disordered, without any stable secondary structures (Fig. S2*A*). Furthermore, we analyzed the NH, N, Cα, and Cβ chemical shifts by both Delta2D (21) and secondary structure propensity (SSP) (22) programs. The results by the SSP program indicate that the unfolded form has no well-formed secondary structure over the whole N-domain (1–102) (Fig. S3*A*). However, it is interesting to note that many short segments have positive SSP scores, implying that these regions have populated helical/loop conformations. Similarly, the results by the Delta2D program also suggest that the unfolded form has random coil conformations over the whole sequence (Fig. S3*B*). Further analysis of both ^{15}N- and ^{13}C-edited NOESY spectra indicated that only several $d_{\alpha N(i,\ i+2)}$ but no long-range nuclear Overhauser effects (NOEs) could be found for the unfolded form (Fig. S2*A*). The backbone of the unfolded form appears much more flexible than that of the folded form on the picosecond-to-nanosecond (ps-ns) time scale as judged from the small or even negative heteronuclear NOEs (hNOEs), with an average value of 0.2 for all 102 residues (Fig. 2*A*) (20).

By contrast, the residues 2–78 of the folded form have very large (ΔCα–ΔCβ), characteristic of a well-folded protein (Fig. 1*D*). To confirm this, we made a construct only consisting of residues 1–80, which still has CD spectra overall similar to those of the

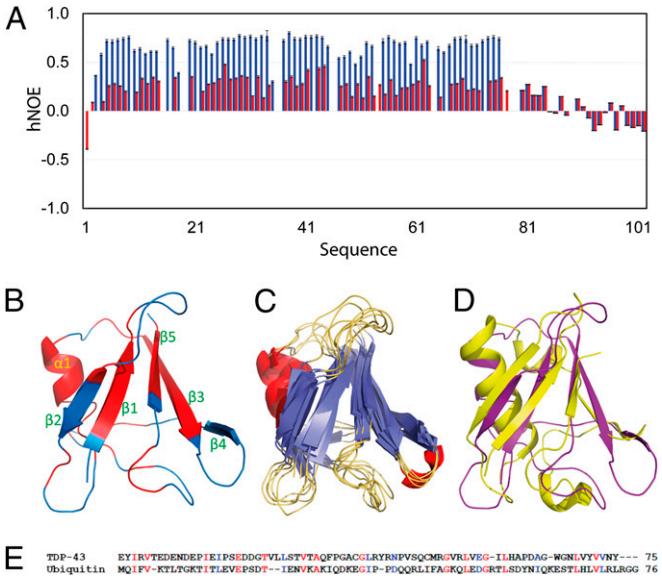

Fig. 2. NMR structure and dynamics of the TDP-43 N-domain. (*A*) {^{1}H}-^{15}N heteronuclear steady-state NOE (hNOE) of the N-domain (1–102) in the folded (blue) and unfolded (red) forms. (*B*) The lowest energy structure of the TDP-43 N-domain calculated by CS-Rosetta program with chemical shifts and unambiguous long-range NOEs. Residues having hNOE >0.7 are colored in red. (*C*) Overlay of five lowest energy structure of the TDP-43 N-domain. Blue is used for coloring β-strands, red for helix, and light yellow for loop. (*D*) Overlay of structures of the TDP-43 N-domain (purple) and ubiquitin (yellow; PDB ID code 3EHV). (*E*) Sequence alignment of TDP-43 N-domain and ubiquitin. Identical residues are colored in red and conserved in blue.

N-domain (1–102), suggesting the coexistence of the well-folded and unfolded conformations (Fig. 1*B*). We further analyzed the CD spectra by the CDPro software package, and the results showed that for the N-domain (1–80), the secondary structures are very similar at pH 4.0 and 7.5, whereas for the N-domain (1–102), slightly more turn structure is formed at pH 7.5 (Table S1), probably involved in the region over the residues 81–102. In the HSQC spectra of the N-domain (1–80), two sets of peaks could still be observed to result from the well-folded and unfolded states at a protein concentration of 500 μM (Fig. S2 *B* and *C*). However, at 40 μM, the set of HSQC peaks corresponding to the unfolded state was too weak to be detected by NMR (Fig. S2*D*), implying that the removal of the disordered region over residues 81–102 stabilized the folded state or/and destabilized the unfolded state. On the other hand, the identical residues in N-domain (1–80) and N-domain (1–102) have almost superimposable HSQC peaks at both protein concentrations of 500 μM (Fig. S2*C*) and 40 μM (Fig. S2*D*), implying that two exchanging conformations are very similar in the contexts of two constructs.

We conducted pulsed field gradient NMR self-diffusion measurements on the TDP-43 (1–102) (23). The results showed that the diffusion coefficients are almost identical (~0.80 ± 0.04 × 10^{-10} m^2/s), although the calculations have been carried out by monitoring different NMR resonances: at −0.1 and −0.21 ppm, resulting from the folded state, as well as over 7.63–8.59 ppm containing mixed NMR resonances from both folded and unfolded forms (Table S2). For a well-packed globular protein, a diffusion coefficient of 0.80 × 10^{-10} m^2/s has been characterized to have a molecular weight of ~38 kDa (23). On the one hand, this implies that the folded form of the TDP-43 N-domain (1–102) might not have a very tight packing as a classic globular protein and/or undergo the dynamic oligomerization. On the other hand, the unfolded form may represent a compact disordered state. Indeed, many residues in the unfolded form still have hNOEs close to or even larger than 0.3 (Fig. 2*A*), implying that the backbone motions on the ps-ns time scale are restricted to some degree.

Furthermore, by quantitatively analyzing the well-resolved auto- and cross-peaks of nine residues in the 3D HSQC-NOESY spectrum (24), we derived that at 25 °C and a protein concentration of 300 μM, the unfolded population is 29.75 ± 1.87%, whereas the exchange rate constant (kex) is 13.37 ± 0.27 Hz (Table S3). The high similarity of the population and kex values for nine residues implies that the whole TDP-43 (1–80) undergoes a two-state slow conformational exchange with similar kinetic parameters. In the further, it is fundamentally interesting to map out how these exchanges are mediated by protein concentrations, salt concentrations, temperatures, and pH values.

The Well-Folded State Adopts a Novel Ubiquitin-Like Fold. The well-folded state has large hNOEs for the majority of residues 1–78 (Fig. 2*A*), with an average value of 0.7, suggesting that the backbone motions in the folded form are significantly restricted. In particular, it also has a large number of medium- and long-range NOEs over residues 3–77, indicating that this form has well-formed secondary structures and tight tertiary packing. On the other hand, the cross-peaks have been identified to result from the exchange of amide protons of two states in NOESY spectra (Fig. S4*A*), thus imposing a significant challenge to assign all NOEs to calculate NMR structures with conventional methods. Indeed, we attempted to calculate NMR structures with conventional methods using dihedral angles and all unambiguous NOEs whose chemical shifts of two protons have no overlap with each other (Fig. S4*B* and Table S4). However, the obtained structures have a similar overall topology but show very large RMSDs among different structures, as exemplified by the 10 CYANA structures with the lowest target functions (Fig. S4 *C* and *D*).

As a consequence, in the present study, the CS-Rosetta program was used to generate the NMR structure of residues 3–77 of the folded form with NMR chemical shifts and a set of 78 unambiguous long-range NOEs (Fig. S4*B*). For the Rosetta methodology, a very limited set of NMR data, which are too sparse for conventional methods, is sufficient to accurately determine solution structures of proteins up to 40 kDa (25–28). In fact, the NMR data serve only to guide conformational search toward the lowest-energy conformations in the folding landscape, whereas the details of the solution structures are determined by the physical chemistry implicit in the Rosetta all-atom energy function (25–28). We calculated 2,000 Rosetta structures and selected the five lowest-energy structures for further analysis (Fig. S4 *E* and *F*). The five structures well converged into the same topology, and Fig. 2*B* presents the lowest-energy structure, which is composed of a short α-helix over residues Thr30-Gln34 and five β-strands over residues Tyr4-Thr8, Ile16-Pro19, Cys39-Asn45, Gln49-Arg52, and Val72-Asn76. The tertiary contact map of the structure is completely consistent with the long-range NOEs used for the calculation by the CS-Rosetta program (Fig. S4*E*), indicating that the structure reflects the NMR experimental constraints. Fig. 2*C* shows the superimposition of the five lowest-energy structures with the backbone RMS deviation (RMSD) of 0.85 Å over the secondary structure regions (Table S4). Subsequently, we searched the Protein Data Bank (PDB) by Dali server (29) and found 952 structure homologs (with RMSD < 3.5 Å), which are all ubiquitin-like folds. Fig. 2*D* shows the superimposition of the lowest-energy Rosetta structure of the N-domain and the crystal structure of ubiquitin at a resolution of 1.8 Å (30) (PDB ID code 3EHV), which has an overall RMSD of 2.4 Å. Despite previous failures in recognizing any sequence homology to ubiquitin, here our alignment does shows that the TDP-43 N terminus and ubiquitin have 18% identity and 27.6% homology between their sequences (Fig. 2*E*).

In the structure, the first nine residues, which were previously shown to be critical for normal functions and formation of TDP-43 inclusion (15), adopt the first β-strand to have extensive contacts with β-strands 2 and 5 (Fig. 2*B*). To explore its structural and functional roles, we made a construct N-domain (10–102), which was found to suddenly become highly soluble in the supernatant of *Escherichia coli* cell lysis. On the other hand, the N-domain (10–102) was highly disordered as judged from its CD spectra (Fig. 1*B*), as well as narrowly dispersed HSQC spectrum with only one set of peaks (Fig. S5). Remarkably, the majority of HSQC peaks of the N-domain (10–102) are largely superimposable to those of the unfolded state of the N-domain (1–102), implying that the N-domain (10–102) has a conformation similar to the unfolded form of the N-domain (1–102) over the identical 93 residues. This result underscores a complexity of the determinants for protein insolubility (18). Indeed, we found a similar case before. A transcriptional activator, ApLLP, was characterized to be insoluble in buffer and intrinsically disordered (18). However, the deletion of the 10-residue nuclear localization signal at its N terminus rendered the truncated form to become soluble in buffers. Intriguingly, this nuclear localization signal sequence even has a slightly lower hydrophobicity than the rest of the protein (18).

Therefore, an interesting question arises: why does the ubiquitin-like fold adopted by the TDP-43 N terminus have the coexistence of the folded and unfolded states? To address this, we analyzed the sequences of the TDP-43 (1–80) and ubiquitin by Database of Disordered Proteins (DISPROT) (31) and Prediction of Intrinsically Unstructured Proteins (IUPRED) (32). Although two programs give different patterns of the disorder score, both of them indicate that the N-terminal 25 residues of TDP-43 have significantly higher disorder scores than those of ubiquitin (Fig. S6), implying that the difference of the N-terminal residues of TDP-43 may at least partly contribute to the low thermodynamic stability of its folded state. Indeed, the results by the Delta2D program indicate that, based on the chemical shifts of the folded form, N-terminal residues Asn12-Glu21 all have random coil populations > 0.6 (Fig. S3*C*). However, this region contains the second β-strand over residues Ile16-Pro19, which are predicted by the Delta2D program to have low populations of the extended strand (Fig. S3*C*). In the future, it is of significant interest to introduce mutations to stabilize the folded state and subsequently to evaluate the biological/pathological consequences for these mutations.

TDP-43 Ubiquitin-Like Fold Binds ssDNA, Which Shifts the Conformational Equilibrium. Two TDP-43 RNA recognition motifs, RRM1 and RRM2, have been extensively demonstrated to primarily recognize single-stranded (TG)n/(UG)n repeats (8, 16, 33), and the TDP-43 N terminus might also be involved in the binding to nucleic acids (16). As such, we examined whether the TDP-43 N-domain (1–102) is able to bind the single-stranded $(TG)_6$ DNA. Interestingly, as seen in Fig. 3*A*, the gradual addition of $(TG)_6$ ssDNA led to the significant changes of its CD spectra: the CD signal at 195 became more positive, whereas the signal at 208 more negative. This result clearly indicates that the N-domain did bind $(TG)_6$ ssDNA, and the binding triggered the increase of the folded population. Indeed, detailed inspection of HSQC spectra in the presence of ssDNA at different ratios confirmed this. The addition of $(TG)_6$ ssDNA first reduced the HSQC peak intensity of the unfolded conformation (Fig. 3*B*). Further addition resulted in the significant broadening of many HSQC peaks without considerable perturbations of chemical shifts. On the other hand, once the molar ratio (protein:ssDNA) was increased to 1:2, almost all HSQC peaks disappeared and visible aggregates started to form in NMR tubes. Therefore, the binding of the TDP-43 N terminus to ssDNA appears to trigger its oligomerization, thus strongly

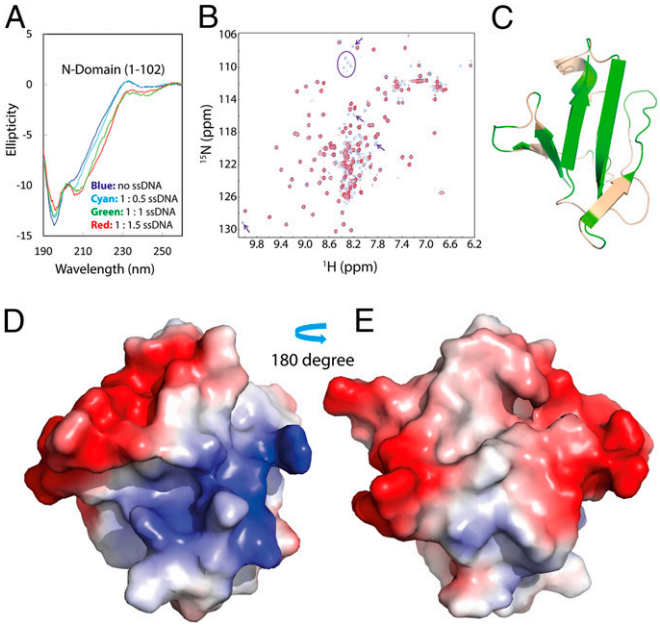

Fig. 3. Binding of ssDNA to N-domain shifts the conformational equilibrium. (*A*) Far-UV CD spectra of the N-domain (1–102) at a protein concentration of 15 μM in the absence (blue) and in the presence of the single-stranded $(TG)_6$ DNA at molar ratio of 1:0.5 (cyan), 1:1 (green), and 1:1.5 (red) (N-domain:DNA). (*B*) Superimposition of NMR ^{1}H-^{15}N HSQC spectra of the N-domain (1–102) at a protein concentration of 40 μM in the absence (blue) and in the presence of the $(TG)_6$ DNA at molar ratio of 1:1 (red). Purple arrows are used to indicate HSQC peaks from the unfolded form which disappeared in the HSQC spectra in the presence of the $(TG)_6$ DNA at molar ratio of 1:1 (red). (*C*) The structure of the TDP-43 N-domain in ribbon with residues having HSQC-peak intensity ratios < average value (0.65) colored in green. (*D*) The electrostatic potential surface of the TDP-43 N-domain structure with the identical orientation as in *C* and (*E*) with 180° rotation.

implying the involvement of the TDP-43 N terminus in forming reversible ribonucleoprotein granules (34).

Detailed analysis uncovered that the significantly broadened HSQC peaks were mostly from the residues of the ubiquitin-like fold. This observation suggests that the $(TG)_6$ ssDNA binds the ubiquitin-like fold but not the disordered region (81–102) containing the nuclear localization signal (Fig. 1*A*). As such, we calculated the intensity ratio of HSQC peaks (Fig. S7*A*) in the presence of $(TG)_6$ ssDNA at a molar ratio of 1:1.5 (protein:ssDNA), and the residues with the intensity ratio < the average value (0.65) were colored in green in the structure (Fig. 3*C*). Interestingly, the majority of the residues perturbed by binding to ssDNA is located over the electrostatically positive surface of the N-domain ubiquitin-like fold (Fig. 3 *D* and *E*), similar to the situations observed on other nucleic acid binding proteins including two TDP-43 RRMs (8).

The $(TG)_6$ ssDNA most likely does not bind to the unfolded form as the addition of the $(TG)_6$ ssDNA triggered no significant changes of the HSQC spectrum of the N-domain (10–102) (Fig. S7*B*). Furthermore, the binding is also specific for the ssDNA sequence as we failed to detect significant changes of the HSQC spectrum of the N-domain (1–102) on addition of the single-stranded $(TT)_6$ DNA (Fig. S7*C*), which was previously selected as a control oligonucleotide for binding to two TDP-43 RRMs (33). On the other hand, we also titrated with UG-rich RNA, which was previously used for determining the complex structure of two TDP-43 RRMs (8). Unfortunately, even at a molar ratio of 1:0.5 (protein:RNA), most HSQC peaks of the TDP-43 N-domain (1–102) disappeared, and visible aggregates formed. This result implies that RNA has a much higher capacity than

ssDNA in inducing the specific oligomerization of the TDP-43 N terminus, which might be correlated to the higher binding affinity of the RNA/DNA to the TDP-43 N terminus (16).

Discussion

The extensive involvements in a variety of human neurodegenerative diseases, in particular in ALS and FTLD-TDP, rank TDP-43 to be the important target not only for understanding the underlying disease mechanism but also for developing therapeutic strategies and agents to treat these diseases. However, previous research attention has mostly been directed to the recognized domains including two RRMs and prion-like domains. Only a very recent study defined the TDP-43 extreme N terminus to be a central player in both physiology and pathogenesis (15). Unfortunately despite intense efforts, the structure of the TDP-43 N terminus remains unknown due to its severe aggregation (16). In the present study, as facilitated by our previous discovery, we successfully deciphered that the TDP-43 N terminus encodes both folded and unfolded states in concentration-dependent equilibrium. This result not only explains the unusual CD spectrum of the N terminus, but also rationalizes its high tendency to aggregate. Briefly, the unfolded state with an exposure of hydrophobic patches is expected to be highly prone to aggregation in high-salted buffers following the mechanism we previously revealed for other insoluble proteins (17, 18).

Most strikingly, despite the previous failure in detecting that the TDP-43 N terminus is a homolog of ubiquitin, here for the first time, to our knowledge, we successfully decrypted that it adopts a novel ubiquitin-like fold, as aided by the powerful CS-Rosetta program with a set of unambiguous NMR constraints. Further alignment shows that the TDP-43 N terminus and ubiquitin do have 27.6% sequence homology (Fig. 3*E*). Moreover, we found that the TDP-43 ubiquitin-like fold was capable of binding to ssDNA, and amazingly, the binding shifted the conformational equilibrium toward reducing the unfolded population. Although proteins with the ubiquitin-like fold have been extensively identified to be involved in DNA/RNA associated activities (35), to the best of our knowledge, the TDP-43 N-domain represents the first ubiquitin-like fold that can directly bind to nucleic acid. This result raises the possibility that the TDP-43 N-domain and two RRMs may cooperatively interact with nucleic acids to implement the physiological functions. Furthermore, the binding to ssDNA might serve as a switch to change the functional roles of TDP-43 in cells.

The formation of the pathological and irreversible TDP-43 inclusions is a histological hallmark found in ~97% ALS and ~45% FTD patients (3). Mysteriously, the TDP-43 N terminus appears to be not only indispensable for its splicing activity, but also for its ability in forming pathological inclusions associated with impaired neurite outgrowth (15). Now our results provide a molecular mechanism for this unique dichotomy (Fig. 4). Briefly, as in cells under normal conditions, the TDP-43 concentration may be very low, the major population of the TDP-43 N-domain is the folded form, and therefore no significant TDP-43 inclusion is formed. Furthermore, if TDP-43 is properly delivered into nucleus, the binding of the ubiquitin-like fold with nucleic acids should significantly reduce the unfolded population, thus facilitating the formation of the functional and reversible oligomers but preventing the formation of the irreversible aggregates in nucleus. On the other hand, if under pathological conditions including overexpression and/or reduced degradation, the TDP-43 protein gets accumulated in cytoplasm, the conformational equilibrium will shift toward increasing the population of the unfolded state. Consequently, in the cytoplasm with high salt concentrations, the unfolded state of N-domain, together with the intrinsically disordered prion-like domain, may initiate the irreversible aggregation of the full-length TDP-43 (17, 18). This event may further destabilize two RRMs to become disordered. Consequently the whole TDP-43 molecule will be involved in the ultimate formation of a large amount of irreversible aggregates/inclusions, characteristic of the disease pathogenesis.

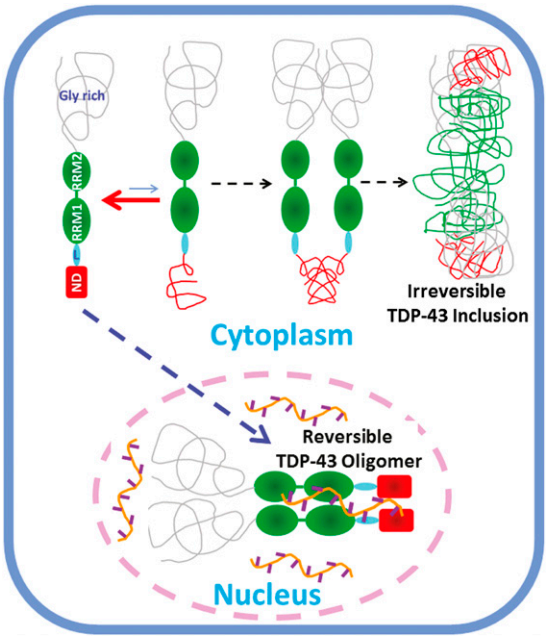

Fig. 4. Proposed model for the functional dichotomy of TDP-43 mediated by N-domain.

On the other hand, it has been recently found that TDP-43 also participates in forming RNP granules in the cytoplasm, which include processing bodies (P-bodies) and stress granules (SGs) (3, 34). The TDP-43 molecules in these granules are shuttling from the nucleus, which are bound to RNA/DNA (3, 34). Moreover, TDP-43 isolated from the cytoplasmic fraction of human brain was also shown to bind numerous mRNAs, suggesting a role of TDP-43 in mRNA stability and/or transport (36). Therefore, fundamentally different from the pathological inclusions, TDP-43 in these physiological and reversible structures appears to always be complexed with RNA/DNA (3, 34, 36). Although the relationship between these structures and inclusions currently remains unclear, it has been proposed that the physiological and reversible structures may serve as precursors, which may transform into TDP-43 pathological and irreversible inclusions with unclear mechanisms (3, 34, 36, 37). Our present results may shed light on this process, and the coexistence of the folded and unfolded states observed on the TDP-43 N terminus may be one basis for its ability to form both reversible high-order structures and irreversible inclusions. Briefly, the physiological structures formed with TDP-43 complexed with RNA/DNA might be largely reversible, as the majority of the TDP-43 N terminus population is well folded. However, on loss of RNA/DNA, the unfolded state of the TDP-43 N terminus will become populated, and consequently, the reversible structures might transform into the pathological and irreversible inclusions. Nevertheless, other TDP-43 domains may also play key roles in mediating the processes. For example, the C-terminal prion-like domain of TDP-43 is essential for interaction with other RNA-binding proteins such as hnRNP A1 and hnRNP A2/B1 to form granules. Interestingly, the ALS-causing mutations on the prion-like domain have been identified to disrupt the binding ability of TDP-43 to these proteins (3, 34, 38). As a consequence, the ALS-causing mutations on the TDP-43 prion-like domain might enhance the pathological aggregation by freeing the aggregation-prone prion-like or/and

N-domains from forming the reversible structures together with other RNA-binding proteins.

In summary, our study reveals that, despite low sequence homology, the TDP-43 N terminus encodes a novel ubiquitin-like fold in equilibrium with its unfolded form under native conditions. This unique property may be the basis for the functional dichotomy of the TDP-43 N terminus, as well as for the mysterious ability of TDP-43 in forming both reversible high-order structures and irreversible inclusions relevant to physiological and pathological processes, respectively. Our results thus imply that the loss or/and inability for TDP-43 in binding RNA/DNA might be a key factor in triggering pathogenesis of ALS/FTD and other neurodegenerative diseases. Therefore, one therapeutic strategy to treat TDP-43–causing diseases may be to use RNA/DNA or design molecules to stabilize the TDP-43 ubiquitin-like fold.

Methods

SI Materials and Methods provide detailed methods about (*i*) cloning, expression, and purification of different N-domain constructs; (*ii*) CD and NMR experiments for characterizing conformations of N-domains and their bindings to $(TG)_6$ ssDNA and $(TT)_6$ ssDNA; and (*iii*) analysis of CD and NMR data, as well as structure determination of the TDP-43 ubiquitin-like fold by the CS-Rosetta program.

ACKNOWLEDGMENTS. We thank Ms. Linlin Miao for cloning the N-domain constructs, as well as Dr. Jingsong Fang for assistance in acquiring NMR spectra. This study is supported by Ministry of Education of Singapore Tier 2 Grant 2011-T2-1-096 (R154-000-525-112) (to J.S.).

1. Arai T, et al. (2006) TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Biochem Biophys Res Commun* 351(3):602–611.
2. Neumann M, et al. (2006) Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science* 314(5796):130–133.
3. Ling SC, Polymenidou M, Cleveland DW (2013) Converging mechanisms in ALS and FTD: Disrupted RNA and protein homeostasis. *N*euron 79(3):416–438.
4. Lee EB, Lee VM, Trojanowski JQ (2012) Gains or losses: Molecular mechanisms of TDP43-mediated neurodegeneration. *Nat Rev Neurosci* 13(1):38–50.
5. Winton MJ, et al. (2008) Disturbance of nuclear and cytoplasmic TAR DNA-binding protein (TDP-43) induces disease-like redistribution, sequestration, and aggregate formation. *J Biol Chem* 283(19):13302–13309.
6. Ayala YM, et al. (2005) Human, Drosophila, and C.elegans TDP43: Nucleic acid binding properties and splicing regulatory function. *J Mol Biol* 348(3):575–588.
7. Buratti E, Baralle FE (2001) Characterization and functional implications of the RNA binding properties of nuclear factor TDP-43, a novel splicing regulator of CFTR exon 9. *J Biol Chem* 276(39):36337–36343.
8. Lukavsky PJ, et al. (2013) Molecular basis of UG-rich RNA recognition by the human splicing factor TDP-43. *Nat Struct Mol Biol* 20(12):1443–1449.
9. Buratti E, et al. (2005) TDP-43 binds heterogeneous nuclear ribonucleoprotein A/B through its C-terminal tail: An important region for the inhibition of cystic fibrosis transmembrane conductance regulator exon 9 splicing. *J Biol Chem* 280(45): 37572–37584.
10. Johnson BS, et al. (2009) TDP-43 is intrinsically aggregation-prone, and amyotrophic lateral sclerosis-linked mutations accelerate aggregation and increase toxicity. *J Biol Chem* 284(30):20329–20339.
11. Fuentealba RA, et al. (2010) Interaction with polyglutamine aggregates reveals a Q/N-rich domain in TDP-43. *J Biol Chem* 285(34):26304–26314.
12. Budini M, et al. (2012) Cellular model of TAR DNA-binding protein 43 (TDP-43) aggregation based on its C-terminal Gln/Asn-rich region. *J Biol Chem* 287(10):7512–7525.
13. Zhang YJ, et al. (2009) Aberrant cleavage of TDP-43 enhances aggregation and cellular toxicity. *Proc Natl Acad Sci USA* 106(18):7607–7612.
14. Hasegawa M, et al. (2011) Molecular dissection of TDP-43 proteinopathies. *J Mol Neurosci* 45(3):480–485.
15. Zhang YJ, et al. (2013) The dual functions of the extreme N-terminus of TDP-43 in regulating its biological activity and inclusion formation. *Hum Mol Genet* 22(15): 3112–3122.
16. Chang CK, et al. (2012) The N-terminus of TDP-43 promotes its oligomerization and enhances DNA binding affinity. *Biochem Biophys Res Commun* 425(2):219–224.
17. Song J (2009) Insight into "insoluble proteins" with pure water. *FEBS Lett* 583(6): 953–959.
18. Song J (2013) Why do proteins aggregate? "Intrinsically insoluble proteins" and "dark mediators" revealed by studies on "insoluble proteins" solubilized in pure water. *F1000 Res* 2:94.
19. Zhang O, Kay LE, Olivier JP, Forman-Kay JD (1994) Backbone ^{1}H and ^{15}N resonance assignments of the N-terminal SH3 domain of drk in folded and unfolded states using enhanced-sensitivity pulsed field gradient NMR techniques. *J Biomol NMR* 4(6): 845–858.
20. Dyson HJ, Wright PE (2004) Unfolded proteins and protein folding studied by NMR. *Chem Rev* 104(8):3607–3622.
21. 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(11):2224–2231.
22. 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(12):2795–2804.
23. Altieri AS, Hinton DP, Byrd RA (1995) Association of biomolecular systems via pulsed field gradient NMR self-diffusion measurements. *J Am Chem Soc* 117(28):7566–7567.
24. Palmer AG III, Kroenke CD, Loria JP (2001) Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. *Methods Enzymol* 339:204–238.
25. Shen Y, et al. (2008) Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA* 105(12):4685–4690.
26. Lange OF, et al. (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(27):10873–10878.
27. Raman S, et al. (2010) NMR structure determination for larger proteins using backbone-only data. *Science* 327(5968):1014–1018.
28. Warner LR, et al. (2011) Structure of the BamC two-domain protein obtained by Rosetta with a limited NMR data set. *J Mol Biol* 411(1):83–95.
29. Holm L, Rosenström P (2010) Dali server: Conservation mapping in 3D. *Nucleic Acids Res* 38(Web Server issue):W545–9.
30. Falini G, Fermani S, Tosi G, Arnesano F, Natile G (2008) Structural probing of Zn(II), Cd(II) and Hg(II) binding to human ubiquitin. *Chem Commun (Camb)* 7(45):5960–5962.
31. Sickmeier M, et al. (2007) DisProt: The database of disordered proteins. *Nucleic Acids Res* 35(Database issue):D786–D793.
32. Dosztányi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21(16):3433–3434.
33. Mackness BC, Tran MT, McClain SP, Matthews CR, Zitzewitz JA (2014) Folding of the RNA recognition motif (RRM) domains of the amyotrophic lateral sclerosis (ALS)-linked protein TDP-43 reveals an intermediate state. *J Biol Chem* 289(12):8264–8276.
34. Li YR, King OD, Shorter J, Gitler AD (2013) Stress granules as crucibles of ALS pathogenesis. *J Cell Biol* 201(3):361–372.
35. Ulrich HD (2014) Two-way communications between ubiquitin-like modifiers and DNA. *Nat Struct Mol Biol* 21(4):317–324.
36. Tollervey JR, et al. (2011) Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat Neurosci* 14(4):452–458.
37. Parker SJ, et al. (2012) Endogenous TDP-43 localized to stress granules can subsequently form protein aggregates. *Neurochem Int* 60(4):415–424.
38. King OD, Gitler AD, Shorter J (2012) The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease. *Brain Res* 1462:61–80.