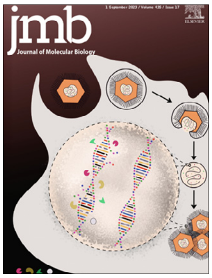
Research Article



Not Only Expansion: Proline Content and Density Also Induce Disordered Protein Conformation Compaction

Milan Kumar Hazra, Yishai Gilron and Yaakov Levy *

Department of Chemical and Structural Biology, *Weizmann Institute of Science, Rehovot, Israel*

Correspondence to Yaakov Levy: *Department of Chemical and Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel. *Koby.Levy@weizmann.ac.il (Y. Levy)*

https://doi.org/10.1016/j.jmb.2023.168196
Edited by Monika Fuxreiter

Abstract

Intrinsically disordered proteins (IDPs) adopt a wide array of different conformations that can be constrained by the presence of proline residues, which are frequently found in IDPs. To assess the effects of proline, we designed a series of peptides that differ with respect to the number of prolines in the sequence and their organization. Using high-resolution atomistic molecular dynamics simulations, we found that accounting for whether the proline residues are clustered or isolated contributed significantly to explaining deviations in the experimentally-determined gyration radii of IDPs from the values expected based on the Flory scaling-law. By contrast, total proline content makes smaller contribution to explaining the effect of prolines on IDP conformation. Proline residues exhibit opposing effects depending on their organizational pattern in the IDP sequence. Clustered prolines (*i.e.*, prolines with $\leq$2 intervening non-proline residues) result in expanded peptide conformations whereas isolated prolines (*i.e.*, prolines with >2 intervening non-proline residues) impose compacted conformations. Clustered prolines were estimated to induce an expansion of ~20% in IDP dimension (via formation of PPII structural elements) whereas isolated prolines were estimated to induce a compaction of ~10% in IDP dimension (via the formation of backbone turns). This dual role of prolines provides a mechanism for conformational switching that does not rely on the kinetically much slower isomerization of *cis* proline to the *trans* form. Bioinformatic analysis demonstrates high populations of both isolated and clustered prolines and implementing them in coarse-grained molecular dynamics models illustrates that they improve the characterization of the conformational ensembles of IDPs.

Introduction

Intrinsically disordered proteins (IDPs) lack a stable three-dimensional structure and therefore are described by an ensemble of rapidly convertible conformations. The conformational ensembles of IDPs are therefore much more expanded than those of folded proteins, but are commonly more compact than the ensemble of random-coil polymers. Similarly to folded proteins, the conformational preferences of IDPs are dictated by their sequence, however, these two classes of proteins differ with respect to the organization of amino acids in their sequences. IDPs often exhibit simple and redundant amino acid patterns that correspond to a large conformational space compared with the much smaller sequence space associated with the unique three-dimensional structure of folded proteins. The sequences of IDPs also differ from those of folded proteins by tending to use a narrower variety of amino acids. Compared with folded proteins, IDP sequences are often more enriched in charged residues (*i.e.*, Glu, Asp, Lys, and Arg), small residues (*i.e.*, Gly, Ala, Ser, and Gln), and particularly in proline.[1,2]

The presence of proline residues is expected to affect the conformational ensembles of IDPs not only because of their high abundance but also because of their unique characteristics compared with other amino-acids. Proline is unique because of its ability to impose a conformational bias owing to the cyclic nature of the pyrrolidine ring that connects the backbone nitrogen to its side-chain, resulting in the absence of the usual backbone nitrogen proton and thereby preventing participation in common secondary structural elements.[1,3] Consequently, prolines disrupt the propagation of regular secondary structural elements and are often treated as "disorder promoters".[4] Although prolines tend to break secondary structures, they can assist in nucleating alpha helices and in promoting turn formation.[5,6] Furthermore, they may participate in the formation of alternative local structural elements. In particular, the intrinsic chain propensities of proline residues enable them to form helical secondary structural elements referred to as polyproline II (PPII).[7–10] The PPII conformation results in a locally expanded backbone conformation.[11] For proline-based PPIIs, propagation of the helix depends on proline content and organization, with PPII content decreasing as the number of spacing residues between prolines increases.[12,13] However, proline residues are not an obligatory component of PPIIs and PPIIs comprised entirely of non-proline sequences have also been found.[14]

Prolines are recognized as encoding a local stiffening of the peptide backbone. This arises from the steric constraint that the cyclic nature of proline's side-chain imposes on the Cα–CO angle (psi, Ψ) in the peptide backbone. The introduction of a steric constraint reduces the energetic barrier to converting the omega (Ω) angle around the C(O)–N peptide bond from 180° (as per a *trans* configuration) to 0° (as per a *cis* configuration) and thereby increases the probability that the backbone will adopt a *cis* configuration, while also potentially affecting the phi (Φ) angle around the Cα–N bond. Similarly to any other amino-acid (X), proline strongly favors the *trans* configuration compared with the *cis* configuration, however the preference is weaker for X–P bonds compared with others. Consequently, 5–10% of X–P bonds in solved protein structures are found to be in the *cis* configuration, which is about 10 times greater than for X–X bonds.[15–17] The *cis* configuration of prolines induces a more compact state and it was therefore suggested that *cis* to *trans* isomerization may act as a switch between compact and expanded structural ensembles.[18–21] Indeed, proline has been found to play various roles, including as a switch in cell signaling[21–23] and folding kinetics.[24,25] Prolines are the second most abundant residue in loops within globular domains.[26] Particularly, an isolated proline was found to mediate loop formation in transmembrane proteins separating two transmembrane segments through a fairly tight turn by reducing the conformational freedom.[27] The presence of a high proline content of *trans* proline configurations in an IDP increases its tendency to form PPII that can expand conformations.

Proline has been extensively investigated because of its unique and versatile chemical and structural features and the central role it plays in various biomolecular processes, such as protein folding,[25] protein–protein interactions,[28] post-translational modifications,[19,24] secondary structure formation[13,29–31] and even in liquid–liquid phase separation.[32,33] However, the anomalous features of proline that are linked with different structural outcomes mean that proline is not fully understood in all its complexity, particularly with respect to its role in IDPs.

Although the multifaceted roles of prolines in modulating the structures of folded proteins for various functions have been well explained,[3,11,34] the essence of prolines in IDPs remains the subject of a long-running debate.[22,35,36,30,37] The much higher abundance of proline in IDPs compared with folded proteins may indicate that it plays a crucial role in the former. Indeed, prolines were shown to affect the conformations of some IDPs[38,32] and to contribute to their ability to form amyloids.[39] In many other cases, the contribution of prolines to the conformational preferences of proteins is disregarded. For example, the molecular grammar that is often considered to dictate the molecular properties of IDPs and their interactions in liquid–liquid phase separation includes primarily charge–charge and cation–π interactions[40–50] without an explicit representation for the unique nature of proline residues. Such effective residue–residue potentials that are scaled based on the hydrophobicity of all pairwise interactions that are supplemented by electrostatic interactions were found to be powerful to estimate the dimension (quantified by the radius of gyration, Rg) of many IDPs.[51]

In some instances, prolines were acknowledged to affect the dimensions of IDPs. However, in such cases, only total proline was used as a measure to estimate the effect of proline residues on protein dimension, assuming that greater numbers of prolines produce more expanded conformations.[52] However, given the complexity of proline residues, as found in folded proteins, it is suspected that the number of prolines in a given sequence may not be a reliable predictor of their effect on conformational preferences. In other cases, the effect of prolines on IDP conformation was argued to be coupled to the effect of charged residues[19] or of adjacent aromatic residues.[15]

In this study, we examined the effect of prolines on the conformational ensemble of IDPs by studying a series of 11–21 residue peptides that varied with respect to the number of proline residues they contained and their organization. Using atomistic molecular dynamics simulations, we quantified how the spacing between the prolines affects the conformational ensemble of the peptides and particularly examined whether prolines induce expansion or compaction compared with peptides that lack any prolines. The results of the atomistic simulations suggest that prolines should be classified in two groups depending on their organizational pattern in the amino acid sequence. In light of the novel biophysical insights obtained from the atomistic simulations, we constructed and calibrated coarse-grained models to isolate the contribution proline residues make to the global dimensions, as measured by the radius of gyration (Rg) of 33 selected IDPs, and compared the results with Rg values modelled or estimated by other means.

Results

Proline residues are highly abundant in IDP sequences, with their content in IDPs being ~80% higher than in folded proteins.[2] However, the exact nature of their contributions to the conformational ensemble of IDPs remains poorly understood. To better understand the effect of proline residues on IDPs, two series of intrinsically disordered peptides were designed, with variation with respect to proline content and organization. These peptide series comprised 0–3 prolines (total proline content ~0–30%) separated by 0–5 Gly or Ser residues. Because the trans configuration of proline is much more populated than the cis configuration, particularly in IDPs,[16] we modeled all proline in the trans configuration. The conformational ensemble of each of these peptides was sampled using atomistic molecular dynamics simulations and was quantified in terms of its mean Rg and S values.

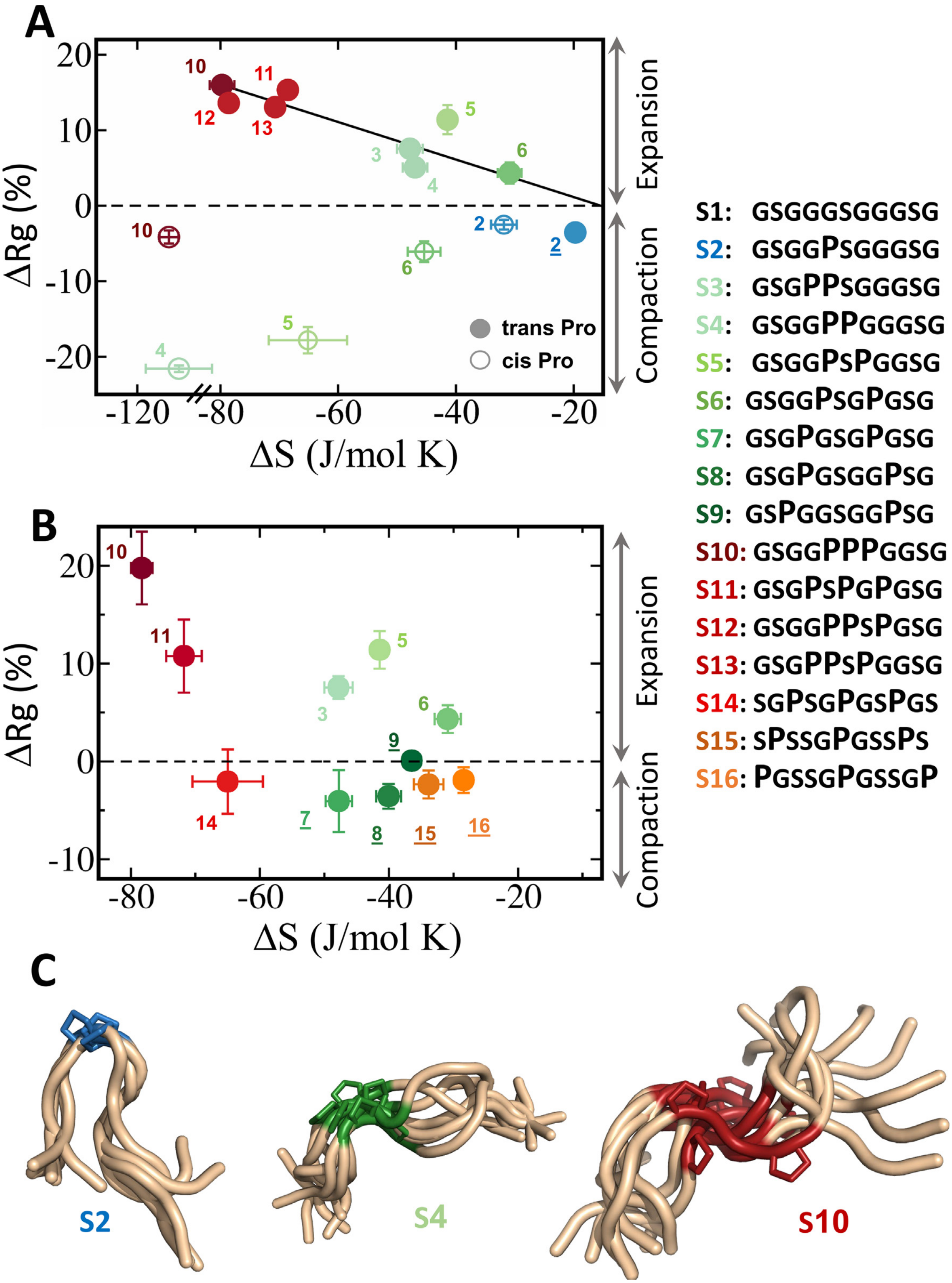
Proline content is sometimes, but not always, correlated with the dimensions of peptide conformations

The effect of the presence of proline on the dimensions of the peptide conformation was examined by comparing sequences that contained proline with the proline-free S1 sequence consisting solely of Gly and Ser residues, which served as a control. Variation of radius of gyration (ΔRg) and entropy (ΔS) of sequences containing prolines was computed as $\Delta Rg = [Rg^{Si}\text{-}Rg^{S1}]/Rg^{S1}$ and $\Delta S = [S^{Si}\text{-}S^{S1}]$, where superscript 'Si' represents all the designed sequences with prolines and 'S1′ represents a pure G/S repeat that follows an ideal Flory scaling. Figure 1(A) shows ΔRg versus ΔS for each of the studied disordered peptides compared with the control S1 sequence. A linear correlation is observed between ΔRg and ΔS as proline content increases, indicating that, in these sequences, increasing the number of prolines has a gradual effect. The presence of two or three prolines increased Rg by approximately 7–12% for sequences S3–S6 and by about 10–15% in sequences S10–S13, with this expansion accompanied by a decrease in ΔS of approximately 75 J/mol K (about 25 J/mol K per proline residue). We note that the sequences in Figure 1(A) include several peptides containing either two or three prolines, yet in all cases the prolines are quite clustered and the spacing between consecutive prolines does not exceed two Gly or Ser residues. The expansion of the peptide conformations (as indicated by ΔRg > 0) in the presence of several clustered prolines separated by $\leq$ 2 residues may correspond to the formation of PPII structural elements that are known to be extended and to increase Rg.[11,53]

The effect of a single proline (sequence S2) on the conformational space of the peptide is surprising (Figure 1(A)), as it exhibits a negative ΔRg, indicating compaction rather than the expansion effect observed for peptides with two or three prolines (sequences S3–S6 and S10–S13). Accordingly, a single proline induces compaction whereas two or three prolines induce expansion. The size of the compaction effect (ΔRg) of a single proline residue is about 4% and it is accompanied by a decrease in ΔS of about 20 J/mol K.

A visual representation of the effect of the number of prolines on peptide conformation is shown in Figure 1(C). Whereas an isolated proline introduces compaction at its position in the sequence, clustered prolines introduce expansion. Figure 1 shows the results obtained using the CHARMM36m force field, with similar results obtained from simulations performed using the Amber-99-SB-ILDN forcefield (Figure S3).

To improve our understanding of the compaction induced by a single proline in the trans configuration, we simulated some of the designed peptides with all the prolines modeled in the rarer cis configuration, which is acknowledged to yield compact conformations.[15] Figure 1(A) shows, as expected, that introducing cis prolines compacted the modelled IDPs. A single cis proline produced compaction (ΔRg) of ~4%, similarly to the effect of a single trans proline. Two cis prolines had a greater compaction effect (ΔRg) of 5–20%, depending on the number of intervening residues. The largest effect, of about 20%, was obtained for two consecutive cis prolines (sequence S4), with the two prolines most likely having a synergistic effect. For three consecutive cis prolines (sequence S10), a smaller compaction effect of only about 4% was observed, possibly because the effects of the prolines partially canceled each other out (the effect of cis or tarns proline of the ψ dihedral angles are shown in Figures S11 and S12).


A
ΔRg (%)
ΔS (J/mol K)
20
10
0
-10
-20
-120
-80
-60
-40
-20
10
11
12
13
3
4
5
6
2
2
10
6
5
4
Expansion
Compaction
● trans Pro
○ cis Pro
B
ΔRg (%)
ΔS (J/mol K)
20
10
0
-10
-80
-60
-40
-20
10
11
3
5
6
9
14
7
8
15
16
Expansion
Compaction
C
S2
S4
S10
S1: GSGGGSGGGSG
S2: GSGGPSGGGSG
S3: GSGPPSGGGSG
S4: GSGGPPGGGSG
S5: GSGGPSPGGSG
S6: GSGGPSGPGSG
S7: GSGPGSGPGSG
S8: GSGPGSGGPSG
S9: GSPGGSGGPSG
S10: GSGGPPPGGSG
S11: GSGPSPGPGSG
S12: GSGGPPSPGSG
S13: GSGPPSPGGSG
S14: SGPSGPGSPGS
S15: SPSSGPGSSPS
S16: PGSSGPGSSGP

In addition to the effect of proline content on the conformational ensemble of disordered peptides, their organizational pattern in the sequence, as defined by the spacing between consecutive prolines, may also play an important role. Figure 1 (B) shows the values of ΔRg and ΔS for peptides containing two or three prolines separated by 0–5 Gly or Ser residues and showing both expansion and compaction effects. An expansion in conformation is observed for peptides with $\leq$2 non-proline residues between consecutive prolines whereas compaction is observed when proline residues are spaced by >2 residues.

Overall, we found that the dimensions of peptides with two clustered prolines (S3–S6) expanded by ΔRg = 5–10% and the dimensions of peptides containing three clustered prolines (S10–S13) expanded by ΔRg = 15–20% (Figure 1(A)), whereas most peptides with 1–3 isolated prolines (S2, S7–S8, S15, and S16) underwent compaction of 3–5%. These observations suggest that proline content and the ΔRg of disordered peptides are correlated only when the prolines are clustered (Figure 1(A)). For isolated prolines, not only is there no correlation, but the proline often exerts an opposite compaction effect on the conformations (Figure 1(B)).

The specific organizational pattern of proline is linked to compaction or expansion of IDPs

To better understand the effect of organizational pattern (in terms of the spacing between consecutive prolines) on IDP conformation, we measured free energy (ΔG) where Rg served as a reaction coordinate. Figure 2(A) shows the conformational free energy landscape for sequences with two clustered prolines separated by $\leq$2 Gly or Ser residues (S4–S6; upper panel) and the corresponding landscapes for sequences containing two isolated prolines separated by 3–5 Gly or Ser residues (S7–S9; lower panel). A comparison between the free energy profiles for sequences with *trans* or *cis* proline is shown in Figures S14 and S15. It is evident that different organizational patterns have differently populated conformational free energy landscapes. Peptides with two clustered prolines (S4–S6) are mostly characterized by a single energy basin that corresponds to extended conformations with a high Rg value (Figures 2(A) and S4). Peptides with isolated prolines (S7–S9) are characterized by dual-basin energy landscapes and the basin corresponding to the more compact state is significantly more populated. These dual-basin landscapes indicate that peptides with isolated prolines may exist in equilibrium between two major classes of conformations with distinct Rg values and that the transition between them may be affected by details of the sequence (e.g., the number of intervening residues between the prolines or the identity of the residues adjacent to the prolines). A peptide with a single proline is also characterized by a dual-basin energy landscape, with the basin corresponding to the more compact state being the more stable

◀

Figure 1. Proline content and patterns affect peptide conformations. The effect of proline residues on the conformation of intrinsically disordered peptides was assessed by plotting $\Delta Rg = [Rg^{Sn}-Rg^{S1}]/Rg^{S1}$, where Rg^{Sn} is the backbone's radius of gyration of the proline-containing sequences Sn, with n = 2–16 and Rg^{S1} is the radius of gyration of the non-proline-containing sequence (S1) against change in entropy (ΔS), which was calculated in a similar manner. The 16 studied peptides (right panel) contain 0–3 proline residues in the *trans* configuration (filled circles) or *cis* configuration (empty circles) arranged in different organizational patterns, with 0–5 Gly or Ser residues placed between consecutive proline residues. Peptide sequences containing two proline residues are represented in shades of green-olive whereas those containing three proline residues are represented in shades of orange-red. A) Plot of ΔRg versus ΔS for sequences with clustered prolines (defined as those with $\leq$2 Gly or Ser residues between consecutive proline residues). ΔRg and ΔS are linearly correlated, reflecting conformational expansion of the protein with increasing numbers of proline residues. For peptide S2, ΔRg is negative, reflecting a compaction effect from a single proline. Some selected peptides that were modeled with prolines in cis configurations and show compaction by the prolines were added to the plot. B). Plot of ΔRg versus ΔS for two series of sequences with two or three prolines with gradual increase in the number of intervening Gly or Ser between consecutive proline residues. Each of the series include sequences with isolated proline residues (defined as those with >2 Gly or Ser residues between consecutive proline residues; *i.e.,* S7–S9, S15, and S16) and clustered prolines (defined as those with $\leq$2 Gly or Ser residues between consecutive proline residues; *i.e.,* S4–S6, S10, S11 and S14). It is evident that isolated peptides have the effect of compacting the protein conformation, regardless of the total number of prolines in the sequence. C). Conformational snapshots of example peptides with 1, 2 or 3 proline residues that illustrate compaction (for S2) and expansion (for S4 and S10). Sequences with isolated prolines are designated by an underline. The simulated sequences were comprise 11 residues, except S10-S16 in (B) that comprise 21 residues to accommodate the many intervening residues between the consecutive prolines. These results were obtained from atomistic molecular dynamics simulations using CHARMM36m. Similar results using Amber are shown in Figure S3.

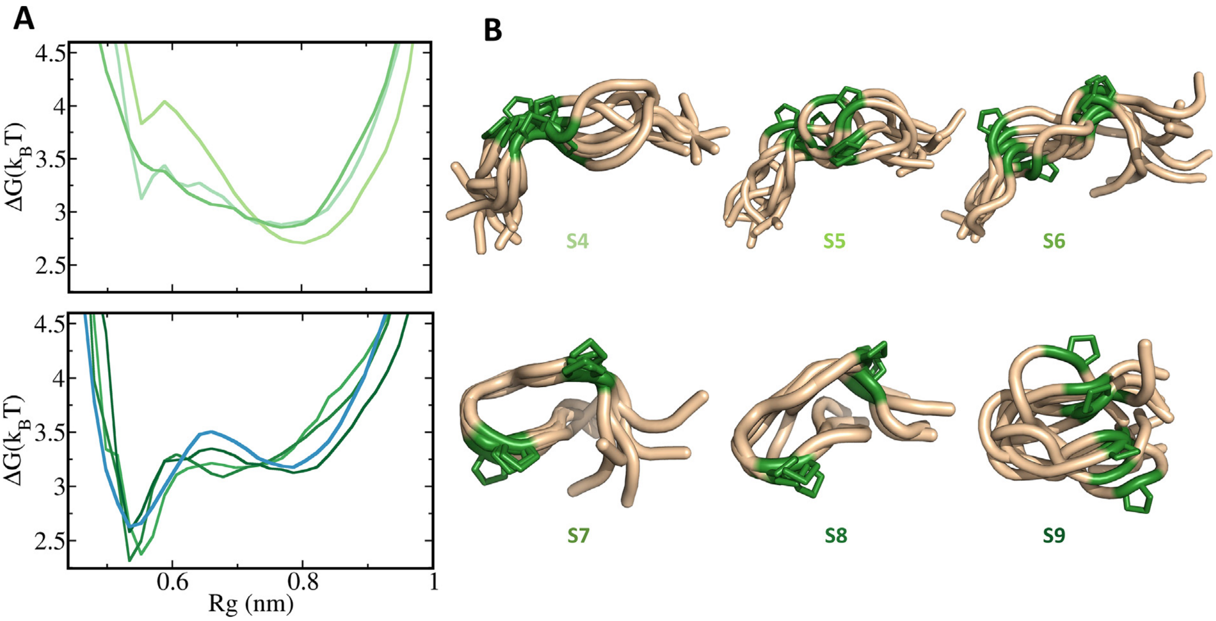


Figure 2. Free energy profiles for peptides containing two proline residues in different organizational patterns. The free energy (ΔG) profiles of intrinsically disordered proteins S3–S8, in which the proline residues (green) are separated by 0–5 Gly or Ser residues (see Figure 1 for exact sequences) are plotted as a function of their radius of gyration (Rg), as an order parameter. A) The free energy profiles for peptides with $\leq$2 residue between the two prolines (S4–S6) are characterized by a single dominant basin that corresponds to expanded conformations due to the formation of PPII structural elements (upper panel). Peptides with prolines separated by >2 residues (S7-S9) are characterized by dual-basin landscapes similar to that of the peptide with a single isolated proline (S2) (bottom panel, blue line). In these peptides with isolated prolines, the basin with the lower energy has the smaller Rg and more-compact conformation. B) Representative snapshots depicting the conformational ensembles of peptides with clustered prolines and thus more expanded conformations (upper panel) and peptides with isolated prolines and thus more compact conformations (bottom panel).

(Figure 2(A)), similarly to the free-energy landscape of peptides with two isolated prolines. For isolated prolines, the deeper free energy basin of the compact state results in such systems having a compact conformation (Figure 1).

The local effect of proline residues on IDP conformation is pattern dependent

At this point, the effect of local changes potentially introduced by consecutive prolines arranged in different organizational patterns (isolated versus clustered) on global IDP conformational variety remained unclear. To understand how prolines tune global peptide conformations, we calculated the ψ dihedral angles for sequences with different patterns of prolines in *cis* or *trans* configurations (Figures S11 and S12). The structural consequence of proline content and patterns can be better probed by the individual backbone Cα dihedral quartets (δ angles) involving proline residues. To correlate between structural changes and the backbone Cα dihedral angles of residue quartets, we plotted joint probability distributions of individual dihedral segments affected by prolines and their end-to-end distance for several sequences with 0–2 prolines as contour plots. Figure 3 shows such contour plots for the angles XPXX and XXPX as a function of their corresponding end-to-end distances (see also Figure S5). Similar plots for angles PXXX and XXXP (which were found to be less affected by whether the proline residues were isolated or clustered) are shown in Figure S2.

The dihedral angles XPXX and XXPX behave differently for isolated and clustered prolines (namely, as spacing between consecutives prolines increases). The identity of the residues that define each Cα dihedral angles are highlighted by frames in the sequences analyzed in Figure 3 (right panel). For the angle XPXX (Figure 3(A)), a flattened distribution is observed for the sequence that lacks proline residues (S1; left panel), illustrating that it almost equally populates all values of the angle. However, the sequence with a single proline (S2) spans two distinct basins with similar probability: δ_{XPXX} ranges from -2 to 2 rad. The basin at $\delta_{XPXX} = \pm 2$ rad corresponds to expanded conformations ($R^{XPXX} \sim 1.0$ nm), whereas the basin at $\delta_{XPXX} = 0$ rad corresponds to more compact conformations ($R^{XPXX} \sim 0.5$ nm). For the peptide with two adjacent prolines (S4), only a single distinct basin is observed at $\delta_{XPXX} = 2$ rad,

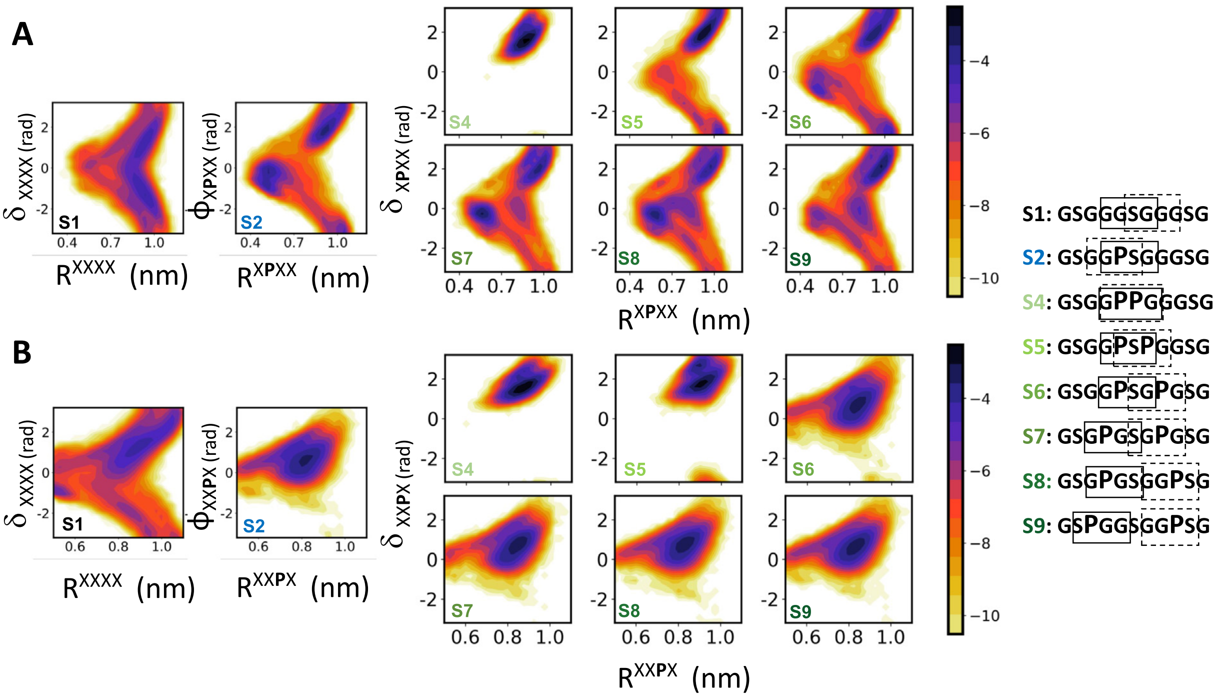


Figure 3. Prolines impose local conformational preferences depending on their organizational pattern. The local conformational consequence of proline is probed by calculating the joint probability distribution of the backbone dihedral angles, Φ, between four consecutive C α carbons in the vicinity of the proline residue and the end-to-end distance, R, of these four residues. Maps of Φ versus R for a peptide that lacks prolines (S1), a peptide with a single proline residue (S2), and peptides with two prolines in different organizational patterns (S4–S9) are displayed for: A) the XPXX quartet and B) the XXPX quartet (where P stands for proline and X for a Gly or Ser residue). We focus on the dihedral angles XPXX and XXPX (marked by solid and dashed rectangles, respectively, on each sequence; right panel) as they are strongly affected by structural changes mediated by the spacing between two prolines. Similar analysis for the angles between PXXX and XXXP are shown in Figure S2. Probability is indicated by the color bar (higher probabilities have bluer shades) on a log scale.

indicating a significant expansion of the quartet to an end-to-end distance (R^{XPXX}) of nearly 1.0 nm. For the peptides with two consecutive prolines that are separated by one and two amino-acids (sequences S5 and S6, respectively), an additional basin (at $\delta_{XPXX} = 0$ rad and R^{XPXX} = 0.5 nm) starts to be populated, but the dominant basin is the one corresponding to the expanded conformation, consistently with the domination of the energy landscape by a single basin when it is projected along peptide Rg (Figure 2). Upon increasing the spacing between the two prolines to 3–5 residues (sequences S7–S9), the basin that corresponds to the more compact conformation becomes more populated and the contour plots resembles that of a single proline (sequence S2). The contour plots illustrate that the transition from peptides with clustered prolines to peptides with isolated prolines is accompanied by the appearance of another state that corresponds to compact conformations.

The contour plots for the dihedral angles between the quartet XXPX versus their corresponding end-to-end distance (Figure 3(B)) exhibit somewhat different behaviors compared with the quartet XPXX, yet they support a similar conclusion. For the peptides with clustered prolines (sequences S4–S6), the population is centered on a restricted region of the space (defined by $\delta_{XXPX} = 2$ rad and R^{XXPX} = 0.9 nm) that corresponds to expanded conformations. The population starts to shift to lower angle δ_{XXPX} values and an end-to-end distance of ~0.5 nm as the separation between the two proline increases (*i.e.*, through the transition from clustered to isolated prolines). A separation of three residues or more (sequences S7–S9) between the two prolines enhances the population of $\delta_{XXPX} = 0$ rad and therefore increases the population of more-compact overall peptide conformations.

The effect of isolated prolines on IDP conformations may depend on neighboring residues

Although the effect of proline residues on the conformational preferences of disordered peptides is related to the intrinsic molecular properties of proline, the identify of neighboring residues may modulate proline's effects. The effect of

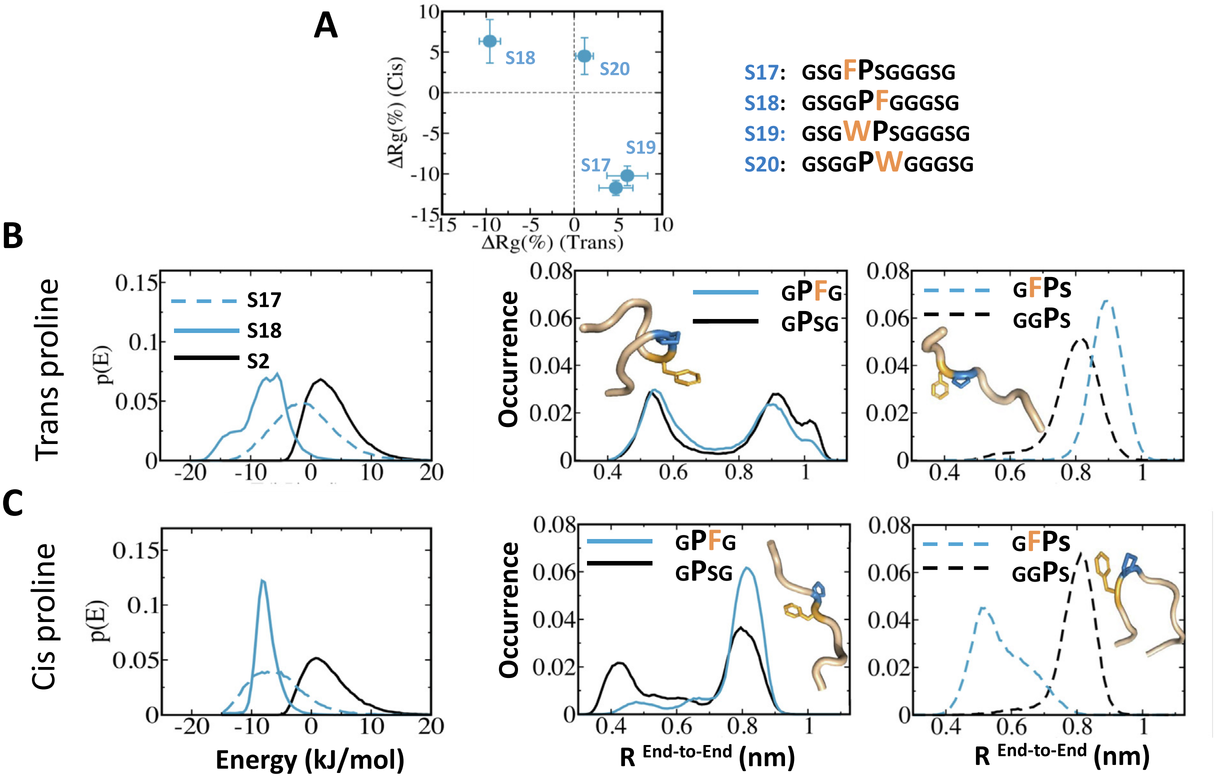


Figure 4. The influence of neighboring aromatic residues on proline's effect on peptide conformational preferences. An aromatic residue (Trp (W) or Phe (F)) was positioned next to a single isolated proline residue (*i.e.*, in position *i* ± 1 relative to proline), which was modeled in either the *cis* or *trans* configuration. A) The effects of the specific aromatic residue and its position assessed in terms of ΔRg, being the deviation of the radius of gyration (Rg) of peptides containing Phe (S17 and S18) or Trp (S19 and S20) from that of S1. Each of the four peptides was simulated where the proline is either in *cis* or in *trans* configuration. B–C). Energetic and distance analysis comparing peptides S17 (F in the *i*-1 position; dashed blue line) and S18 (F in the *i* + 1 position; solid blue line) containing a proline residue in the *trans* configuration (B) or the *cis* configuration (C) with S2, which lacks aromatic residues and whose single isolated proline can be viewed as within a G-P-S-G (solid black line) or G-G-P-S (dashed black line) quartet. The energetic analysis is for the interaction between the side chains of the adjacent aromatic and proline residues. The distance (R) refers to the end-to-end distance of the four residues centered on the proline and aromatic residues. Snapshots illustrate the conformational preferences of the IDPs. When the aromatic residue is positioned *i* + 1 relative to *trans* proline residue *i*, it induces parallelization of sidechain rings, which results into a more-compact overall conformation (S17 in panel B), whereas when the aromatic residue is positioned *i*-1 from the *trans* proline the side-chains interact weakly, resulting in an extended conformation (S18 in panel B).

neighboring residues is expected to be more applicable to isolated prolines because of their dual-basin energy landscape whose population can be shifted by the local environment. Following a recent report on the effect of aromatic residues on the conformational preferences of prolines in their vicinity,[15] we focused on the effect of Phe and Trp in positions *i* + 1 (S18 and S20) and *i*-1 (S17 and S19) relative to a proline residue *i* modeled in the *trans* or *cis* configuration.

We found that Phe and Trp interact with proline residue *i* to a different extent depending on whether the aromatic residues are located closer to the C-terminus (*i.e.*, at the *i* + 1 position) or the N-terminus (*i.e.*, at the *i*-1 position) relative to the proline and on whether the proline is in a *cis* or *trans* configuration (Figures 4 and S6-S8). For *trans* configured prolines, the interaction proceeds more readily when the aromatic residues are located closer to the C-terminus (sequences S18 and S20) because this position enables preferential side-chain interactions to occur, leading to compaction.

The stabilization obtained from such interactions between the aromatic side chain and the *trans* proline can lock the conformation of the IDPs in a compact state leading to 10% reduction in Rg compared with an IDP lacking prolines (sequence S1) (Figure 4(A)), which corresponds to a stabilization energy that can reach −20 kJ/mol (Figure 4(B), left panel). Interactions driven by Phe or Trp at position (*i*-1) relative to the *trans*

A

CFTR (ΔRg=-18%; Tot Pro=4.32%; Clus Pro=0%)
SAERRNSILTETLHRFSLEGDAPVSWTETKKQSFKQTGEFGEKRKNSILNPINSIRKFSIVQKTPLQMNGIE
EDSDEPLERRLSLVPDSEQGEAILPRISVISTGPTLQARRRQSVLNLMTHSVNQGQNIHRKTTASTRKVSL
APQANLTELDIYSRRLSQETGLEISEEINEEDLKECFFDDME

MeCP2 (ΔRg=-13%;Tot Pro=10%;Clus Pro=39%)
EGSGSAPAVPEASASPKQRRSIIRDRGPMYDDPTLPEGWTRKLKQRKSGRSAGKYDVYLINPQGKAFR
SKVELIAYFEKVGDTSLDPNDFDFTVTGRGSPSRREQKPPKKPKSPKAPGTGRGRGRPKGSGTTRPKA
ATSEGVQVKRVLEKSPGKLLVKMPFQTSPGGKAEGGGATTSTQVMVIKRPGRKRKAEADPQAIPKKRG
RKPGSVVAAAAAEAKKKAVKESS

Human NCBD (ΔRg=15%; Tot Pro=15%;Clus Pro=75%)
ISPLKPGTVSQQALQNLLRTLRSPSSPLQQQQVLSILHANPQLLAAFIKQRAAKYANSNPQPIPGQPG
MPQGQPGLQPPTMPGQQGVHSNPAMQNMNPMQAGVQR

II1(ΔRg=31%; Tot Pro=37%; Clus P=94%)
ISGKPVGRRPQGGNQPQRPPPPPPGKPQGPPPQGGNQSQGPPPPPGKPEGRPQGRNQSQGPP
PHPGKPERPPPQGGNQSQGTPPPPGKPERPPPQGGNQSHRPPPPPGKPERPPPQGGNQSRG
PPPHRGKPEGPPPQEGNKSR

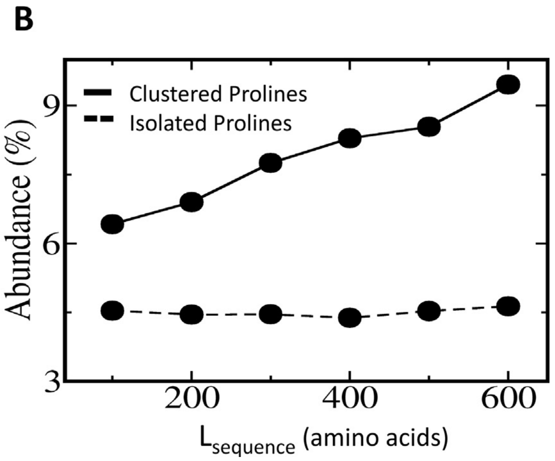


Figure 5. Abundance of isolated and clustered prolines across the human proteome. A) The organizational patterns of proline residues in four human intrinsically disordered proteins (IDPs) are shown for proteins whose proline content ranges from 4.3% to 37%. Isolated prolines (*i.e.*, where consecutive proline residues are separated by >2 other residues) are colored blue and clustered prolines (*i.e.*, where consecutive proline residues are separated by ≤2 other residues) are colored red. Aromatic residues are shown in green. The percentage of clustered prolines in these proteins is 0–100%. Parentheses next to each IDP name present data for the deviation of the experimental radius of gyration (Rg) of these IDPs from that estimated by the Flory scaling law for IDPs [ΔRg=(Rg^{Exp}-Rg^{IDP})/Rg^{IDP})]; proline residues as a percentage of the total number of residues in the sequence (Tot Pro); and clustered prolines as a percentage of all prolines in the sequence (Clus Pro). B) A bioinformatic analysis of the abundance of isolated (dashed line) and clustered (solid line) prolines in intrinsically disordered regions of the human proteome plotted against sequence length (L).

proline residue (sequences S17 and S19) are weaker and more transient than those occurring when the aromatic residue is at position (*i* + 1) and they favor an extended state due to flanking sidechain–sidechain cross interactions (Figure 4 (B)).

Prolines in *cis* configurations interact strongly with Phe and Trp residues in the *i*-1 position (S19 and S19), in contrast to their *trans* counterparts, through sidechain–sidechain interactions that lead to ΔRg of about (-12%) compared with sequence S1 (Figure 4(A)). When the aromatic residue is positioned at (*i* + 1) (S18 and S20), *cis* prolines can even mediate strong cross-interactions between the backbone and sidechains (even when sidechain–sidechain interactions are negligible; shown in Figure 4 inset), with such interactions leading to an elongated state characterized by ΔRg of about 5%.

A bioinformatic analysis indicates that both isolated and clustered prolines are common in natural IDPs

IDPs sequences may comprise prolines organized in different patterns, such that they may be enriched with a single type of proline organization (isolated or clustered) or with a mixture of both. Figure 5(A) shows the sequences of four natural IDPs with proline contents of 4–37% and including 0–94% clustered prolines. The IDPs CFTR and MeCP2, which contain exclusively or mostly isolated proline residues, have lower experimentally measured Rg values than are predicted by the typical scaling law for IDPs. On the contrary, the IDPs NCBD and Il1, which contain a high percentage of clustered prolines, have higher experimentally measured Rg values than are predicted by the typical-used IDP scaling law. This suggests a correlation between the pattern of prolines in the sequences and the DRg between the experimental and predicted Rg values of typical IDPs. To isolate the effects of proline, these four IDPs were selected for their low net charge content. It is expected that, for IDPs with a high net charge content, the interplay between proline and charged residues will affect overall protein conformation and Rg.[19]

To characterize the abundance of isolated and clustered prolines in a larger protein dataset, we performed a bioinformatic analysis of all intrinsically disordered regions (IDRs) in the human proteome. We found that both isolated and clustered prolines are common but show some differences, particularly for long IDRs. While the mean abundance of isolated proline content remains stable at nearly 4% for any IDR length, the abundance of clustered prolines is typically greater and it rises from about 6% up to about 10% for long IDPs of about 600 amino acids (Figure 5(B)). This suggests that the Rg of longer IDRs is more expanded than could be expected based on the IDP scaling law.

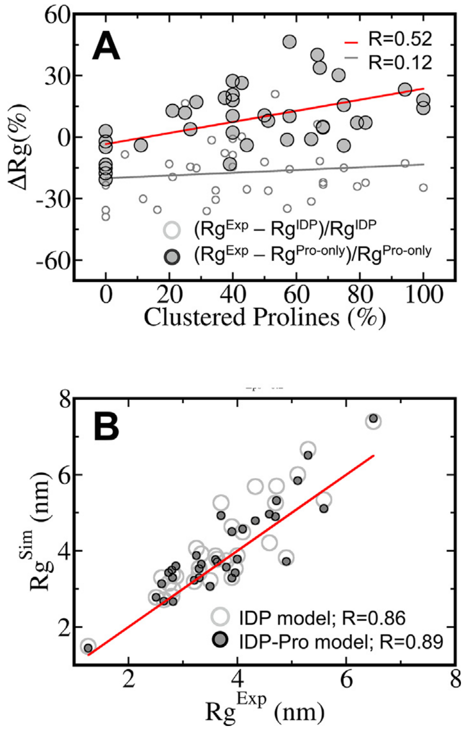


Figure 6. Radius of gyration is best predicted by proline organizational pattern. The correlation between the ΔRg values of 33 IDPs (see Table S1) and three sequence-related Rg-influencing factors. ΔRg was calculated as [Rg^{Exp}-Rg^{IDP}]/Rg^{IDP}, where Rg is the radius of gyration of the IDP backbone, Rg^{Exp} refers to experimentally-observed IDP radius of gyration values and Rg^{IDP} refers to the Rg estimated from the scaling law (filled circles) or as [Rg^{Exp}-$Rg^{Pro\text{-}only}$]/$Rg^{Pro\text{-}only}$, where $Rg^{Pro\text{-}only}$ refers to the Rg obtained from the Pro-only coarse-grained model, in which non-local residue–residue interactions were modelled as excluded interactions (empty circle). A) ΔRg versus the percentage of clustered proline residues. Prolines were considered clustered when they were separated from each other by ≤2 non-proline residues, with the percentage of clustered prolines calculated versus the total number of proline residues in the sequence. Improved correlation of the ΔRg was obtained also when the content of clustered prolines was normalized against the total number of residues in the sequence (Figure S9). B) Comparison between the simulated and experimental Rg of 33 IDPs simulated using the IDP and IDP-Pro models. The red line is the linear fit for the Rg from the IDP-Pro.

The experimental Rg values of IDPs are better explained by proline organizational pattern than by proline content

Often the total proline content in IDPs is used to predict the ΔRg between experimental and estimated values. For example, the ΔRg between experimental values and estimates based on the scaling law was shown to be correlated with both charged-residue and proline content.[52] We sought to examine whether the classification of prolines in a protein sequence as isolated or clustered could better explain the compaction and expansion behavior of IDPs than the approaches used to date. To that end, we compared the goodness of correlation between ΔRg (experimental Rg compared with the value expected from the scaling law estimate) and proline clustering (Figure 6(A) filled circles and Figure S9(A)), proline content (Figure S9(B)), and charge per residue (Figure S9(C)). We found that ΔRg correlated only weakly with percent proline content (Figure S9(B)) and net charge per residue (Figure S9(C)), with Pearson correlation coefficients or R = 0.20 and R = 0.36, respectively. However, correlation improved significantly to R = 0.52 when ΔRg was plotted against the percentage of clustered proline residues (Figure 6(A)). We also examined correlations for various definitions of clustering that differed in terms of the numbers of intervening Gly or Ser residues. The results (Figure S15) confirmed that the greatest correlation was obtained when a cluster was defined as proline residues separated by ≤2 intervening residues.

To elucidate the extent to which proline content and organization contribute to the conformational ensemble of natural IDPs, we used coarse-grained models to simulate the curated dataset of 33 natural IDPs for which experimentally determined Rg values are available (Table S1).[54,55] The coarse-grained models incorporated different parameters for isolated versus clustered prolines by specifying their Cα dihedral angles as identified in the atomistic simulations of the 11-residue peptides (S2–S11; see Methods). The strength of each dihedral angle in the coarse-grained model was calibrated by simulating the 11-residue peptides and matching their Rg values to those measured in the atomistic simulations (the detailed criteria for proline representation in the coarse-grained models and their calibration can be found in the SI and Figure S1).

The selected IDPs were simulated using two different coarse-grained models. In the Pro-only model, non-local residue–residue interactions were modeled solely as excluded interactions and ΔRg was calculated by comparing the Pro-only coarse-grained model results with the random-coil values. In The IDP-Pro model, non-local residues–residue interactions were represented by Lennard-Jones potentials with the strength represented by the mean hydrophobicity of the two interacting residues as well as by electrostatic interactions. For the IDP-Pro model, ΔRg was calculated by comparing the Rg values produced by the IDP-Pro model with the results obtained from the IDP model that lacks the effect of proline backbone.

Simulating with the Pro-only model yielded IDPs that are either more compact or more expanded than their Rg predicted when assuming that they follow polymer physics of random coil. Table 1 illustrates the effect of prolines on the backbone for several selected IDPs in comparison to their modeling as random-coil demonstrating compaction up to 4% and expansion up to 27%. When ΔRg was calculated by comparing the Pro-only model with experimental data (i.e., ΔRg= (Rg^{Exp}- $Rg^{Pro\text{-}only}$ /$Rg^{Pro\text{-}only}$), it correlated only very weakly with the percentage of clustered prolines in the sequence (Figure 6(A), empty circles), indicating that the representation of clustered and isolated prolines in the Pro-only model captures the effect of proline measured experimentally. Although the Pro-only model captures the ambivalent effect of proline depending on its pattern of organization along the sequence, it cannot predict the absolute value of Rg as it does not include any favorable residue–residue interactions for transient packing of the IDPs. Several such models that include hydrophobic and charge–charge interactions between interacting residues have been reported to successfully reproduce the experimental Rg of various IDPs.[51,56,57]

Simulating the 33 selected IDPs using the IDP-Pro model was found to introduce compaction or expansion compared with the Rg obtained from common IDP models. The Pro-only and IDP-Pro models yielded expansion or compaction effects of similar magnitude. More importantly, the Rg values of IDPs simulated with the IDP-Pro model are in better agreement with the experimental results than those obtained from the FB IDP model that lacks the information on prolines, with correlation increasing from R = 0.86 to R = 0.89 (Figure 6(B)).

Conclusions

The proline residue, being unique among the amino-acid residues, has been highly studied both experimentally and theoretically with the aim of quantifying the consequences of the cyclization of its sidechain with the backbone on conformational preferences and therefore on function. Despite extensive research, the molecular biophysics of prolines is not fully understood, particularly in IDPs, which are much more enriched in prolines than are folded proteins. The role of prolines in IDPs, therefore, may differ from its roles in folded proteins. To decipher the mechanism by which prolines modulate the IDP conformations, we designed and computationally studied a series of 11- and 21-residue peptides containing 1–3 prolines, thus corresponding to a local proline content of 10–30%. These peptides also varied with respect to proline organizational pattern, with the spacing between consecutive prolines being 0–5 residues.

We found that prolines may exert opposite effects on the conformational preferences of IDPs. Namely, prolines may impose either compaction or expansion on IDPs. This dual role of prolines on the structural properties of IDPs depends on the total proline content in the sequence and, especially, on its organizational pattern. Prolines that are clustered tend to expand IDP conformation via the formation of PPII-like secondary structural elements. The degree of expansion depends on the number of prolines that comprise the proline cluster. Prolines that are isolated may result in compacted conformations.

Table 1 Experimentally and computationally measured Rg of natural IDPs with different content and sequence pattern of proline residues.

Protein	Length (amino acids)	Total proline (%)	Clustered proline (%)	Experimental Rg (nm)	Experimental ΔRg [a] (%)	Simulated ΔRg [b] (%) (Pro-only model)	Simulated ΔRg [c] (%) (IDP-Pro model)
FHua	143	7	0	3.34	−4.8	−1.9	−6.9
CFTR	185	4.3	0	3.25	−18.9	−3.2	−4.6
Rec1 Resilin	310	13.9	16	4.33	−14.4	−4.0	−15.8
CornID	267	9.7	35	4.72	−3.3	−0.6	−6.7
MeCP2	228	10.1	39	3.7	−13	−2.0	−6.4
hNHe1cdt	131	14.5	58	36.3	8.4	2.3	−1.9
Human NCBD	105	15.2	75	3.3	10.8	5.1	3.3
p53	93	23.7	81.8	2.87	2.8	8.2	8.2
Ash1	83	14.5	91.7	2.89	9.9	8.6	8.8
Il1	143	36.4	94	4.59	30.9	19.0	17.8
IB5	72	38.9	100	2.8	14.8	27.0	24.5

[a] The deviation of the experimentally measured Rg from that calculated using a scaling law for predicting the Rg of IDPs[71]; i.e., ΔRg=(Rg^{Exp} – Rg^{IDP})/Rg^{IDP}.

[b] The deviation of the Rg using the Pro-only coarse grained model is relative to a random-coil model; i.e., ΔRg=($Rg^{Pro\text{-}only}$ – Rg^{RC})/Rg^{RC}.

[c] The deviation of the Rg using the IDP-Pro coarse-grained model is relative to an IDP model where the residue–residue interactions are modeled based on their hydrophobic and electrostatic interactions (model FB,[56] ΔRg=($Rg^{IDP\text{-}Pro}$ – Rg^{FB})/Rg^{FB}.

We found that proline should be classified as clustered if the spacing between consecutive proline residues is ≤ 2 non-proline residues. Otherwise, the proline should be classified as isolated.

The energy landscape for peptides with isolated prolines comprises two major basins that correspond to more- and less-compacted conformations, whereas the energy landscape for peptides with clustered prolines is characterized by a single basin that corresponds to the expanded conformation. The dual-basin energy landscapes for peptides with isolated prolines indicates that a conformational switch between the two conformational ensembles can be achieved even between prolines in the *trans* configuration and that its timescale is much shorter than that associated with *cis* to *trans* proline isomerization. Isolated prolines are reported to mediate a hairpin loop formation in transmembrane protein separating two transmembrane segments separated by a tight turn.[27] Furthermore, proline frequency in loops (i.e., a compact conformation) in globular proteins has been measured to be the second highest after Glycine.[26] A similar conformational transition by a single proline in the *trans* configuration located in the hinge region of a two-domain protein has been reported using atomistic molecular dynamics simulations.[21]

Whereas isolated *trans* prolines may induce compaction of about 10%, clustered *trans* prolines may expand IDPs by about 30%. The degree of expansion may be reduced depending on whether the clustered prolines are separated by one or two residues. This is consistent with reduced PPII propensity in the presence of spacing residues.[12] We found that the degree of compaction due to isolated *cis* prolines can reach about 20%, but this is achieved only when there are two clustered *cis* prolines. For other *cis* proline patterns, and particularly for isolated *cis* prolines, a more modest compaction effect of about 5% is found in our simulations.

The classification of prolines as isolated or clustered in natural IDPs explains the deviation of their experimental Rg values from the estimated Rg based on Flory scaling law of IDPs. Accordingly, IDPs that contain a high fraction of isolated prolines are often more compact than the estimated Rg value based on the scaling law whereas IDPs with a higher fraction of clustered prolines are often more expanded. The Pro-only and IDP-Pro coarse-grained molecular dynamics simulations, which accounted for the effect of proline residues as found through atomistic simulations, better captured the backbone compaction or expansion effects of proline residues compared with a model that neglected the effect of prolines on backbone conformations. The compaction or expansion observed in the coarse-grained models is consistent with that identified experimentally.

The effects exerted on the overall dimensions of IDPs by the local effects of isolated and clustered proline residues may depend on proline's interactions with other residues in the IDP sequence. Charge–charge interactions may affect the ability of prolines to induce compaction or expansion, as was shown for an 81-residue IDR from the *S. cerevisiae* transcription factor, Ash1.[19] Polyampholytic IDPs (*i.e.*, IDPs with close to zero net charge but a high fraction of positively and negatively charged residues[58,59] can enhance compaction when the IDP is rich in isolated prolines or enhance expansion when the IDP is rich in clustered prolines. Polyelectrolytic IDPs (*i.e.*, with either a positive or negative net charge) adopt expanded conformations to maximize charge separation, which may restrict the compaction effect of isolated prolines and enhance it for clustered prolines. We show that aromatic residues adjacent to isolated prolines can affect the degree of compaction induced by the proline via direct interactions, primarily between the aromatic sidechain and proline. We found evidence for this effect not only for *cis*[15] but also for *trans* proline residues. This effect is sensitive to the nature of the aromatic residue, its location, and the proline isomer. An aromatic residue at position *i*-1 relative to a trans proline is associated with reduced compaction compared to aromatic residues at position $i + 1$. An aromatic residue next to a *cis* proline exhibits different effect than that for *trans* proline as a stronger compaction effect is observed when the aromatic residue is at position $i + 1$ relative to the *cis* proline than compare to position *i*-1. These examples suggest that the effect of prolines on peptide conformations may depend on the IDP sequence complexity (e.g., pattern and content of charged and aromatic residues in its vicinity) due to either direct interactions with prolines or indirectly by biasing the local environment of the prolines. Quantification of such cross-talks between proline and other residues may refine its degree of compaction or expansion on IDP conformations.

In conclusion, proline in the *trans* configuration is identified to have an ambivalent behavior that may impose either compaction or expansion on the overall conformational ensemble of an IDP. Our simulation study using atomistic and coarse-grained models shows that this dual role depends on their organization as isolated or clustered prolines, suggesting that total proline content is not a good measure to probe the effect of proline on IDPs. Local conformations of IDPs may vary depending on whether they contain isolated or clustered prolines. IDPs that are rich in isolated prolines tend to be more compact than expected on the basis of their length. IDPs that are rich in clustered prolines, which are often found in long IDPs, tend to be more expanded than expected on the basis of their length. These new insights suggest that proline residues can tune protein conformations not only via *cis* to *trans* isomerization. Indeed, whereas the mouse CrkII kinase adopts the functional closed state via *trans* to *cis* isomerization, the human CrkII achieves the functional closed state while the proline remains in the *trans* configuration.[60,61]

Models and Simulations

Designed peptides

To understand the roles played by prolines in IDPs, we designed a series of 13 peptides (S1–S13) comprised of 11 amino acid residues, specifically, Gly, Ser, and varying amounts of proline (see Figure 1 for the exact sequences). These flexible peptides differed from each other with respect to the number (0–3) and organizational pattern (isolated or clustered) of the *trans* proline residues they contained. We also designed a second series of four 11-residue proteins (S17–S20) containing a single *trans* proline residue immediately adjacent to a single aromatic Trp or Phe residue, with the remaining sequence consisting of Gly or Ser residues as in the first series (see Figure 4 for the exact sequences). An additional series of five 21-residue peptides (S10, S11, S14-S16) containing up to 3 *trans* proline residues was designed to allow modeling of greater spacing between consecutive proline residues (see Figure 1). Overall, the designed peptide sequences included prolines spaced by 0–5 residues. For comparison, we also designed some of the sequences with proline in the *cis* configuration (Figures S10-S13).

Dataset of natural intrinsically disordered proteins

A dataset of 33 natural IDPs with experimentally determined Rg was curated (see Table S1). The selected IDPs have low net charges (measured as absolute net charge per residue) to avoid the presence of confounding charge-related effects on IDP dimensions. The IDPs differ in the proline content (i.e., the number of prolines in the sequence). These IDPs were simulated using several models (as described below) to estimate the potential improvements of common approaches to estimate IDPs ensemble by incorporation of proline effects on backbone dynamics.

Molecular dynamics simulations

Atomistic simulations. To study the effect of prolines on the conformational ensemble of the designed peptides, they were studied using atomistic molecular dynamics simulations. Each designed peptide was placed in a cubic box of length 6 nm and solvated using the extended simple point charge (SPC/E) water model. The system was energy minimized with the steepest descent algorithm and therefore equilibrated using constant number, volume and temperature (NVT) protocols with velocity rescaling to attain the desired temperature. Three independent 1 μs runs were carried out for each of the sequences. All simulations were performed with the GROMACS package v. 2020[62] employing the Charmm36m force field.[63] To ensure that the results were independent of the selected force field, we also studied the peptide series with the Amber-99-SB-ILDN force field.[64] The results presented in the main text are for simulations performed using the CHARMM36m force field whereas the results obtained using the Amber-99-SB-ILDN force field are presented in the Supporting Information. Selected sequences were also simulated with CHARMM36m force field with modified TIP3P water model to examine the robustness of our results as CHARMM36m model was calibrated with modified TIP3P water model that enhances protein-water interactions for better modeling of IDP conformational ensemble. We have observed similar results from simulations of modified TIP3P water model to that obtained using the SPC/E water model (See Figure S12). Simulations were carried out at a temperature of 300 K. Full atomistic details were retained for every molecule. The simulations were carried out with 2 fs timesteps and trajectories were saved every 10 ps. Periodic boundary condition was applied. Non-bonded interactions were calculated using a grid search for neighbors. Each sequence was capped with an acetyl group at the N-terminus and an N-methyl group at the C-terminus in order to avoid electrostatic attraction between the two termini that may affect the overall dynamics and conformation of the studied peptides.

The global effect of prolines was estimated by the radius of gyration, Rg, of each of the simulated peptides. The local effect of prolines was probed by examining changes in the various dihedral angles around four consecutive Cα atoms in the vicinity of any prolines (designated as δ angles). Namely, the dihedral angles between four consecutive Cα atoms of the sequences PXXX, XPXX, XXPX, and XXXP were probed for the designed sequences. The dihedral angles with prolines at the 2nd and 3rd angles are more sensitive to the organizational pattens of prolines in the sequences than for prolines in the 1st and 4th positions. Also, the end-to-end distances between the ends of such four consecutive Cα were measured. To estimate the effect of prolines on the conformational flexibility of the peptides, their configurational entropy was calculated. The configurational entropy (S) was estimated by focusing on the Cα backbone and was evaluated using Schlitter's formula starting from a quasi-harmonic approximation.[65–66] Accordingly, $S = \frac{k}{2}\ln(det\frac{k_BT}{h^2}\boldsymbol{M\sigma} + 1)$; where k_B is the Boltzmann constant, M is the mass matrix, 1 is the unity matrix, and $\boldsymbol{\sigma}$ corresponds to the covariance matrix of atomic fluctuations in cartesian coordinates $\sigma_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle$.

Coarse-grained simulations. To elucidate the bare effect of prolines on natural IDP sequences, we performed coarse-grained simulations of selected proteins (see SI) whose Rg values are known from experimental measurement. In the coarse-grained model, each residue is represented by a single bead and its local and non-local interactions with neighboring beads are described by the energy function V:

$$V = \sum_{i<j\le N} \frac{1}{2} k^b_{ij}(d_{ij} - d^0_{ij})^2 + \sum_{i<j<k\le N-1} \frac{1}{2} k^a_{ijk}(\theta_{ijk} - \theta^0_{ijk})^2 + \sum_{i<j<k<l\le N-2} [k^d_{ijkl}(1 + \cos(n\phi_{ijkl} - \phi^0_{ijkl})]+$$

$$\sum_{i,j\in non-contacts} 4\varepsilon_{nc}(\frac{\sigma}{r_{ij}})^{12} + \sum_{i,j\in contacts} \varepsilon_c(5(\frac{\sigma}{r_{ij}})^{12} - 6(\frac{\sigma}{r_{ij}})^{10})$$

The first two terms represent bonded interactions and angular interactions (defined between three consecutive beads) with uniform force constants, with the third term representing the simplest form of a proper dihedral (defined between four consecutive beads). The fourth term represents excluded volume interactions between all bead pairs and the fifth term represents the Lennard-Jones potential for interactions between two beads that are separated by at least three beads. The excluded volume term is applied between beads that do not interact via Lennard-Jones potential. Additional details of the coarse-grained models can be found in earlier publications.[67–69]

We performed simulations of the selected IDPs using two different types of coarse-grained models: the Pro-only and IDP-Pro. The Pro-only and IDP-Pro models represent the effect of proline on the backbone in a similar manner (as described below), but differ with respect to their modeling of non-local residue–residue interactions. In the Pro-only model, the non-local residue–residue interactions are represented solely by excluded volume interactions between beads (with the backbone modelled as a random coil (RC) chain defined solely by bonds between adjacent beads). The Pro-only model examines whether a model that accounts for proline's effects on backbone conformations yields a better estimation of IDP Rg than can be obtained from regarding the IDP as a random-coil polymer. In the IDP-Pro model, non-local residue–residue interactions are also incorporated as Lennard-Jones potentials whose strength depends on the mean hydrophobicity of the two interacting residues and on their ability to engage in electrostatic interactions.[70] For hydrophobic pairwise residue interactions, we followed three different models (FB,[56] M3[57] and HPS-Urry[51] which were shown to successfully estimate the experimental Rg of various IDPs. Thus, the IDP-Pro model aims to examine the contribution of proline's effects on backbone conformations using a more-realistic representation of IDPs and to yield a better estimation of IDP Rg than can be obtained from the Pro-only model.

In addition to performing simulations using these two coarse-grained model, a random coil (RC) model was used to simulate each IDP solely by its bonded terms (*i.e.*, without proline effects and residue–residue interactions). All three models were used to estimate the Rg values of 33 IDPs with a proline content greater than 4% for which experimental Rg values have been published.

The effects of proline residues were incorporated into the dihedral angles defined between four consecutive backbone beads (criteria shown in Figure S1). As we observed in our atomistic modeling that dihedral angles of the form XPXX, XXPX, PXXX, XXXP strongly dominate the structural features of prolines (Figures 3 and S2), we imposed three different parameter sets depending on the nature of the proline residue involved in the dihedral angle. Prolines separated by $\le$ 2 non-proline residues were defined as clustered whereas those separated by a larger spacing were defined as isolated. This threshold of 2 intervening residues to define clustered and isolated prolines is based on the difference of the potential of mean force (Figure 2). In a subsequent bioinformatic experiment (see below), we specifically explored the number of intervening residues that generally define these two terms in naturally-occurring proteins.

Since atomistic modelling (see Results) showed that clustered proline residues had larger Rg values indicating conformational extension, we set the dihedral angle XPXX of clustered IDPs (S3–S6; S10–S14) to an equilibrium value of δ_{XPXX} = 2.0 rad with a force constant of K_{XPXX} = 1.0, whereas the other dihedrals were set as follows, K_{XXPX} = 0.4 and δ_{XXPX} = 2.0 rad, K_{PXXX} = 0.4 and δ_{PXXX} = 2.0 rad, K_{XXXP} = 0.4 and δ_{XXXP} = 1.0 rad. By contrast, for isolated proline residue(s), we set the dihedral angle XPXX to δ_{XPXX} = 0.0 rad with K_{XPXX} = 0.5, whereas the other dihedral angles involving proline residues were set as follows, K_{XXPX} = 0.1 and δ_{XXPX} = 0.0 rad, K_{PXXX} = 0.1 and δ_{PXXX} = 1.0 rad and K_{XXXP} = 0.1 and δ_{XXXP} = 0.0 rad. The structural effects of separating two proline residues by two non-proline residues lies between those found for clustered versus isolated prolines. Consequently, for PXXP we imposed a parameter set combining aspects of those utilized for dihedral angles involving clustered and isolated prolines. We set K_{XPXX} = 0.7 and δ_{XPXX} = 2.0 Rad, K_{XXPX} = 0.4 and δ_{XXPX} = 2.0 Rad, and most importantly K_{PXXP} = 0.4 and δ_{PXXP} = 0.0 Rad to impose a certain mix of bent conformations. We observed excellent correlation (R = 0.98) between the Rgs from our coarse-grained model and its all-atom counterpart (Figure S1).

Structural analysis of the simulated peptides

The sampled conformations of the simulated peptides were analyzed to probe the local and global effects imposed by proline residues. The local effects of prolines were analyzed by the end-to-end distance between four consecutive residues around prolines as well as the dihedral angle, δ, defined by such quartets. The global effect of prolines on the structural ensemble was estimated by the Rg.

The Rg was calculated for the series of peptides simulated using the atomistic models (i.e., with either the Charmm36m or Amber-99-SB-ILDN force-fields). The coarse-grained models include the IDP-Pro models (where the IDP is modeled using FB, M3 or Urry potentials) or the Pro-only models. The Rg of the peptides was also predicted by typical Flory scaling law for IDPs (i.e., that does not account specifically for proline effects), where Rg = aN^{ν}; when a = 0.254, ν = 0.52 and N is the IDP chain length.[54,71] In a similar way, the predicted Rg for peptides that follow RC characteristics was estimated by applying ν = 0.66. These two estimated Rgs are referred here as Rg^{IDP} and Rg^{RC}, respectively.

To quantify the proline effect the, we compared the Rg values of constructed and natural protein sequences in various ways via the term $\Delta Rg = [Rg^1\text{-}Rg^2]/Rg^2$; where Rg^1 is the Rg of the sequence of interest and Rg^2 is a specific comparator. In the atomistic simulations of the designed sequences, the ΔRg of each sequence was compared to the Rg of sequence that lacks prolines (i.e., sequence S1). For the natural protein, the experimentally measured Rg, Rg^{Exp}, is compared to Rg^{IDP} (i.e. $\Delta Rg=(Rg^{Exp} - Rg^{IDP})/Rg^{IDP}$). The effect of prolines in the simulations using the IDP-Pro model was performed by comparing its corresponding Rg to that obtained from IDP models (i.e, FB, M3 or Urry) that lacks a direct effect of prolines (i.e., $\Delta Rg=(Rg^{IDP\text{-}Pro} - Rg^{IDP})/Rg^{IDP}$). Similarly, the Rgs from the Pro-only model was compared to that from the RC model (i.e. $\Delta Rg=(Rg^{Pro\text{-}only} - Rg^{RC})/Rg^{RC}$). To estimate the representation of prolines in the model, the $Rg^{Pro\text{-}only}$ was compared to Rg^{Exp}, which is expected to eliminate the effect of prolines from the experimentally determined Rg and therefore not to be correlated with proline content (Figure 6 (A)).

Bioinformatic analysis

To obtain a detailed view of how prolines are distributed in naturally occurring disordered sequences, we analyzed disordered regions of the human proteome, which we defined as those with an IUPred2A disorder score > 0.5. In this manner, we obtained 34,811 disordered regions. Each disordered region was analyzed to probe its proline content (total, isolated, and clustered).

CRediT authorship contribution statement

Milan Kumar Hazra: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft. **Yishai Gilron:** Data curation, Investigation, Formal analysis. **Yaakov Levy:** Conceptualization, Supervision, Writing – review & editing.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the Israeli Science Foundation (ISF center of excellence grant no. 2253/18 and 2072/22) for funding. Y. L. holds The Morton and Gladys Pickman professional chair in Structural Biology.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2023.168196.

Received 25 March 2023;
Accepted 5 July 2023;
Available online 11 July 2023

Keywords:
intrinsically disordered proteins;
proline;
radius of gyration;
cis-trans isomerization

References

1. Theillet, F.-X., Kalmar, L., Tompa, P., Han, K.-H., Selenko, P., Dunker, A.K., et al., (2013). The alphabet of intrinsic disorder. *Intrinsically Disordered Proteins* **1**, e24360 -e.
2. Uversky, V.N., (2019). Intrinsically disordered proteins and their “Mysterious” (Meta)Physics. *Front Phys.-Lausanne* **7**
3. Macarthur, M.W., Thornton, J.M., (1991). Influence of proline residues on protein conformation. *J. Mol. Biol.*, 397–412.
4. Morgan, A.A., Rubenstein, E., (2013). Proline: the distribution, frequency, positioning, and common functional roles of proline and polyproline sequences in the human proteome. *PLoS One* **8**

5. Rose, G.D., Glerasch, L.M., Smith, J.A., (1985). Turns in peptides and proteins. In: Anfinsen, C.B., Edsall, J.T., Richards, F.M. (Eds.), *Advances in Protein Chemistry*. Academic Press, pp. 1–109.
6. Aurora, R., Rose, G.D., (1998). Helix capping. *Protein Sci.* **7**, 21–38.
7. Cubellis, M.V., Caillez, F., Blundell, T.L., Lovell, S.C., (2005). Properties of polyproline II, a secondary structure element implicated in protein-protein interactions. *Proteins* **58**, 880–892.
8. Stapley, B.J., Creamer, T.P., (1999). A survey of left-handed polyproline II helices. *Protein Sci.* **8**, 587–595.
9. Perez, R.B., Tischer, A., Auton, M., Whitten, S.T., (2014). Alanine and proline content modulate global sensitivity to discrete perturbations in disordered proteins. *Proteins* **82**, 3373–3384.
10. Perez, R.B., Tischer, A., Auton, M., Whitten, S.T. 2014. In disordered proteins. 3373–84.
11. Reiersen, H., Rees, A.R., (2001). The hunchback and its neighbours: proline as an environmental modulator. *Trends Biochem. Sci* **26**, 679–684.
12. Kelly, M.A., Chellgren, B.W., Rucker, A.L., Troutman, J.M., Fried, M.G., Miller, A.F., et al., (2001). Host-Guest study of left-handed polyproline II helix formation. *Biochemistry-Us.* **40**, 14376–14383.
13. Vila, J.A., Baldoni, H.A., Ripoll, D.R., Ghosh, A., Scheraga, H.A., (2004). Polyproline II helix conformation in a proline-rich environment: A theoretical study. *Biophys. J* . **86**, 731–742.
14. Chellgren, B.W., Miller, A.F., Creamer, T.P., (2006). Evidence for polyproline II helical structure in short polyglutamine tracts. *J. Mol. Biol.* **361**, 362–371.
15. Mateos, B., Conrad-billroth, C., Schiavina, M., Beier, A., Kontaxis, G., Konrat, R., et al., (2020). The ambivalent role of proline residues in an intrinsically disordered protein : from disorder promoters to compaction facilitators. *J. Mol. Biol.* **432**, 3093–3111.
16. Alderson, T.R., Lee, J.H., Charlier, C., Ying, J., Bax, A., (2018). Propensity for cis-proline formation in unfolded proteins. *Chembiochem* **19**, 37–42.
17. Alcantara, J., Stix, R., Huang, K., Connor, A., East, R., Jaramillo-Martinez, V., et al., (2021). An unbound proline-rich signaling peptide frequently samples cis conformations in gaussian accelerated molecular dynamics simulations. *Front. Mol. Biosci.*, 8.
18. Creamer, T.P., Campbell, M.N., (2002). Determinants of the polyproline II helix from modeling studies. *Adv. Protein Chem.* **62**, 263–282.
19. Martin, E.W., Holehouse, A.S., Grace, C.R., Hughes, A., Pappu, R.V., Mittag, T., (2016). Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J. Am. Chem. Soc.* **138**, 15323–15335.
20. Andreotti, A.H., (2003). Native state proline isomerization: an intrinsic molecular switch. *Biochemistry-Us.* **42**, 9515–9524.
21. Xia, J., Levy, R.M., (2014). Molecular dynamics of the proline switch and its role in Crk signaling. *J. Phys. Chem. B* **118**, 4535–4545.
22. Sarkar, P., Saleh, T., Tzeng, S.R., Birge, R.B., Kalodimos, C.G., (2011). Structural basis for regulation of the Crk signaling protein by a proline switch. *Nat. Chem. Biol.* **7**, 51–57.
23. Sarkar, P., Reichman, C., Saleh, T., Birge, R.B., Kalodimos, C.G., (2007). Proline cis-trans isomerization controls autoinhibition of a signaling protein. *Mol. Cell* **25**, 413–426.
24. Gibbs, E.B., Lu, F., Portz, B., Fisher, M.J., Medellin, B.P., Laremore, T.N., et al., (2017). Phosphorylation induces sequence-specific conformational switches in the RNA polymerase II C-terminal domain. *Nat. Commun.* **8**, 1–11.
25. Osváth, S., Gruebele, M., (2003). Proline can have opposite effects on fast and slow protein folding phases. *Biophys. J.* **85**, 1215–1222.
26. Leszczynski, J.F., Rose, G.D., (1986). Loops in globular proteins: a novel category of secondary structure. *Science* **234**, 849–855.
27. Nilsson, I., von Heijne, G., (1998). Breaking the camel's back: proline-induced turns in a model transmembrane helix. *J. Mol. Biol.* **284**, 1185–1189.
28. Fernandez-Ballester, G., Blanes-Mira, C., Serrano, L., (2004). The tryptophan switch: changing ligand-binding specificity from Type I to Type II in SH3 domains. *J. Mol. Biol.* **335**, 619–629.
29. Crabtree, M.D., Borcherds, W., Poosapati, A., Shammas, S.L., Daughdrill, G.W., Clarke, J., (2017). Conserved helix-flanking prolines modulate intrinsically disordered protein: target affinity by altering the lifetime of the bound complex. *Biochemistry-Us.* **56**, 2379–2384.
30. Pappu, R.V., Rose, G.D., (2002). A simple model for polyproline II structure in unfolded states of alanine-based peptides. *Protein Sci.* **11**, 2437–2455.
31. Elam, W.A., Schrank, T.P., Campagnolo, A.J., Hilser, V.J., (2013). Evolutionary conservation of the polyproline II conformation surrounding intrinsically disordered phosphorylation sites. *Protein Sci.* **22**, 405–417.
32. Ahuja, P., Cantrelle, F.-x., Huvent, I., Hanoulle, X., Lopez, J., Smet, C., et al., (2016). Proline conformation in a functional tau fragment. *J. Mol. Biol.* **428**, 79–91.
33. Zhang, X., Vigers, M., McCarty, J., Rauch, J.N., Fredrickson, G.H., Wilson, M.Z., et al., (2020). The proline-rich domain promotes Tau liquid-liquid phase separation in cells. *J. Cell Biol.* **219**
34. Patriarca, E.J., Cermola, F., D'Aniello, C., Fico, A., Guardiola, O., De Cesare, D., et al., (2021). The multifaceted roles of proline in cell behavior. *Front. Cell Dev. Biol.* **9**, 728576.
35. Creamer, T.P., (1998). Left-handed polyproline II helix formation is (very) locally driven. *Proteins Struct. Funct. Genet.* **33**, 218–226.
36. Shi, Z., Woody, R.W., Kallenbach, N.R., (2002). Is polyproline II a major backbone conformation in unfolded proteins? *Adv. Protein Chem.* **62**, 163–240.
37. Stapley, B.J., Creamer, T.P., 1999. A survey of left-handed polyproline II helices.
38. Boze, H., Marlin, T., Durand, D., Perez, J., Vernhet, A., Canon, F., et al., (2010). Proline-rich salivary proteins have extended conformations. *Biophys. J.* **99**, 656–665.
39. Taler-Vercic, A., Hasanbasic, S., Berbic, S., Stoka, V., Turk, D., Zerovnik, E., (2017). Proline residues as switches in conformational changes leading to amyloid fibril formation. *Int. J. Mol. Sci.* **18**
40. Dignon, G.L., Zheng, W., Kim, Y.C., Best, R.B., Mittal, J., (2018). Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.* **14**, e1005941.

41. Wu, H., Wolynes, P.G., Papoian, G.A., (2018). AWSEM-IDP: A coarse-grained force field for intrinsically disordered proteins. *J. Phys. Chem. B* **122**, 11115–11125.

42. Benayad, Z., von Bulow, S., Stelzl, L.S., Hummer, G., (2021). Simulation of FUS protein condensates with an adapted coarse-grained model. *J. Chem. Theory Comput.* **17**, 525–537.

43. Das, S., Lin, Y.H., Vernon, R.M., Forman-Kay, J.D., Chan, H.S., (2020). Comparative roles of charge, pi, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *P. Natl. Acad. Sci. USA* **117**, 28795–28805.

44. Das, S., Amin, A.N., Lin, Y.H., Chan, H.S., (2018). Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters. *Phys. Chem. Chem. Phys.* **20**, 28558–28574.

45. Baul, U., Chakraborty, D., Mugnai, M.L., Straub, J.E., Thirumalai, D., (2019). Sequence effects on size, shape, and structural heterogeneity in intrinsically disordered proteins. *J. Phys. Chem. B* **123**, 3462–3474.

46. Cragnell, C., Rieloff, E., Skepo, M., (2018). Utilizing coarse-grained modeling and Monte Carlo simulations to evaluate the conformational ensemble of intrinsically disordered proteins and regions. *J. Mol. Biol.* **430**, 2478–2492.

47. Jephthah, S., Staby, L., Kragelund, B.B., Skepo, M., (2019). Temperature dependence of intrinsically disordered proteins in simulations: what are we missing? *J. Chem. Theory Comput.* **15**, 2672–2683.

48. Fagerberg, E., Lenton, S., Skepo, M., (2019). Evaluating models of varying complexity of crowded intrinsically disordered protein solutions against SAXS. *J. Chem. Theory Comput.* **15**, 6968–6983.

49. Shrestha, U.R., Smith, J.C., Petridis, L., (2021). Full structural ensembles of intrinsically disordered proteins from unbiased molecular dynamics simulations. *Commun. Biol.* **4**, 243.

50. Shrestha, U.R., Juneja, P., Zhang, Q., Gurumoorthy, V., Borreguero, J.M., Urban, V., et al., (2019). Generation of the configurational ensemble of an intrinsically disordered protein from unbiased molecular dynamics simulation. *P. Natl. Acad. Sci. USA* **116**, 20446–20452.

51. Regy, R.M., Thompson, J., Kim, Y.C., Mittal, J., (2021). Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. *Protein Sci.* **30**, 1371–1379.

52. Marsh, J.A., Forman-Kay, J.D., (2010). Sequence determinants of compaction in intrinsically disordered proteins. *Biophys. J.* **98**, 2383–2390.

53. Adzhubei, A.A., Sternberg, M.J.E., Makarov, A.A., (2013). Polyproline-II helix in proteins: Structure and function. *J. Mol. Biol.* **425**, 2100–2132.

54. Bernadó, P., Svergun, D.I., (2012). Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol. Biosyst.* **8**, 151–167.

55. Cordeiro, T.N., Herranz-Trillo, F., Urbanek, A., Estaña, A., Cortés, J., Sibille, N., et al., (2017). Structural characterization of highly flexible proteins by small-angle scattering. In: Chaudhuri, B., Muñoz, I.G., Qian, S., Urban, V.S. (Eds.), *Biological Small Angle Scattering: Techniques, Strategies and Tips*. Springer Singapore, Singapore, pp. 107–129.

56. Dannenhoffer-Lafage, T., Best, R.B., (2021). A data-driven hydrophobicity scale for predicting liquid-liquid phase separation of proteins. *J. Phys. Chem. B* **125**, 4046–4056.

57. Tesei, G., Schulze, T.K., Crehuet, R., Lindorff-Larsen, K., (2021). Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc. Nat. Acad. Sci.* **118** e2111696118.

58. Hazra, M.K., Levy, Y., (2020). Charge pattern affects the structure and dynamics of polyampholyte condensates. *Phys. Chem. Chem. Phys.*.

59. Hazra, M.K., Levy, Y., (2021). Biophysics of phase separation of disordered proteins is governed by balance between short- and long-range interactions. *J. Phys. Chem. B* **125**, 2202–2211.

60. Kobashigawa, Y., Sakai, M., Naito, M., Yokochi, M., Kumeta, H., Makino, Y., et al., (2007). Structural basis for the transforming activity of human cancer-related signaling adaptor protein CRK. *Nat. Struct. Mol. Biol.* **14**, 503–510.

61. Schmidpeter, P.A., Schmid, F.X., (2014). Molecular determinants of a regulatory prolyl isomerization in the signal adapter protein c-CrkII. *ACS Chem. Biol.* **9**, 1145–1152.

62. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., Berendsen, H.J., (2005). GROMACS: fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–1718.

63. Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., De Groot, B.L., et al., (2016). CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73.

64. Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J.L., Dror, R.O., et al., (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958.

65. Schlitter, J., (1993). Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett.* **215**, 617–621.

66. Andricioaei, I., Karplus, M., (2001). On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.* **115**, 6289–6292.

67. Azia, A., Levy, Y., (2009). Nonnative electrostatic interactions can modulate protein folding: molecular dynamics with a grain of salt. *J. Mol. Biol.*.

68. Givaty, O., Levy, Y., (2009). Protein sliding along DNA: dynamics and structural characterization. *J. Mol. Biol.* **385**, 1087–1097.

69. Clementi, C., Nymeyer, H., Onuchic, J.N., (2000). Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **298**, 937–953.

70. Wessen, J., Das, S., Pal, T., Chan, H.S., (2022). Analytical formulation and field-theoretic simulation of sequence-specific phase separation of protein-like heteropolymers with short- and long-spatial-range interactions. *J. Phys. Chem. B* **126**, 9222–9245.

71. Bernadó, P., Blackledge, M., (2009). A self-consistent description of the conformational behavior of chemically denatured proteins from NMR and small angle scattering. *Biophys. J.* **97**, 2839–2845.