Sequence-Dependent Backbone Dynamics of Intrinsically Disordered Proteins

Souvik Dey^{1,3}, Matthew MacAinsh^{1,3}, and Huan-Xiang Zhou^{1,2,*}

¹Department of Chemistry and ²Department of Physics, University of Illinois at Chicago,
Chicago, IL 60607, USA

³Equal contribution

*Correspondence should be addressed to H.X.Z. (e-mail: hzhou43@uic.edu).

Abstract

For intrinsically disordered proteins (IDPs), a pressing question is how sequence codes for function. Dynamics serves as a crucial link, reminiscent of the role of structure in sequence-function relations of structured proteins. To define general rules governing sequence-dependent backbone dynamics, we carried out long molecular dynamics simulations of eight IDPs. Blocks of residues exhibiting large amplitudes in slow dynamics are rigidified by local inter-residue interactions or secondary structures. A long region or an entire IDP can be slowed down by long-range contacts or secondary-structure packing. On the other hand, glycines promote fast dynamics and either demarcate rigid blocks or facilitate multiple modes of local and long-range inter-residue interactions. The sequence-dependent backbone dynamics endows IDPs with versatile response to binding partners, with some blocks recalcitrant while others readily adapting to intermolecular interactions.

Introduction

The sequence-structure-function paradigm has guided protein biophysics for many decades. Intrinsically disordered proteins (IDPs) account for 30% to 50% of proteomes and perform myriad cellular functions including signaling and regulation ^{1,2}. Many IDPs, including amyloid- β peptides (e.g., A β 40), tau, and α -synuclein, are also central players in human diseases such as Alzheimer's and Parkinson's ^{3,4}. The lack of well-defined structures poses the fundamental question of how IDP sequences code for functions. Many studies have shown that, for IDPs, dynamics serves as a crucial link between sequence and function, replacing the role ascribed to structure for structured proteins. For example, fast dynamics allows IDPs to rapidly adapt to target proteins in forming stereospecific complexes ⁵. Likewise, uncoupled dynamics between neighboring blocks of an IDP enables fast dissociation from its target protein, leading to a desired short lifetime for the signaling complex ⁶. The level of dynamic coupling between neighboring blocks may also dictate the competition between IDPs in binding to the same target protein ^{7,8}.

Sequence-dependent backbone dynamics on the ps-ns timescales of many IDPs has been characterized by NMR relaxation ⁹⁻¹⁷. Whereas relaxation properties of structured proteins can be easily interpreted by well-separated ns global motions and ps internal motions ¹⁸, dynamics of IDPs occur on many scales that are intricately linked and therefore analyzing the resulting NMR relaxation properties is not trivial. One approach is to fit the time correlation function of each backbone N-H bond vector to a sum of exponentials ¹¹⁻¹³, but then one is still at a loss regarding the molecular origins of any sequence dependence of the backbone dynamics. Another approach is to incorporate results from molecular dynamics (MD) simulations ^{14,16}. Force fields for MD simulations have become accurate for predicting conformational properties ¹⁹⁻²¹ but, up to the recent past, have not been reliable for predicting dynamic properties ^{14,16}. Therefore MD results for relaxation rates or time constants had to be rescaled in order to match with NMR

relaxation data, which unfortunately introduces uncertainty to MD-derived interpretations of backbone dynamics. For α -synuclein, both residues with small amplitudes (the NAC region, i.e., residues 62-95) and large amplitudes (e.g., Tyr39) on the slow timescale play important roles in fibrillation, but the connection between backbone dynamics and fibrillation is unclear¹⁴. In A β 40, residues Leu16-Phe20 have large amplitudes on the slow time scale, and noting that these residues form direct contact with the protofibril surface²², it was speculated that the higher rigidity of these residues reduces the entropic cost for monomer-protofibril binding¹⁶. Similarly, a block of residues around Lys311 of tau nucleates its aggregation²³; in the K18 fragment (residues 244-372) these residues exhibit restricted fast motions¹⁰.

Recent advances in the implementation of force fields and water models in MD simulations have led to accurate predictions of not only conformational but also dynamic properties of IDPs^{24,25}. We tested multiple force fields in simulations of the disordered N-terminal region (NT) of ChiZ²⁵, a member of the cell division machinery of *Mycobacterium tuberculosis*. The AMBER ff14SB force field²⁶ in conjunction with the TIP4P-D water model²⁷ predicted well not only the small-angle X-ray scattering (SAXS) profile and chemical shifts but also NMR relaxation rates. Two blocks of residues in the N-terminal half (N-half) of the 64-residue ChiZ-NT exhibit large amplitudes on the slow timescale, due to salt bridges, hydrogen bonds, cation-pi interactions, as well as polyproline II (PPII) formation. It was proposed that the N-half, rigidified by intramolecular interactions, would be recalcitrant to interactions with binding partners, whereas the C-terminal half (C-half) would readily adapt to binding partners. This model is validated in a subsequent study on the association of ChiZ with acidic membranes, showing that the C-half of NT is dominant in forming membrane contacts²⁸. Our study on ChiZ backbone dynamics further hinted that the thermostat needed for regulating temperature in MD simulations can distort dynamics, and correcting for such distortions

can further improve agreement with NMR relaxation data ²⁵. We have now developed a method for removing thermostat distortions of protein dynamics ²⁹.

To define general rules governing sequence-dependent backbone dynamics, here we performed long MD simulations of eight IDPs with varying sequence lengths and extents of secondary structure formations (Table 1 and Fig. 1). The MD results were validated by a variety of experimental data, including SAXS profiles, chemical shifts, paramagnetic relaxation enhancements (PREs), and NMR relaxation rates. The time correlation function of each backbone NH bond vector was fit to a sum of three exponentials, with time constants of 3-8 ns (τ_1), 1-2 ns (τ_2), and 0.1-0.2 ns (τ_3). Blocks exhibiting large amplitudes in slow dynamics (τ_1) are rigidified by side chain-side chain interactions or secondary structures, whereas glycines promote fast dynamics (τ_3) and either demarcate rigid blocks or facilitate multiple modes of side chain-side chain interactions. The sequence-dependent backbone dynamics endows IDPs with versatile response to binding partners, with some blocks recalcitrant while others readily adapting to intermolecular interactions.

Results

The eight IDPs studied here range in sequence length from A β 40 with 40 residues to α -synuclein with 140 residues (Table 1). Five of the proteins, A β 40, tau K18, the low-complexity domain of heterogeneous nuclear ribonucleoprotein A1 (A1-LCD), α - and β -synuclein, are fully disordered ^{10, 17, 30, 31} (Fig. 1A,C), but two others, HOX transcription factors DFD and SCR, have three stable α -helices ¹⁵ (Fig. 1B), and one, the C-terminal domain of the nucleoprotein of Sendai virus (SEV-NT), has a well-populated α -helix ¹¹. Following our previous study on ChiZ-NT ²⁵, we used the AMBER ff14SB force field ²⁶ and the TIP4P-D water model ²⁷ to run four replicate simulations for each of the eight IDPs. Each replicate simulation was 3.2 μ s, resulting in a total simulation time of 102.4 μ s. The total numbers of atoms in the simulation systems ranged from 240,000 to

370,000 atoms (Table S1). Below we present validation of the MD results by a variety of experimental data (Table S1) and identify the origins of the sequence-dependent backbone dynamics.

Experimental Validation of Conformational Ensembles From MD Simulations.

SAXS profiles report on the overall size of a protein and have been determined for three of the IDPs, tau K18³², A1-LCD¹⁷, and α -synuclein³³. In all the three cases, the SAXS profiles calculated on conformations sampled by our MD simulations show good agreement with the experimental counterparts (Fig. S1). To provide a direct measure of the sizes of the IDPs, we calculated the distributions of the radius of gyration (R_g ; Fig. S2). For three of the IDPs, A β 40, tau K18, and α -synuclein, the mean R_g values agree with those predicted from a scaling relation,

$$R_g = 2.54N^{0.522} \quad [1]$$

deduced from a set of IDPs (N : number of residues)³⁴. The mean R_g values of HOX-DFD, HOX-SCR, and SEV-NT are lower than predicted by Eq [1], due to the presence of α -helices. The mean R_g value of A1-LCD is also lower than predicted, due to extensive local and long-range interactions (see below).

Secondary chemical shifts indicate the formation of α -helices and β -strands, and have been measured for all the eight IDPs^{9-11, 15, 17, 30, 35}. In Figs. 2A-C and S3A-E, we compare the experimental data with the secondary chemical shifts calculated on conformations sampled by our MD simulations. Overall, there is reasonable agreement. Most of the secondary chemical shifts are close to 0, indicating complete disorder. The exceptions are three blocks in both of the HOX transcription factors and one block in SEV-NT, with large positive secondary chemical shifts indicating high propensities for α -helices. In later figures, we also report the helix and β -strand propensities of individual residues, which confirm the lack of well-populated secondary structures other than the α -helices in HOX-DFD, HOX-SCR, and SEV-NT.

Inter-residue nuclear Overhauser effect (NOE) cross-peaks report on pairs of residues that can form close contact between hydrogen atoms; for A β 40, NOEs between H $_{\alpha}$ at residue i and H $_N$ at residue $i + 2$ “were seen within the Asp7-Glu11, Phe20-Ser26, and Gly29-Ile31 regions”³⁰ (Fig. S3F). We calculated the effective interproton distances, $\langle r^{-6} \rangle^{-1/6}$ ³⁶, which implicated NOEs for His6, Asp7, His13, Asp23, and Asn27 (Fig. S3F). The calculated results, except for the addition of His13, are in agreement with the observed $\alpha N(i, i + 2)$ NOEs.

PREs can reveal long-range interactions, and have been measured for spin labels placed at 12 residues distributed throughout the sequence of α -synuclein^{31, 37-39}. Consistent with the experimental data, a site in the N-terminal region (residues 1-61) can interact with the C-terminal region (residues 96-140) and vice versa; likewise a site in the central NAC region can interact with the C-terminal region and vice versa (Fig. 2D). For example, our MD results correctly predict that a spin label at residue 120³¹ has propensities for interacting with residues around position 40 and residues around position 90. The calculated PREs for β -synuclein do not agree as well with the experimental values³¹ (Fig. S4A-C). The MD results correctly predict the absence of contacts between a spin label at position 20 with the central region, but incorrectly predict interaction propensities between a spin label at residue 114 with residues around position 10 and residues around position 75. For tau K18, consistent with experimental data⁴⁰, PREs calculated from the MD conformations indicate absence of long-range interactions (Fig. S4D). Next we document the validation of our MD simulations by NMR relaxation data.

Correcting for Dynamic Distortion by the Langevin Thermostat. The backbone amide transverse (R_2) and longitudinal (R_1) relaxation rates and nuclear Overhauser effects (NOE) are determined by the time correlation function of the NH bond vector (see Computational Methods). We fit this correlation function to a sum of three exponentials, with time constants τ_1 , τ_2 , and τ_3 (ordered from long to short) and amplitudes A_1 , A_2 , and

A_3 ²⁵. Illustrative fits are shown in Fig. S5A-C. Note that we did not restrain the sum of the amplitudes to be 1; the missing amplitude represents ultrafast motion (faster than the 20-ps interval for saving snapshots from the MD simulations). The exponential components with the long (τ_1), intermediate (τ_2), and short (τ_3) time constants contribute the most to R_2 , R_1 , and NOE, respectively.

We have recently shown systematic time dilation by the Langevin thermostat⁴¹ used here and elsewhere for regulating temperature in MD simulations, and found a correction scheme²⁹. A raw time constant τ_{raw} obtained in an MD simulation is corrected by a scaling factor,

$$\tau = \frac{\tau_{\text{raw}}}{a + b\tau_{\text{raw}}} \quad [2]$$

The constants a and b depend on the damping constant of the Langevin thermostat and are 1.526 and 0.086, respectively, for the damping constant of 3 ps⁻¹ in the present study. The MD simulations in the present study were also run at constant pressure using the Berendsen barostat⁴²; however, we did not find any additional effect of the barostat on protein dynamics. Note that the scaling factor is proportional to τ_{raw} , and thus has the strongest effect on τ_1 and weakest effect on τ_3 . Figure S5D illustrates the correction for the three time constants of Gln15 in A β 40 and of Ile397 in HOX-DFD. Figure S6 demonstrates that the correction for the thermostat effects dramatically reduces the root-mean-square-error (RMSE), from 3.1 s⁻¹ to 0.5 s⁻¹, of the R_2 values of α -synuclein, while maintaining the sequence-dependent profile. In fact, the R_2 , R_1 , and NOE values calculated from the MD simulations match well with the experimental data for all the eight IDPs (Figs. 3 and S7A to S11A).

Backbone Dynamics of A β 40. Figure 3A displays the comparison between the calculated and experimental¹⁶ R_2 , R_1 , and NOE values for A β 40 residues. We make the following observations on the sequence dependences of the calculated results. First, the profile of each relaxation property along the sequence has a bell shape, with the terminal

residues having much lower values than internal residues. This observation holds for all the IDPs, and the corresponding elevated fast dynamics of the terminal residues on the fast timescale can be attributed to the fact that these residues are linked to the polypeptide chain only on one side. Second, the N-half of the sequence has higher R_2 and NOE values than the C-half; the mean R_2 values are $\sim 4 \text{ s}^{-1}$ and 3 s^{-1} for the N- and C-halves, respectively. This asymmetry, corresponding to more significant slow dynamics in the N-half, is similar to what was observed for ChiZ-NT²⁵. Third, the highest R_2 values are found for residues His13-His14-Gln15. The experimental R_2 values are even higher, likely due to exchange between unprotonated and protonated histidines¹⁶ that is not modeled by our MD simulations. Fourth, a local R_2 peak is observed in His6-Gln7 of the N-half, while the first block of eight residues in the C-half (Ala21-Lys28) has higher R_2 values than the next block of eight residues.

To uncover the origins of the sequence-dependent backbone dynamics (of the non-terminal residues), we calculated the secondary structure propensities and contact numbers of individual residues (displayed above the relaxation properties in Fig. 3A). The latter was defined as the number of residues, other than the immediate neighbors, that form contacts with a given residue; contacts were defined as between heavy atoms within a cutoff distance of 3.5 Å. In line with the secondary chemical shifts (Fig. 2A), helices and β -strands are only sampled infrequently ($< \sim 10\%$ in all residues). PPIIs are also only sampled at low levels. The lowly populated secondary structures in A β 40 do not explain the sequence-dependent backbone dynamics. Instead, the contact numbers provide the explanations. It is immediately clear that the N-half residues have higher contact numbers than the C-half residues. Moreover, His6, His13, and Gln15 that are in the local or global R_2 peaks are among the residues with the highest contact numbers.

To further characterize residue-residue contacts, we calculated the contact map, which displays the frequencies of contact formation between any pair of side chain heavy atoms²⁵. The contact frequencies above a threshold of 0.005 are displayed in Fig. 4A.

Two blocks of residues, Glu3- Glu11 and Glu11-Lys16, show tendencies to form extensive contacts, leading to rigidification. That these two blocks overlap and are both in the N-half underlies the asymmetry in backbone dynamics between the N- and C-halves. In the Glu3-Glu11 block, the most frequent contacts include salt bridges of Arg5 with Glu3, Asp7, and Glu11 (Fig. 4A inset); hydrogen bonds of Ser8 with His6 (Fig. 4A inset) and Asp7; and cation- π interaction between Arg5 and Tyr10. In the Glu11-Lys16 block, the most frequent contacts include a salt bridge of Glu11 with Lys16; π - π stacking between His13 and His14 (Fig. 4A inset); a hydrogen bond between His13 and Lys16; and amino- π interactions of Gln15 with His13 and His14. These side chain-side chain interactions explain the local R_2 peak in residues 6-7 and the global R_2 peak in residues 13-15. In the C-half, the Glu22-Lys28 block has a tendency to form a less extensive network of contacts, including a salt bridge between Glu22 and Lys28 (Fig. 4A inset) and hydrogen bonds of Ser26 with Glu22, Asp23, Asn27, and Lys28. These interactions explain why the first eight residues of the C-half have higher R_2 values than the rest of the C-half. Note that the two serine residues, Ser8 and Ser26, can both hydrogen bond with multiple neighboring residues. These hydrogen bonds may be the reason for the $\alpha N(i, i + 2)$ NOEs described above.

All the foregoing interactions involve polar (including charged) residues. Val18 and Phe20 are the only two nonpolar residues that form contacts with elevated frequencies. The N-and C-halves contain 12 and 5 polar residues, respectively, and thus the former sequence is much more polar than the latter sequence. Therefore the origin of the asymmetry in backbone dynamics can be traced to the disparity in number of polar residues between the two halves.

Backbone Dynamics of HOX-DFD. In Fig. 3B we compare the calculated and experimental 15 NMR relaxation data for HOX-DFD. In line with the experimental data, the most noticeable feature of the calculated relaxation data along the sequence is the

significantly higher R_2 and NOE of the helical C-half (starting at residue Arg375) relative to the disordered N-half. The mean R_2 values are $\sim 4 \text{ s}^{-1}$ and 9 s^{-1} for the disordered and helical regions, respectively. In contrast, R_1 is relatively uniform, except for a dip around residue Gly354. In the helical region, R_2 shows dips in the two inter-helical linkers. In the disordered region, R_2 , R_1 , and NOE all show a local peak in residues Trp346-Met347, which is a part of the conserved YPWM motif. After the dip around Gly354, R_2 rises steadily from Tyr360 to Arg368, and then rapidly until reaching the helical region.

The high R_2 and NOE of the C-half can be easily explained by the three α -helices, which remain intact in the MD simulations (Fig. 3B, top row). On the other hand, the N-half has low secondary structure content and forms few contacts. The contact numbers do show higher values in the YPWM motif and in the 14 residues (starting at Tyr360) preceding the helical region, and thus track well the sequence dependence of R_2 in the N-half. The contact map, with a frequency threshold of 0.02, in Fig. 4B provides more details. In the C-half, 112 side-chain heavy atoms, or 41%, form contacts with other side-chain heavy atoms with frequencies above the threshold. The counterparts in the N-half are only 33 and 22%. The side chain-side chain contacts in the C-half show a well-packed 3-helix bundle, including salt bridges between the helices (Glu384-Arg395, Glu382-Arg417, and Arg396-Glu407) and cation- π interactions between the inter-helical linkers and the helices (Arg389-Trp421 and Tyr390-Arg418). In the N-half, most of the atoms with elevated contact frequencies belong to two blocks: the YPWM motif and Tyr360 to Arg368. The contacts in the YPWM motif are anchored by the Tyr344-Trp346 π - π stacking (Fig. 4B inset). In addition, this motif can occasionally form long-range contacts with the second α -helix in the C-terminal region, including a hydrogen bond between Tyr344 and Arg393 and a cation- π interaction between Trp346 and Arg394 (Fig. 4B inset). These interactions explain the local R_2 peak in residues Trp346-Met347. The interactions within the Tyr360 to Arg368 block (e.g., a Glu365-Arg368 salt bridge) and

between Tyr373 with the third α -helix and between Thr374 and the first α -helix explain the rise in R_2 to the level of the C-terminal helical region.

Backbone Dynamics of α -Synuclein. The calculated and experimental ¹⁴ R_2 , R_1 , and NOE values for α -synuclein are shown in Fig. 3C. In both sets of data, R_1 and NOE are rather uniform along the sequence, but R_2 , with a mean value of $\sim 3 \text{ s}^{-1}$, shows local peaks at Tyr39, Lys96, and Asp121. α -Synuclein contains six imperfect repeats with a consensus sequence KTKEGV (Table 1). In five of these repeats, the fifth position is a glycine. R_1 dips at the sixth position in each of these five repeats (vertical lines in Fig. 3C). Two adjacent Gly residues occupy positions 67 and 68. Each of the three relaxation properties dips prominently at Gly68.

The secondary structure propensities and contact numbers provide some explanations for the R_2 local peaks. Both Leu38 and Val95 have relatively high propensities ($\sim 20\%$) for β -strands, while residues near Asp121 have PPII propensities close to 50%. Four proline residues, at positions 108, 117, 120, and 128, surround Asp121. Moreover, Tyr39 and those near Lys96 have the highest contact numbers. The contact map, with a frequency threshold of 0.005, provides more details (Fig. S12). Tyr39 is upstream of the fourth KTKE motif and can hydrogen bond with Ser42 (Fig. S12 inset); the latter can also form additional hydrogen bonds with Thr44 and Lys45 from the fourth KTKE motif. These interactions explain the elevated R_2 of Tyr39. Lys96 starts a block of residues with elevated frequencies for contact formation, including a salt bridge between Lys97 and Asp98 and hydrogen bonding of Lys96 with Gln99 (Fig. S12 inset) and of Asn103 with Gln 99 and Glu105. Furthermore, Lys96, along with Gln99, can also form long-range contacts with Asp121 (Fig. S12 inset). These interactions explain the elevated R_2 of Lys96. In addition to the just-noted long-range contacts, the elevated R_2 of Asp121 is also contributed by its location inside a block of residues, Pro108-Pro128, that contain four Pro residues and thus have high propensities for forming PPII.

Backbone Dynamics of the Other Five IDPs. HOX-SCR is a very close homologue of HOX-DFD and their relaxation properties show very similar sequence dependences (Figs. 3B and S7A). Likewise β - and α -synuclein have significant sequence similarity, and their relaxation properties are all pretty uniform along the sequence, except for minor peaks in R_2 (Figs. 3C and S8A). Compared to the HOX proteins and the synucleins, SEV-NT is intermediate both in the level of secondary structure formation and in the variation of relaxation properties over the sequence (Fig. S9A). Whereas the three α -helices of the HOX proteins are intact, the α -helix (Val476 to Gln492) in SEV-NT is only transient: the helical propensities peak at 90% over three residues but taper off on both sides. R_2 exhibits a significant elevation in the middle of this transient helix, though the MD results do not rise as much as the experimental values¹¹. NOE also shows a modest rise whereas R_1 shows a depression over the helical region. NOE and R_1 are otherwise very uniform, but R_2 shows a local peak at Ala450 and dips at Gly415 and Gly456. That glycines promote fast dynamics is now a familiar occurrence. The elevated R_2 of Ala450 can be explained by its proximity to a block of residues, Gln436-Lys448, that tends to form extensive interactions, including Arg439-Asp437, Arg439-Glu444, and Arg448-Glu444 salt bridges, Gln436-Arg439 and Trp441-Glu444 hydrogen bonds, and Arg448-Trp441 cation- π interaction. Arg439 can also form salt bridges with upstream Asp427, Glu429, and Glu431, as well as Asp494 that is downstream of the transient helix. Glu429 can also form a salt bridge with Arg482 that is within the helical region. These long-range interactions help keep SEV-NT compact (Fig. S2D).

Tau K18 and AI-LCD happen to have similar sequence lengths (129 and 131 residues, respectively). While AI-LCD has propensities (~20%) for β -strands in some residues, both of these IDPs lack stable secondary structures (Figs. S10A and S11A, top). However, as already noted above, whereas the overall size of tau K18 is typical of IDPs, AI-LCD is much more compact (Figs. S2E,F). Correspondingly, slow dynamics is largely

absent in tau K18 but prominent in A1-LCD. In tau K18, R_2 is below 5 s^{-1} for all but a few residues (e.g., Lys311 and Val313) (Figs. S10A). In contrast, in A1-LCD, R_2 is above 5 s^{-1} for essentially all non-terminal residues, and reaches 9 s^{-1} for Tyr61-Phe64 and Asn79-Gln80 (Figs. S11A). The contact maps explain the difference. In tau K18, contacts with frequencies above 0.01 are sparse and all local (Fig. S13A). The sequence of this IDP can be divided into four imperfect repeats ending with PGGG (Table 1 and Fig. S13A). All the three relaxation properties show prominent dips at the end of each repeat (Figs. S10A). That observation, along with the lack of any significant contact between the repeats, suggests that the four repeats behave dynamically as independent units. In the third repeat, the relatively higher R_2 values of Lys311 and Val313 can be accounted for by local contacts including a salt bridge between Lys311 and Asp314 and a hydrogen bond between Asp314 and Ser316 (Fig. S13A), and additionally by some rigidification provided by Pro312.

The sequence of A1-LCD contains three GGGG motifs, with the third Gly at positions 52, 74, and 103 (Table 1). R_2 dips at these positions; the dips are particularly prominent for the last two GGGG motifs, where dips in NOE also occur (Figs. S11A). The contact map of A1-LCD shows three blocks of residues, Phe19-Arg49, Ser57-Phe71, and Asn76-Phe84, with elevated frequencies for extensive contacts (Fig. S13B). It is clear that the GGGG motifs demarcate these blocks. Within the second block, hydrogen bonds can be formed by Ser57 with Asp59, Asp59 with Asn62, Asn62 with Asn66, Asn66 with Asn70, and Asp67 with Ser69, and an amino- π interaction can be formed between Asn70 and Phe71. This block contains five Gly residues, and it appears that the flexibility of these Gly residues facilitates neighboring residues in forming multiple modes of interactions, e.g., Asp59 hydrogen bonding with either Ser57 or Asn62. In addition, residues in this block can also form long-range interactions, including hydrogen bonding of Tyr54 with Ser16, Asn62 with Gln113 and Asn79, and Asn66 with Asn91, cation- π interactions of Tyr61 (and Tyr54) with Arg95 (Fig. S13B inset) and of Phe64

with Arg95, and an amino- π interaction of Tyr61 with Asn91. Again, intervening Gly residues, e.g., between Asn91 and Arg95, appear to facilitate the multiple modes of interactions. These extensive networks of interactions explain the elevated R_2 of Tyr61-Phe64. In the third block, hydrogen bonds can form between Asn76 and Asn78 or Gln80, between Ser81 and Asn83, and an amino- π interaction can form between Asn83 and Phe84, along with the already noted inter-block hydrogen bond between Asn62 and Asn79. These interactions explain the elevated R_2 of Asn79-Gln80.

Time Constants and Amplitudes of Three-Exponential Fits. Our NMR relaxation properties were calculated from the three-exponential fits of backbone NH time correlation functions. The time constants and amplitudes of these fits provide additional insight into the backbone dynamics of the IDPs. We display the distributions of the three time constants for each IDP in Fig. S14, showing good separation. The ranges of τ_1 , τ_2 , and τ_3 are 3-8 ns, 1-2 ns, and 0.1-0.2 ns, respectively. The means and standard deviations of each time constant and the corresponding amplitude, calculated among all the residues of each IDP, are collected in Table S2; the values for individual residues are shown in Figs. 5 and S7B,C to S11B,C. The mean τ_1 values separate the 8 IDPs into two groups. The first group, with mean τ_1 close to or even above 6 ns, consists of HOX-DFD, HOX-SCR, and SEV-NT, all of which have stable or well-populated α -helices, and A1-LCD, which form numerous side chain-side chain contacts. The other four IDPs, with mean τ_1 around 4 ns, are fully disordered and form much fewer side chain-side chain contacts than A1-LCD (β -synuclein has a mean τ_1 at 5.2 ns but that is probably overestimated). The mean τ_2 values of the first group are close to 1.6 ns and somewhat higher than those the second group, which are close to 1 ns. However, there is no distinction in mean τ_3 between the two groups of IDPs. The first group has approximately double the mean A_1 (at ~ 0.6) but half the mean A_2 (at ~ 0.2) and mean A_3 (at 0.1) of the second group. Correspondingly, A_1 is the dominant amplitude in the first group but A_1 and A_2 are

roughly equal in the second group. It is clear that the formation of secondary structures and side chain-side chain interactions promotes slow dynamics and suppresses fast dynamics.

Next we examine the sequence dependences of the time constants and amplitudes of the IDPs. Figures 5C and S7B show that, for HOX-DFD and HOX-SCR, the helical C-half has moderately higher τ_1 , slightly higher τ_2 , and similar τ_3 when compared with the disordered N-half. More interestingly, as shown in Fig. 5D and S7C, going from the N-half to the C-half, there is a significant increase in A_1 along with significant decreases in A_2 and A_3 . Correspondingly, A_1 and A_2 are approximately equal in the N-half but A_1 dominates in the C-half. The contrasts in time constants and amplitudes between the C- and N-halves of the two HOX proteins mirror precisely the contrasts between the IDPs as two separate groups, one rich in structures or interactions and the other lacking such features. The same contrasts apply to SEV-NT when the helical region is compared to the non-helical region, except that the dominance of A_1 extends to most of the non-helical region as well (Fig. S9B,C), due to the long-range contacts detailed above. Likewise, due to the extensive local and long-range contacts, A_1 dominates nearly the entire sequence of A1-LCD (Fig. S11B,C). The opposite scenario is represented by α -synuclein and tau K18, which are disordered and lack significant long-range contacts (Fig. S12 and S13A), and correspondingly motions on the intermediate timescale have the largest amplitudes (Figs. 5F and S10C). β -Synuclein is similar to α -synuclein, except that A_1 is overestimated and becomes comparable to A_2 (Fig. S8C). A β 40 represents an intermediate scenario, where the N-half with extensive contacts has higher A_1 than A_2 but the C-half has comparable A_1 and A_2 (Fig. 5B).

Lastly let us take a closer look at the time constants and amplitudes of some residues noted above for elevated or depressed relaxation rates. The elevated R_2 at Tyr39, Lys96, and Asp121 in α -synuclein (Fig. 3C) and at Lys311 and Val313 in tau K18 (Fig. S10A) can both be attributed to elevated τ_1 (Figs. 5E and S10B), but the elevated R_2 at Ala450 of

SEV-NT (Fig. S9A) and at Tyr61-Phe64 and Asn79-Gln80 in A1-LCD (Fig. S11A) is due to elevated A_1 (Figs. S9C and S11C). Many Gly residues have been seen to promote fast dynamics, as evidenced by dips in R_2 and/or R_1 , including six in α -synuclein (Fig. 3C), five in β -synuclein (Fig. S8A), two in SEV-NT (Fig. S9A), four in tau K18 (Fig. S10A), and three in A1-LCD (Fig. S11A). All these Gly residues (or their immediate neighbors) can be recognized by dips in A_{sum} (Figs. 5F and S8C to S11C). In some cases the dips in A_{sum} are specifically due to dips in A_1 (Figs. S9C to S11C) and occasionally are also accompanied by dips in τ_1 (Fig. S9B).

Discussion

Based on experimentally validated MD simulations, we have characterized the sequence-dependent backbone dynamics of a variety of IDPs. The dynamics spans a wide range of timescales, from sub-ps to 10 ns, and the distribution of amplitudes on these timescales can vary greatly from IDP to IDP. Slow dynamics can be promoted by two mechanisms. The first mimics what happens in structured proteins, through the formation and packing of secondary structures, as exemplified by the helical C-half of the two HOX proteins. The second is through extensive local and long-range contacts, including hydrogen bonds, salt bridges, and cation- and amino- π interactions. Local contacts can rigidify a block of residues, as illustrated by the N-half of A β 40. When long-range contacts are also formed, as typified by A1-LCD, the entire IDP becomes relatively compact and tumbles as a globule without any secondary structures. SEV-NT represents a hybrid of the above two mechanisms, where an α -helix is well-populated and also engages in long-range contacts.

Glycine is flexible and generally promotes fast dynamics. The flexibility allows glycine to play opposite roles in IDP dynamics. On the one hand, Gly repeats are very effective in interrupting rigid blocks, as illustrated by three GGGG motifs in A1-LCD and GGG and GG in SEV-NT. In tau K, four GGG motifs effectively break the IDP into

four dynamically independent units. On the other hand, as seen in A1-LCD, Gly residues facilitate multiple modes of local interactions within rigid blocks and multiple modes of long-range contacts between rigid blocks.

We have fit NH time correlation functions to a sum of three exponentials. The three time constants fall in the ranges of 3-8 ns (τ_1), 1-2 ns (τ_2), and 0.1-0.2 ns (τ_3) for all the IDPs. The greatest difference in dynamics between different regions of an IDP and among different IDPs is captured by the motional amplitudes of the different timescales. A_1 , the amplitude in the slow timescale, is dominant in the entire A1-LCD and SEV-NT, in the helical C-half of the two HOX proteins, and in the N-half of A β 40, reflecting the effects of secondary structures and local and long-range interactions. In contrast, A_2 , the amplitude in the intermediate timescale, is either dominant or on par with A_1 in the entire tau K18, α - and β -synuclein, in the disordered N-half of the two HOX proteins, and in the C-half of A β 40, reflecting the lack of long-range or even local interactions. The A_1 -dominated IDPs or regions thereof also have a slightly higher A_{sum} than their A_2 -dominated counterparts, indicating that the suppression of slower dynamics in the former group extends all the way to the ultrafast (i.e., ps and sub-ps) timescale. Lastly the A_1 -dominated group has somewhat longer τ_1 (and τ_2 to a lesser extent) than the A_2 -dominated group.

It is instructive to compare the time constants and amplitudes of the IDPs with those of structured globular proteins. In the latter case the NH time correlation functions could be fit to the sum of two exponentials, with one time constant (τ_G) for ns global tumbling and the other (τ_L) for fast local motion²⁹. τ_G scales linearly with the number of residues ($\tau_G \approx 0.051N$) whereas τ_L fluctuates around 25 ps. The typical amplitudes of these two types of motions are 0.85 (A_G) and 0.05 (A_L). The τ_1 motion of IDPs has some resemblance to the global tumbling of globular proteins, but differs in important ways. First, while the helical region of the two HOX proteins and the few most highly helical residues of SEV-NT have A_1 amplitudes that approach the typical value of 0.85 for A_G

(Figs. 5D, S7C, and S9C), other IDP residues, especially those in the IDPs that lack long-range contacts, have much lower A_1 amplitudes. Second, the mean τ_1 values of the IDPs lack dependence on sequence length. The expected τ_G value for a 40-residue protein like A β 40 is 2.0 ns and that for a 140-residue like α -synuclein is 7.1 ns. However, A β 40 has a mean τ_1 of 3.8 ns and α -synuclein has a mean τ_1 of 4.4 ns. Relative to the expected τ_G , the higher mean τ_1 of A β 40 can be explained by its conformational ensemble being more open than the structure of a globular protein of the same sequence length, whereas the lower mean τ_1 of α -synuclein can be rationalized if each of its residues belongs to a dynamically independent unit that is much shorter than the full-length protein. The fact that both A β 40 and α -synuclein have a mean τ_1 around 4 ns suggests that, for IDPs that lack well-populated secondary structures and long-range contacts, such dynamically independent units may be as short as 30 residues. As noted above, tau K18 may be divided into four such units demarcated by PGGG motifs, and the length of each unit is about 30 residues. So apparently two factors pull the τ_1 values of IDPs in opposite directions: being more open than globular proteins could lead to higher τ_1 , but the accompanying loss of long-range contacts could lead to lower τ_1 .

The fast local motion seen in globular proteins with a very small amplitude does not show up in our three-exponential fits for IDPs, possibly being lumped into the missing amplitude. Instead, τ_2 and τ_3 emerge. An interesting question is motions on these timescales involve how many residues. If a 4-ns τ_1 indeed involves 30 residues, then we could estimate a 1-ns τ_2 could involve perhaps 5-7 residues, and a 0.15-ns τ_3 might involve at most the nearest neighbors.

Sequence-dependent backbone dynamics can potentially code for functions or contribute to disease mechanisms, by endowing IDPs with versatile response to binding partners. In a previous study, based on sequence-dependent dynamics, we proposed that the rigid N-half of ChiZ-NT would be recalcitrant to interactions with binding partners whereas the flexible C-half would readily adapt to binding partners²⁵. This prediction

was verified by a subsequent study on the association of ChiZ with acidic membranes, showing that the C-half is dominant in forming membrane contacts ²⁸. A similar mechanism may be at work in the nucleation of A β 40 fibrils, where the flexible C-half may readily undergo a conformation transition to form a β -sheet between monomers under a concentrated condition (Fig. 6A). Likewise tau K18 may utilize its flexibility to achieve multiple modes of binding with microtubule. On the other hand, the preformed three-helix bundle allows the HOX transcription factors to recognize their DNA target, while the dynamic N-terminal tail, which occasionally forms intramolecular contacts with the helix bundle in the free state, now binds to the co-transcription factor to provide further stabilization ⁴³ (Fig. 6B). Similarly, the nascent α -helix in SEV-NT recognizes the target protein and becomes further rigidified on the target surface ⁴⁴. Note that most of the local and long-range intramolecular contacts in free SEV-NT can still form in the bound state; a notable exception is the salt bridge between Glu429 and Arg482, as Arg482 now engages in intermolecular interactions with the target protein. It is also likely that many intramolecular contacts of A1-LCD become intermolecular under a concentrated condition, thereby enabling condensate formation ¹⁷. Intermolecular interactions in turn dictate dynamic properties of biomolecular condensates ⁴⁵. Knowledge of sequence-dependent dynamics and its origins provides deep insight into functional and disease mechanisms.

Computational Methods

Molecular Dynamics Simulations. MD simulations were run in AMBER18 ^{46,47} using the ff14SB force field ²⁶ for proteins and TIP4P-D for water ²⁷. The eight IDPs span a wide range in sequence length and level of secondary structure. Accordingly we generated their initial structures in several ways. Those of A β 40 were snapshots, with a short α -helix or a short β -sheet, from a previous simulation using a different force field ⁴⁷. Initial structures of HOX-SCR and HOX-DFD were generated by the I-TASSER web

server⁴⁸; that of tau K18 was from the RaptorX web server⁴⁹. The initial structure of SEV-NT consisted of an α -helix for residues 476-492 [built in Pymol (<https://pymol.org/>)] and disordered N- and C-terminal regions (built by *tleap*). Initial structures of A1-LCD and α - and β -synuclein were generated by the TraDES web server⁵⁰. Each initial structure was placed into a rectangular box, with a solvent layer ranging from ~ 20 Å (for the more open) to ~ 50 Å (for the more compact). Na⁺ and Cl⁻ were added to neutralize the systems and provide the experimental salt concentrations. The initial extent of secondary structures, total number of atoms, NaCl concentration, and simulation temperature for each IDP are listed in Table S1.

For each system, with *sander*, energy minimization (2000 steps of steepest descent and 3000 steps of conjugate gradient) was followed by heating from 0 to the final temperature (Table S1) over 100 ps at a 1 fs timestep. Temperature was regulated by the Langevin thermostat⁴¹ with a 3.0 ps⁻¹ damping constant. Bond lengths involving hydrogens were constrained using the SHAKE algorithm⁵¹. Long-range electrostatic interactions were treated by the particle mesh Ewald method⁵². The cutoff distance for nonbonded interactions were 10 Å. The simulations then continued on GPUs using *pmemd.cuda*⁵³ in four replicates with different random seeds at constant temperature and pressure (1 atm) at a 2 fs timestep, initially for 3 ns and then for 3.2 μ s. In the case of A β 40 and A1-LCD, four replicates were prepared from the beginning using different initial structures. Pressure was regulated using the Berendsen barostat⁴². The final 2.5 μ s of each trajectory, with snapshots saved every 20 ps, was used for analysis.

Chemical shifts, Small Angle X-ray Scattering Profiles, and Paramagnetic

Relaxation Enhancements. These properties were averaged over MD snapshots saved every 1 ns. Chemical shifts were calculated using SHIFTX2⁵⁴. Secondary C α and C β chemical shifts were calculated by subtracting random-coil values from POTENCI⁵⁵. SAXS profiles were calculated using FoXS⁵⁶, with calculated profile for each snapshot

scaled to optimize the match with the experimental data. PREs were calculated using DEER-PREdict⁵⁷.

Secondary Structures. Secondary structures were calculated using the *dssp* command in *cpptraj*⁵⁸. These results were expanded to include PPII formation when three or more consecutive coil residues fell into the PPII region of the Ramachandran map.

Contact Numbers and Contact Maps. Distances between heavy atoms (except those in the same residue) were calculated by loading trajectories into MDTraj⁵⁹. A contact was defined when two heavy atoms were within 3.5 Å of each other. The contact number of a residue was the number of residues beyond the immediate neighbors with which at least one contact was formed. This number was calculated for each snapshot and then averaged over all the saved snapshots. For calculating the contact map, the contact frequency between any two heavy atoms from two different side chains was calculated by counting the fraction of snapshots in which that contact was formed.

NMR Relaxation Properties. For each non-proline residue, the time correlation function of the NH bond vector was calculated as

$$C(t) = \langle P_2[\mathbf{n}(t') \cdot \mathbf{n}(t' + t)] \rangle_{t'} \quad [3]$$

where $\mathbf{n}(t')$ is the unit vector along the NH bond at time t' ; $P_2(x)$ is the second order Legendre polynomial; and $\langle \dots \rangle_{t'}$ denotes time average over each trajectory. After further averaging over the four replicate simulations, the correlation function was fit to a tri-exponential function,

$$C(t) = \sum_{i=1}^3 A_i e^{-t/\tau_i} \quad [4]$$

by a custom python code²⁵. The fit was done without constraints on any parameters. In particular, the sum of the amplitudes was not constrained to 1, allowing for missing

amplitudes of ultrafast motions (i.e., faster than 20 ps at which snapshots were saved).

The default for the upper bound of the time range was 15 ns, but for a small number of residues fitting errors were reduced by increasing this upper bound. Following our previous study ²⁹, the three time constants were corrected by a scaling factor according to Eq [2] in order to remove dynamic distortions of the Langevin thermostat used in the MD simulations.

The spectral density for the time correlation function of Eq [4] was

$$J(\omega) = \sum_{i=1}^3 \frac{A_i \tau_i}{1 + (\tau_i \omega)^2} \quad [5]$$

Finally R_1 , R_2 , and NOE were given by

$$R_1 = f_{DD}[J(\omega_H - \omega_N) + 3J(\omega_N) + 6J(\omega_H + \omega_N)] + f_{CSA}J(\omega_N) \quad [6]$$

$$R_2 = \frac{R_1}{2} + f_{DD}[2J(0) + 3J(\omega_H)] + \frac{2}{3}f_{CSA}J(0) \quad [7]$$

$$\text{NOE} = 1 + \frac{f_{DD}}{R_1} \frac{\gamma_H}{\gamma_N} [6J(\omega_H + \omega_N) - J(\omega_H - \omega_N)] \quad [8]$$

where $f_{DD} = \frac{1}{10} \left(\frac{\mu_0 \hbar \gamma_N \gamma_H}{4\pi r_{NH}^3} \right)^2$ and $f_{CSA} = \frac{2}{15} \omega_N^2 \Delta_{CSA}^2$. The meanings of the symbols are:

μ_0 , permittivity of free space; $\hbar$, reduced Plank constant; γ_H and γ_N , gyromagnetic ratios of hydrogen and nitrogen; $\omega_H = \gamma_H B_0$, Larmor frequency of hydrogen; ω_N , counterpart of nitrogen; r_{NH} , NH bond length (set at 1.02 Å); and Δ_{CSA} ($= -170$ ppm), chemical shift anisotropy of nitrogen. The magnetic field strengths were 600 MHz for Aβ40 and α-and β-synuclein, 700 MHz for tau K18, 800 MHz for HOX-DFD, HOX-SCR, and A1-LCD, 850 MHz for SEV-NT. RMSEs for R_1 , R_2 , and NOE were calculated over non-terminal residues.

References

1. Xue, B., Dunker, A. K., Uversky, V. N. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J Biomol Struct Dyn* **30**, 137-149 (2012).

2. Uversky, V. N. Introduction to Intrinsically Disordered Proteins (IDPs). *Chemical Reviews* **114**, 6557-6560 (2014).
3. Busche, M. A., Hyman, B. T. Synergy between amyloid-beta and tau in Alzheimer's disease. *Nat Neurosci* **23**, 1183-1193 (2020).
4. Stefanis, L. alpha-Synuclein in Parkinson's disease. *Cold Spring Harb Perspect Med* **2**, a009399 (2012).
5. Wu, D., Zhou, H. X. Designed Mutations Alter the Binding Pathways of an Intrinsically Disordered Protein. *Sci Rep* **9**, 6172 (2019).
6. Zhou, H. X. Intrinsic disorder: signaling via highly specific but short-lived association. *Trends Biochem Sci* **37**, 43-48 (2012).
7. Berlow, R. B., Martinez-Yamout, M. A., Dyson, H. J., Wright, P. E. Role of backbone dynamics in modulating the interactions of disordered ligands with the TAZ1 domain of the CREB-binding protein. *Biochemistry* **58**, 1354-1362 (2019).
8. Berlow, R. B., Dyson, H. J., Wright, P. E. Multivalency enables unidirectional switch-like competition between intrinsically disordered proteins. *Proc Natl Acad Sci U S A* **119**, (2022).
9. Bertonecini, C. W., *et al.* Structural characterization of the intrinsically unfolded protein beta-synuclein, a natural negative regulator of alpha-synuclein aggregation. *J Mol Biol* **372**, 708-722 (2007).

10. Barre, P., Eliezer, D. Structural transitions in tau k18 on micelle binding suggest a hierarchy in the efficacy of individual microtubule-binding repeats in filament nucleation. *Protein Sci* **22**, 1037-1048 (2013).
11. Abyzov, A., *et al.* Identification of Dynamic Modes in an Intrinsically Disordered Protein Using Temperature-Dependent NMR Relaxation. *J Am Chem Soc* **138**, 6240-6251 (2016).
12. Gill, M. L., Byrd, R. A., Palmer, A. G., III. Dynamics of GCN4 facilitate DNA interaction: a model-free analysis of an intrinsically disordered region. *Phys Chem Chem Phys* **18**, 5839-5849 (2016).
13. Khan, S. N., *et al.* Distribution of Pico- and Nanosecond Motions in Disordered Proteins from Nuclear Spin Relaxation. *Biophys J* **109**, 988-999 (2015).
14. Rezaei-Ghaleh, N., *et al.* Local and global dynamics in intrinsically disordered synuclein. *Angew Chem Int Ed Engl* **57**, 15262-15266 (2018).
15. Maiti, S., Acharya, B., Boorla, V. S., Manna, B., Ghosh, A., De, S. Dynamic Studies on Intrinsically Disordered Regions of Two Paralogous Transcription Factors Reveal Rigid Segments with Important Biological Functions. *J Mol Biol* **431**, 1353-1369 (2019).
16. Rezaei-Ghaleh, N., Parigi, G., Zweckstetter, M. Reorientational Dynamics of Amyloid-beta from NMR Spin Relaxation and Molecular Simulation. *J Phys Chem Lett* **10**, 3369-3375 (2019).

17. Martin, E. W., *et al.* Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **367**, 694-699 (2020).
18. Lipari, G., Szabo, A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. *Journal of the American Chemical Society* **104**, 4546-4559 (2002).
19. Meng, F., Bellaiche, M. M. J., Kim, J. Y., Zerbe, G. H., Best, R. B., Chung, H. S. Highly disordered amyloid-beta monomer probed by single-molecule FRET and MD simulation. *Biophys J* **114**, 870-884 (2018).
20. Robustelli, P., Piana, S., Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc Natl Acad Sci U S A* **115**, E4758-E4766 (2018).
21. Shrestha, U. R., *et al.* Generation of the configurational ensemble of an intrinsically disordered protein from unbiased molecular dynamics simulation. *Proc Natl Acad Sci U S A* **116**, 20446-20452 (2019).
22. Fawzi, N. L., Ying, J., Ghirlando, R., Torchia, D. A., Clore, G. M. Atomic-resolution dynamics on the surface of amyloid-beta protofibrils probed by solution NMR. *Nature* **480**, 268-272 (2011).
23. von Bergen, M., Friedhoff, P., Biernat, J., Heberle, J., Mandelkow, E. M., Mandelkow, E. Assembly of tau protein into Alzheimer paired helical filaments

- depends on a local sequence motif ((306)VQIVYK(311)) forming beta structure. *Proc Natl Acad Sci U S A* **97**, 5129-5134 (2000).
24. Kampf, K., Izmailov, S. A., Rabdano, S. O., Groves, A. T., Podkorytov, I. S., Skrynnikov, N. R. What drives (15)N spin relaxation in disordered proteins? combined NMR/MD study of the H4 histone tail. *Biophys J* **115**, 2348-2367 (2018).
25. Hicks, A., Escobar, C. A., Cross, T. A., Zhou, H. X. Sequence-dependent correlated segments in the intrinsically disordered region of ChiZ. *Biomolecules* **10**, 1-23 (2020).
26. Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput* **11**, 3696-3713 (2015).
27. Piana, S., Donchev, A. G., Robustelli, P., Shaw, D. E. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J Phys Chem B* **119**, 5113-5123 (2015).
28. Hicks, A., Escobar, C. A., Cross, T. A., Zhou, H. X. Fuzzy association of an intrinsically disordered protein with acidic membranes. *JACS Au* **1**, 66-78 (2021).
29. Hicks, A., MacAinsh, M., Zhou, H. X. Removing thermostat distortions of protein dynamics in constant-temperature molecular dynamics simulations. *J Chem Theory Comput* **17**, 5920-5932 (2021).

30. Hou, L., *et al.* Solution NMR studies of the A beta(1-40) and A beta(1-42) peptides establish that the Met35 oxidation state affects the mechanism of amyloid formation. *J Am Chem Soc* **126**, 1992-2005 (2004).
31. Sung, Y. H., Eliezer, D. Residual structure, backbone dynamics, and interactions within the synuclein family. *J Mol Biol* **372**, 689-707 (2007).
32. Mylonas, E., Hascher, A., Bernado, P., Blackledge, M., Mandelkow, E., Svergun, D. I. Domain conformation of tau protein studied by solution small-angle X-ray scattering. *Biochemistry* **47**, 10345-10353 (2008).
33. Ahmed, M. C., *et al.* Refinement of alpha-synuclein ensembles against SAXS data: comparison of force fields and methods. *Front Mol Biosci* **8**, 654333 (2021).
34. Bernado, P., Blackledge, M. A self-consistent description of the conformational behavior of chemically denatured proteins from NMR and small angle scattering. *Biophys J* **97**, 2839-2845 (2009).
35. Rao, J. N., Kim, Y. E., Park, L. S., Ulmer, T. S. Effect of pseudorepeat rearrangement on alpha-synuclein misfolding, vesicle binding, and micelle binding. *J Mol Biol* **390**, 516-529 (2009).
36. Schwalbe, H., *et al.* Structural and dynamical properties of a denatured protein. Heteronuclear 3D NMR experiments and theoretical simulations of lysozyme in 8 M urea. *Biochemistry* **36**, 8977-8991 (1997).

37. Bertoncini, C. W., *et al.* Release of long-range tertiary interactions potentiates aggregation of natively unstructured alpha-synuclein. *Proc Natl Acad Sci U S A* **102**, 1430-1435 (2005).
38. Dedmon, M. M., Lindorff-Larsen, K., Christodoulou, J., Vendruscolo, M., Dobson, C. M. Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *J Am Chem Soc* **127**, 476-477 (2005).
39. Salmon, L., *et al.* NMR characterization of long-range order in intrinsically disordered proteins. *J Am Chem Soc* **132**, 8407-8418 (2010).
40. Peterson, D. W., Zhou, H., Dahlquist, F. W., Lew, J. A soluble oligomer of tau associated with fiber formation analyzed by NMR. *Biochemistry* **47**, 7393-7404 (2008).
41. Pastor, R. W., Brooks, B. R., Szabo, A. An analysis of the accuracy of Langevin and molecular dynamics algorithms. *Molecular Physics* **65**, 1409-1419 (1988).
42. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81**, 3684-3690 (1984).
43. Joshi, R., *et al.* Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* **131**, 530-543 (2007).

44. Houben, K., Marion, D., Tarbouriech, N., Ruigrok, R. W., Blanchard, L.
Interaction of the C-terminal domains of sendai virus N and P proteins:
comparison of polymerase-nucleocapsid interactions within the paramyxovirus
family. *J Virol* **81**, 6807-6816 (2007).
45. Ghosh, A., Kota, D., Zhou, H. X. Shear relaxation governs fusion dynamics of
biomolecular condensates. *Nat Commun* **12**, 5995 (2021).
46. Case, D. A., *et al.* AMBER 2018, University of California, San Francisco.
(2018).
47. Guo, C., Zhou, H. X. Fatty acids compete with Abeta in binding to serum albumin
by quenching Its conformational flexibility. *Biophys J* **116**, 248-257 (2019).
48. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., Zhang, Y. The I-TASSER Suite:
protein structure and function prediction. *Nat Methods* **12**, 7-8 (2015).
49. Kallberg, M., *et al.* Template-based protein structure modeling using the RaptorX
web server. *Nat Protoc* **7**, 1511-1522 (2012).
50. Feldman, H. J., Hogue, C. W. Probabilistic sampling of protein conformations:
new hope for brute force? *Proteins* **46**, 8-23 (2002).
51. Ryckaert, J.-P., Ciccotti, G., Berendsen, H. J. C. Numerical integration of the
cartesian equations of motion of a system with constraints: molecular dynamics of
n-alkanes. *J Comput Phys* **23**, 327-341 (1977).

52. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., Pedersen, L. G. A smooth particle mesh Ewald method. *J Chem Phys* **103**, 8577-8593 (1995).
53. Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S., Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J Chem Theory Comput* **9**, 3878-3888 (2013).
54. Han, B., Liu, Y., Ginzinger, S. W., Wishart, D. S. SHIFTX2: significantly improved protein chemical shift prediction. *J Biomol NMR* **50**, 43-57 (2011).
55. Nielsen, J. T., Mulder, F. A. A. POTENCI: prediction of temperature, neighbor and pH-corrected chemical shifts for intrinsically disordered proteins. *J Biomol NMR* **70**, 141-165 (2018).
56. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A., Sali, A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res* **44**, W424-429 (2016).
57. Tesei, G., Martins, J. M., Kunze, M. B. A., Wang, Y., Crehuet, R., Lindorff-Larsen, K. DEER-PREdict: Software for efficient calculation of spin-labeling EPR and NMR data from conformational ensembles. *PLoS Comput Biol* **17**, e1008551 (2021).

58. Roe, D. R., Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **9**, 3084-3095 (2013).
59. McGibbon, R. T., *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J* **109**, 1528-1532 (2015).

Acknowledgments

This work was supported by National Institutes of Health Grant R35 GM118091. We thank Alan Hicks and Ramash Prasad for technical assistance.

Author contributions

S.D, M.M., and H.-X.Z. designed and performed the research, analyzed the data, and wrote the manuscript.

Competing financial interests: The authors declare no competing financial interests.

Table 1. Sequences of eight IDPs.

IDP name	# of Res	Sequence ^a
Aβ40	40	1 DAEFRHDSGY EVHHQKLVFF AEDVGSNKG A IIGLMVGGVV 40
HOX-SCR	87	298 KKNPPQIYPW MKRVHLGTST VNANGETKRQ RTSYTRYQTL 337 338 <u>ELEKEF</u> HFNR <u>YLTRRRRIEI AHALCLT</u> ERQ IKIWFQNRMM 377 378 <u>KWKKE</u> HK 384
HOX-DFD	90	337 TDGERIIPW MKKIHVAGVA NGSYQPGMEP KRQRTAYTRH 376 377 <u>QILELEKEFH</u> YNRYLT <u>RRRR IEIAHT</u> LVLS <u>ERQIKIWFQN</u> 416 417 <u>RRMKWK</u> NDK 426
SEV-NT	124	401 LSGGDGAYHE PTGGGAIEVA LDNADIDLET EAHADQDARG 440 441 WGGESGERWA RQVSGGHFVT LHGAERLEEE TNDED <u>VSDIE</u> 480 481 <u>RRIAMRLAER</u> RQEDSATHGD EGRNNGVDHD EDDDAAAVAG 520 521 IGGI 524
tau K18	129	244 QTAPVPMPLD KNVKSKIGST ENLKHQ <u>PGGG</u> KVQIINKKLD 283 284 LSNVQSKCGS KDNIKHV <u>PGG</u> GSVQIVYKPV DLSKVTSKAG 323 324 SLGNIHHK <u>PG</u> <u>GG</u> QVEVKSEK LDFKDRVQSK IGSLDNITHV 363 364 <u>PGGG</u> NKKIE 372
A1-LCD	131	1 GSMASASSSQ RGRSGSGNFG GGRGGGFGGN DNFGRGGNFS 40 41 GRGGFGGSRG <u>GGGYGGSGDG</u> YNGFGNDGSN FGGGGNYNNQ 80 81 SSNFGPMKGG NFGGRSSGPY <u>GGGGQYFAKP</u> RNQGGYGGSS 120 121 SSSSYGSGRR F 131
β-synuclein	134	1 MDVFMKGLSM <u>AKEGV</u> VAAAE <u>KTKQGV</u> TEAA <u>EKTKEGV</u> LYV 40 41 GS <u>KTREGV</u> VQ GVASVAE <u>KTK</u> <u>EQASHL</u> GGA V FSGAGNIAAA 80 81 TGLVKREEFP TDLKPEEVAQ EAAEEPLIEP LMEPEGESYE 120 121 DPPQEYQEY EPEA 134
α-synuclein	140	1 MDVFMKGLSK <u>AKEGV</u> VAAAE <u>KTKQGV</u> AEAA <u>GKTKEGV</u> LYV 40 41 GSKTKEGV <u>VH</u> GVATVAE <u>KTK</u> <u>EQVTNV</u> GGA V VTGVTAVAQK 80 81 <u>TVEGA</u> GSIAA ATGFVKKDQL GKNEEGAPQE GILEDMPVDP 120 121 DNEAYEMPSE EGYQDYEPEA 140

^aShaded and underlined residues: olive, α-helices; cyan, PGGG near the end of a microtubule-binding motif of tau; yellow, KTKEGV motifs of α- and β-synuclein; underline, GGGG, GGG, or GG.

Figure Captions

Figure 1. Snapshots from MD simulations of three IDPs. (A) A β -40. (B) HOX-DFD. (C) α -synuclein. Basic and acidic residues are colored blue and red, respectively. The N-terminal and C-terminal residue numbers are shown, as are the residue number marking the start of the helical region of HOX-DFD and the residue numbers marking the end of the N-terminal region and the start of the C-terminal region of α -synuclein.

Figure 2. Experimental validation of MD conformational ensembles. (A-C) Comparison of calculated and experimental secondary chemical shifts for A β 40, HOX-DFD, and α -synuclein. RMSE values are shown in the legends. (D) Comparison of calculated (red curves) and experimental (gray bars) PREs by spin labels at 12 residues in α -synuclein.

Figure 3. Backbone NMR relaxation properties. (A) A β 40. (B) HOX-DFD. (C) α -synuclein. Calculated and experimental R_2 , R_1 , and NOE values are compared, with RMSE values shown in the legends. Horizontal dashed lines show average values over the N-half and C-half of A β 40 or of HOX-DFD (demarcated by black vertical lines), or over the entire sequence of α -synuclein. For the latter protein, blue vertical lines are placed at residues 15, 26, 37, 48, 68, and 85 to indicate dips in R_1 (and R_2 in the case of Gly68); red vertical lines are placed at Tyr39, Lys96, and Asp121 to indicated elevated R_2 . Also shown at the top are secondary structure propensities and average contact numbers. The latter are displayed as bars on a gray scale, with white at 0 and black at 5 or above.

Figure 4. Side chain-side chain contact maps. (A) A β 40. Contacts with frequencies higher than 0.005 are shown, but for Arg5-Tyr10 contacts, the threshold is lowered to 0.002. (B) HOX-DFD. Contacts with frequencies higher than 0.02 are shown. Blocks of residues with tendencies for extensive contacts are highlighted by red boxes. For HOX-DFD, α -helices are identified by blue boxes, and prominent interhelical and linker-helix contacts are indicated by magenta and orange ovals. Insets: snapshots illustrating selected contacts.

Figure 5. Time constants and amplitudes from fitting NH time correlation functions to a sum of three exponentials. (A-B) A β 40. (C-D) HOX-DFD. (E-F) α -synuclein. In panels (A), (C), and (E), the left ordinate displays the scale for τ_1 and τ_2 whereas the right ordinate displays the scale for τ_3 . In panels (B), (D), and (F), the sum of the amplitudes, A_{sum} , was not constrained to 1 and is shown in black; the missing amplitude is due to ultrafast motions. Horizontal dashed lines show average values over the N-half and C-half (demarcated by black vertical lines) or over the entire sequence. For α -synuclein, vertical lines are placed in panel (E) at Tyr39, Lys96, and Asp121 to indicate elevated τ_1 , and in panel (F) at residues 15, 26, 37, 48, 68, and 85 to indicate dips in A_{sum} .

Figure 6. Mechanisms by which sequence-dependent backbone dynamics of IDPs codes for functions or contributes to diseases. (A) The fast dynamics of the C-half of A β 40 allows it to readily form a β -sheet between monomers and nucleate fibrillation. (B) The rigid three-helix bundle of a HOX transcription factor allows it to recognize the DNA target; the dynamic N-terminal region can form intramolecular contacts in the free state but engage with the co-transcription factor to further stabilize the complex.

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

Figure 6