PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Hydrodynamic Radii of Intrinsically Disordered Proteins Determined from Experimental Polyproline II Propensities

Maria E. Tomasso, Micheal J. Tarver, Deepa Devarajan, Steven T. Whitten*

Department of Chemistry and Biochemistry, Texas State University, San Marcos, Texas, United States of America

* steve.whitten@txstate.edu

OPEN ACCESS

Citation: Tomasso ME, Tarver MJ, Devarajan D, Whitten ST (2016) Hydrodynamic Radii of Intrinsically Disordered Proteins Determined from Experimental Polyproline II Propensities. PLoS Comput Biol 12(1): e1004686. doi:10.1371/journal.pcbi.1004686

Editor: Predrag Radivojac, Indiana University, UNITED STATES

Received: August 13, 2015

Accepted: December 1, 2015

Published: January 4, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award R15GM115603 and by the Division of Materials Research of the National Science Foundation under award DMR-1205670. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The properties of disordered proteins are thought to depend on intrinsic conformational propensities for polyproline II (PP_{II}) structure. While intrinsic PP_{II} propensities have been measured for the common biological amino acids in short peptides, the ability of these experimentally determined propensities to quantitatively reproduce structural behavior in intrinsically disordered proteins (IDPs) has not been established. Presented here are results from molecular simulations of disordered proteins showing that the hydrodynamic radius (R_h) can be predicted from experimental PP_{II} propensities with good agreement, even when charge-based considerations are omitted. The simulations demonstrate that R_h and chain propensity for PP_{II} structure are linked via a simple power-law scaling relationship, which was tested using the experimental R_h of 22 IDPs covering a wide range of peptide lengths, net charge, and sequence composition. Charge effects on R_h were found to be generally weak when compared to PP_{II} effects on R_h. Results from this study indicate that the hydrodynamic dimensions of IDPs are evidence of considerable sequence-dependent backbone propensities for PP_{II} structure that qualitatively, if not quantitatively, match conformational propensities measured in peptides.

Author Summary

Molecular models of disordered protein structures are needed to elucidate the functional mechanisms of intrinsically disordered proteins, a class of proteins implicated in many disease pathologies and human health issues. Several studies have measured intrinsic conformational propensities for polyproline II helix, a key structural motif of disordered proteins, in short peptides. Whether or not these experimental polyproline II propensities, which vary by amino acid type, reproduce structural behavior in intrinsically disordered proteins has yet to be demonstrated. Presented here are simulation results showing that polyproline II propensities from short peptides accurately describe sequence-dependent variability in the hydrodynamic dimensions of intrinsically disordered proteins. Good agreement was observed from a simple molecular model even when charge-based

Competing Interests: The authors have declared that no competing interests exist.

considerations were ignored, predicting that global organization of disordered protein structure is strongly dependent on intrinsic conformational propensities and, for many intrinsically disordered proteins, modulated only weakly by coulombic effects.

Introduction

Many proteins, and protein domains, that perform critical biological tasks have disordered structures under normal solution conditions [1–3]. These proteins are referred to as intrinsically disordered [4] and, accordingly, molecular models of disordered protein structures are needed to understand the physical basis for the activities [2,3], roles regulating key signaling pathways [5], and relationships to human health issues [6–9] that have been linked to intrinsically disordered proteins (IDPs).

The properties of disordered protein structures are often associated with conformational propensities for polyproline II (PP_{II}) helix [10–12] and charge-based intramolecular interactions [13–15]. PP_{II} propensities are locally-determined [16] and intrinsic to amino acid type [17–19], while charge-charge interactions seem to be important for organizing disordered structures owing to both long and short range contacts [13–15,20,21]. Since chain preferences for PP_{II} increase the hydrodynamic sizes of IDPs [22,23], and Coulombic interaction energies are distance-dependent, it could be argued that charge effects on IDP structures are modulated locally by intrinsic PP_{II} propensities. A number of issues with that hypothesis, however, are apparent. First, it has not been established if PP_{II} propensities measured in short peptide models of the unfolded states of proteins [17–19] translate to IDPs. It could be that PP_{II} propensities are negligible and unimportant in IDP systems. Second, methods capable of separating the impact of weak to possibly strong local conformational propensities and charge-charge interactions in the context of flexible and disordered protein structures have not been demonstrated, but are required for testing any potential interdependence.

To investigate such issues, a computer algorithm [22–24] based on the Hard Sphere Collision (HSC) model [25] was developed for parsing the contributions of intrinsic PP_{II} propensities and charge to the structures of IDPs, as represented by the hydrodynamic radius (R_h). A HSC model was chosen since PP_{II} propensities and charge effects could be added separately and in steps, to isolate contributions to simulated IDP structures. R_h was chosen since experimental values are available for a wide range of IDP sequences, allowing direct comparisons to model-simulated R_h.

Here we demonstrate that R_h for disordered proteins trend with chain propensities for PP_{II} structure by a simple power-law scaling relationship. Using experimental PP_{II} propensities for the common biological amino acids from Kallenbach [17], Creamer [18], and Hilser [19], this relationship was tested against experimental R_h from 22 IDPs [23,26–42] ranging in size from 73 to 260 residues and net charge from 1 to 43. We observed that the power-law scaling function was able to reproduce IDP R_h with good agreement when using propensities from Hilser, while the Kallenbach and Creamer scales consistently overestimated R_h. The ability to describe R_h from just intrinsic PP_{II} propensities associated with a sequence was supported by simulation results showing that charge effects on IDP R_h are generally weak. Relative to the effects of PP_{II} propensities, charge effects on IDP R_h were substantial only when charged side chains were separated in sequence by 2 or fewer residue positions and if the sequence had higher than typical bias for one charge type (i.e., positive or negative). Overall, these results demonstrated that two seemingly disparate experimental datasets, IDP R_h and intrinsic PP_{II} propensities, are in

qualitative agreement; providing evidence for considerable sequence-dependent conformational preferences for PP_{II} structure in the disordered states of biological proteins.

Results

Computer simulation of R_h dependence on PP_{II} propensity

R_h for IDPs are sensitive to site-specific and general structural perturbations such as amino acid substitutions [23], changes in net charge [13,14], charge rearrangements [15], and temperature changes [22,43,44]. Fig 1 shows that IDP R_h differ substantially from R_h for folded proteins [22,45,46] that have similar residue length, N. R_h from modeling proteins with no strongly preferred conformations [22], which is referred to as a random coil [47], is also provided for comparison to the experimental values. The solid line representing coil R_h was determined from computer simulation of randomly configured polypeptide chains using a HSC model [22]. Owing to favorable native contacts that promote stable globular structures, folded proteins have R_h that are compacted relative to the R_h of simulated random coils. In contrast, the data in Fig 1 indicate that R_h from IDPs are generally larger than random coil estimates.

The dependence of R_h on N for chemically denatured proteins follows a power-law scaling relationship,

$$R_h = R_o \cdot N^{\nu}, \tag{1}$$

where R_o is 2.2 Å and ν is 0.57 [45]. To understand changes in R_o and ν that are required for modeling the dependence of R_h on N for IDPs, it is useful to recognize that unfolded proteins in aqueous solutions absent high concentrations of guanidine hydrochloride or urea show R_h compaction [48] with a concomitant decrease in ν [49]. Consistent with that observation, Marsh and Forman-Kay demonstrated that R_h and N scale with $\nu = 0.509$ for IDPs under normal conditions [49]. R_o for IDPs was found to depend on PRO content and net charge by,

$$R_o = (1.24 \cdot f_{PRO} + 0.904) \cdot (0.00759 \cdot |Q| + 0.963) \cdot 2.49, \tag{2}$$

where f_{PRO} is the fractional number of PRO residues and $|Q|$ the absolute net charge determined from sequence [49]. Since PRO residues have strong propensities for PP_{II} helix, which is an extended structure [50], and repulsive interactions between charged groups likewise favor extended conformations to minimize unfavorable energetics, a simple molecular interpretation of Eq (2) can be offered whereby the R_h dependence on N for IDPs follows a baseline trend of $R_h = (2.17\ \text{Å}) \cdot N^{0.509}$ (i.e., R_o with f_{PRO} and $|Q|$ set to zero) with sequence-dependent increases in R_h from this baseline owing to chain propensities for PP_{II} and repulsive charge-charge interactions. Simulated R_h for random coils were observed to trend with N by $R_h = (2.16\ \text{Å}) \cdot N^{0.509}$ [22], supporting this hypothesis (and reproduced in Fig 1). The effects of ALA to GLY substitutions on IDP R_h also indicated that chain propensities for PP_{II} structure modulate IDP R_h and not simply PRO content [23].

To model the effects of PP_{II} propensities on coil R_h, a sampling bias for PP_{II} structure was applied to random coil simulations and the relationship between R_h, N, and fractional number of residues in the PP_{II} conformation, f_{PPII}, was determined [22,23]. This is shown in Fig 1 by stippled lines to demonstrate that increases in f_{PPII} cause increases in coil R_h. These results were generated from simulations that modeled PP_{II} bias by applying an identical sampling bias for PP_{II} structure at each residue position in a polypeptide chain and, accordingly, did not include effects that could be caused by position-specific variations in PP_{II} propensity.

To test for effects on coil R_h owing to PP_{II} propensity variations within a polypeptide chain, conformational ensembles for $N = 15$, 25, 35, 50, and 75 were generated for poly-ALA with the algorithm modified to allow position-specific sampling rates for PP_{II} structure. It was shown

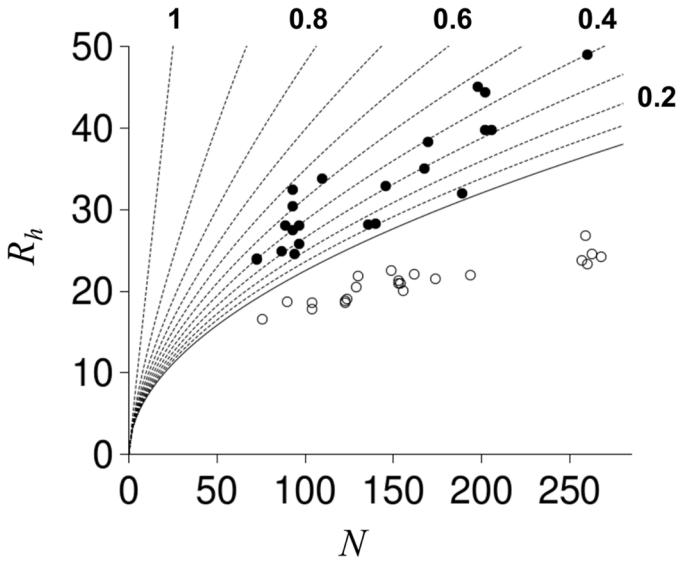

Fig 1. R_h comparison to number of residues, N. Filled and open circles represent experimental R_h for IDPs [23,26–42] and folded proteins [22,45,46], respectively. The solid line is the R_h dependence on N estimated from simulations of randomly configured protein structures [22]. Stippled lines show R_h for randomly configured structures with chain propensities for PP_{II} (f_{PPII}) from 0.1 to 1 in 0.1 increments. Every other stippled line is end-labeled by its f_{PPII} value.

doi:10.1371/journal.pcbi.1004686.g001

previously that the effects of N on R_h were mostly insensitive to amino acid sequence in HSC model simulations of disordered proteins [22] and thus poly-ALA was chosen as a computational simplification. Variations in PP_{II} propensity among residue positions were simulated by applying a sampling bias for PP_{II} structure (S_{PPII}) at every position, every second position, every third position, every fourth position, or every fifth position in the poly-ALA chains. S_{PPII} at values of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 were tested at the indicated residue locations. This PP_{II} sampling strategy resulted in 225 separate simulated ensembles (5 N lengths X 5 patterns X 9 S_{PPII} values).

A set of simulations using randomly determined position-specific bias for PP_{II} structure was also modeled using poly-ALA chains. These additional simulations used N = 15, 25, and 35, with each residue position assigned a different random value for S_{PPII}. Position-specific random assignments were repeated 3 times for S_{PPII} ranging from 0 to 1, 0 to 0.5, 0.25 to 0.75, and 0.5 to 1, resulting in an additional 36 simulated ensembles (3 N lengths X 3 distributions of random position-specific PP_{II} biases X 4 applied ranges in PP_{II} sampling bias).

The ensemble-averaged fractional number of residues in the PP_{II} conformation (i.e., the propensity) can be different from S_{PPII} in these simulations since randomly generated structures containing van der Waals contact violations are removed from the calculation. Differences between the applied sampling rate (i.e., S_{PPII}) and the observed ensemble-averaged rate (i.e., f_{PPII}) at S_{PPII}-targeted positions followed the same Gaussian relationship that was established previously for whole-chain S_{PPII} and f_{PPII} comparisons [22] and thus straight-forward conversion between applied and observed bias rates was available (S1 Fig). f_{PPII} determined from simulation for residue positions with no applied S_{PPII} was 0.012 ± 0.004.

Cumulative results from the >250 separate ensemble simulations were analyzed in terms of the power-law scaling relationship given by Eq (1). Previously, we demonstrated that the

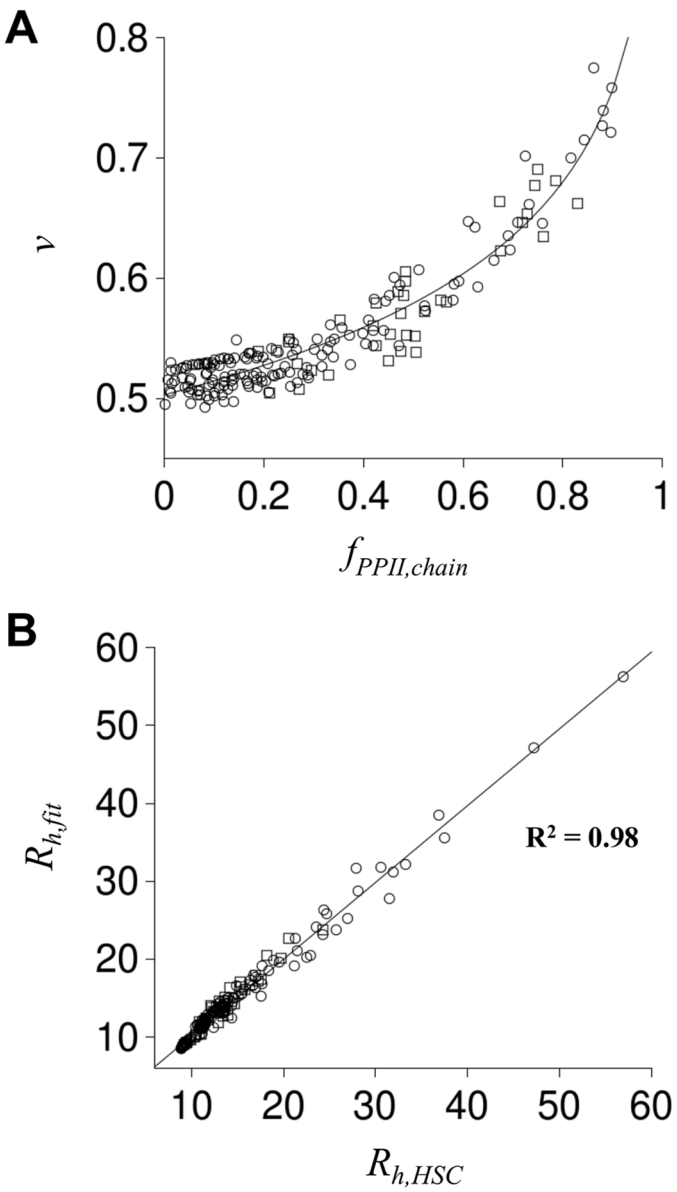

Fig 2. Simulated effect of PP_{II} propensities on coil R_h. Each circle and square represents a simulated disordered polypeptide. Squares are from ensembles simulated with position-specific PP_{II} propensities assigned randomly; circles had PP_{II} propensity assignments that followed the sequence patterns described in the text. In panel **A**, $f_{PPII,chain}$ was calculated as $\langle N_{PPII}\rangle/N$, where $\langle N_{PPII}\rangle$ was the ensemble averaged number of residues with (Φ,Ψ) in the PP_{II} region (-75±10, 145±10), and ν was calculated as $\ln(R_h/R_o)/\ln(N)$ using $\langle L\rangle/2$ for R_h and 2.16 Å for R_o. These data were fit to $\nu = \nu_o + \beta\cdot\ln(1\text{-}f_{PPII,chain})$, with ν_o and β as fit parameters, producing the solid line. In panel **B**, $R_{h,HSC}$ was calculated as $\langle L\rangle/2$. $R_{h,fit}$ was determined from $f_{PPII,chain}$ using $R_{h,fit} = (2.16\text{ Å})\cdot N^{\nu}$ and the panel A fit for ν. $R_{h,HSC}$ and $R_{h,fit}$ correlation (R^2) is provided in the figure.

doi:10.1371/journal.pcbi.1004686.g002

exponential term, ν, was dependent on S_{PPII} while R_o was mostly independent of S_{PPII} with an averaged value of 2.16 Å [22]. Fig 2A shows ν, determined from $\ln(R_h/2.16)/\ln(N)$, for each simulated ensemble and plotted as a function of f_{PPII} calculated for the whole chain. R_h for each simulated ensemble was calculated as,

$$R_h = \langle L\rangle/2, \tag{3}$$

and $f_{PPII,chain}$ as,

$$f_{PPII,chain} = \langle N_{PPII} \rangle / N \cdot \quad (4)$$

In Eq (3), $<L> = \Sigma\, L_i \cdot P_i$, where L_i is the maximum Cα-Cα distance calculated for state i, P_i is the Boltzmann probability for state i, and the summation was over all states i of an ensemble. In Eq (4), $<N_{PPII}> = \Sigma\, N_{PPII,i} \cdot P_i$, where $N_{PPII,i}$ is the number of residues in the PP_{II} conformation for state i. The distinction of "chain" given to f_{PPII} in Eq (4) was provided to limit confusion between f_{PPII} calculated for a whole chain versus f_{PPII} calculated for specific residue positions.

The relationship between ν and $f_{PPII,chain}$ for all simulations followed a logarithmic trend that was fit to the equation,

$$\nu(f_{PPII,chain}) = \nu_o + \beta \cdot \ln(1 - f_{PPII,chain}), \quad (5)$$

using the Levenberg-Marquardt method of nonlinear least squares [51,52]. The parameters ν_o and β were found to be 0.503 ± 0.002 and -0.11 ± 0.003, respectively. Fig 2B shows that R_h determined from $f_{PPII,chain}$ (Eq (4)) and N by combining Eqs (1) and (5) (see Eq (6) below) correlated strongly with R_h calculated directly from a simulated ensemble (Eq (3)). All possible patterns of position-specific PP_{II} bias were not tested in our computer trials. Results in Fig 2 predict, however, that in general a quantitative relationship exists for disordered proteins between R_h, N, and the ensemble-averaged per-residue chain propensity for PP_{II} structure ($f_{PPII,chain}$).

Test of model using experimental PP_{II} propensities

Results from HSC model simulations that are summarized in Figs 1 and 2 can be interpreted as an ideal relationship between R_h and N that includes the general effects of sterics and PP_{II} propensities but is absent other intrinsic and intramolecular factors. Contributions from Coulombic interaction energies to IDP R_h will be discussed below and added to this model. First, the simulation-derived relationship between R_h, N, and $f_{PPII,chain}$ is tested by applying experimental PP_{II} propensities to the sequences of IDPs in Fig 1. The identity, sequence, and experimental R_h for each IDP are given in Supporting Information (S1 and S2 Tables). This dataset includes 22 IDPs containing 3016 total residue positions. Amino acids represented at rates greater than 0.05 in this dataset were, in rank order and listed by their three letter codes, SER (0.104), GLU (0.100), LEU (0.083), PRO (0.080), ASP (0.074), GLY (0.073), ALA (0.073), THR (0.061), LYS (0.055), GLN (0.053), and VAL (0.053).

Amino acid PP_{II} propensities reported by Kallenbach [17], Creamer [18], and Hilser [19] for disordered proteins are reproduced in Table 1 and were used for testing the relationship,

$$R_h = 2.16 \cdot N^{0.503 - 0.11 \cdot \ln(1 - f_{PPII,chain})}\,. \quad (6)$$

These propensity scales were chosen since weak correlations are observed among the group (S2 Fig), indicating a potential for yielding different results when each set is used separately with Eq (6) for a given IDP sequence. A physical explanation for the different PP_{II} propensity values reported for the amino acids is not given here (e.g., the reported ALA PP_{II} propensities are very different when compared), other than to note that their measurements used host peptide sequences that were also very different (Table 1). Kallenbach measured PP_{II} propensities in the background of a GLY-rich host peptide, whereas the scale reported by Creamer was determined for positions flanked on both sides by PRO residues. The propensity scale from

Table 1. Intrinsic backbone PP_{II} propensities measured in disordered peptides.

host[a]	Kallenbach [17] Ac-G_2XG_2-NH_2	Creamer [18] Ac-P_3XP_3GY-NH_2	Hilser [19] Ac-$VP_2XVP_2R_3Y$-NH_2
ALA (A)	0.818	0.61	0.37
CYS (C)	0.557	0.55	0.25
ASP (D)	0.552	0.63	0.30
GLU (E)	0.684	0.61	0.42
PHE (F)	0.639	0.58	0.17
GLY (G)	-	0.58	0.13
HIS (H)	0.428	0.55	0.20
ILE (I)	0.519	0.50	0.39
LYS (K)	0.581	0.59	0.56
LEU (L)	0.574	0.58	0.24
MET (M)	0.498	0.55	0.36
ASN (N)	0.667	0.55	0.27
PRO (P)	-	0.67	1.00
GLN (Q)	0.654	0.66	0.53
ARG (R)	0.638	0.61	0.38
SER (S)	0.774	0.58	0.24
THR (T)	0.553	0.53	0.32
VAL (V)	0.743	0.49	0.39
TRP (W)	0.764	-	0.25
TYR (Y)	0.630	-	0.25
average	0.626	0.58	0.35

[a]sequence of host peptide used to measure PP_{II} propensity at the guest position, X

doi:10.1371/journal.pcbi.1004686.t001

Hilser was measured for positions located in between PRO and valine (VAL). Other PP_{II} propensity scales were not included in these tests due to similarities to the Kallenbach, Creamer, or Hilser reported values. For example, a PP_{II} propensity scale from Zondlo [53] correlated with the Creamer values (coefficient of determination, R^2, gave a correlation of 0.58), likely owing to the use of a host peptide that also flanked the guest position with PRO residues.

Inspection of Table 1 shows that PP_{II} propensities for tryptophan (TRP) and tyrosine (TYR) were not reported by Creamer. For these amino acids, we used the averaged Creamer-reported value calculated from the 18 other amino acids (0.58). In the Hilser set, TRP and TYR had lower than average PP_{II} propensity. In contrast, TRP and TYR had higher than average PP_{II} propensity in the Kallenbach set. Using the Creamer average was a compromise that likely had low significance in our tests since TRP and TYR had very low representation among the IDP sequences; 0.008 and 0.012, respectively. PP_{II} propensities were not reported for PRO and GLY by Kallenbach. Here, we used 1 for PRO since it is generally accepted that PRO has the highest propensity for PP_{II} structure [10,12,17–19]. This gave PRO a larger value than ALA (0.818), which was the amino acid with the highest reported propensity in the Kallenbach set. GLY was given a propensity of 0.50, which is lower than the Kallenbach average (0.626) but higher than the lowest value (0.428). This also was a compromise from observing that GLY had the lowest value in the Hilser set (0.13), but an average value in the Creamer set (0.58).

$f_{PPII,chain}$ was calculated for each IDP by using the amino acid PP_{II} propensity given in Table 1, summing over the IDP sequence, and dividing by N. Fig 3A shows the experimental scales predict different chain propensities for PP_{II} structure for each IDP sequence. The scale from Kallenbach gave $f_{PPII,chain}$ ranging from 0.746 to 0.628, whereas the Creamer and Hilser

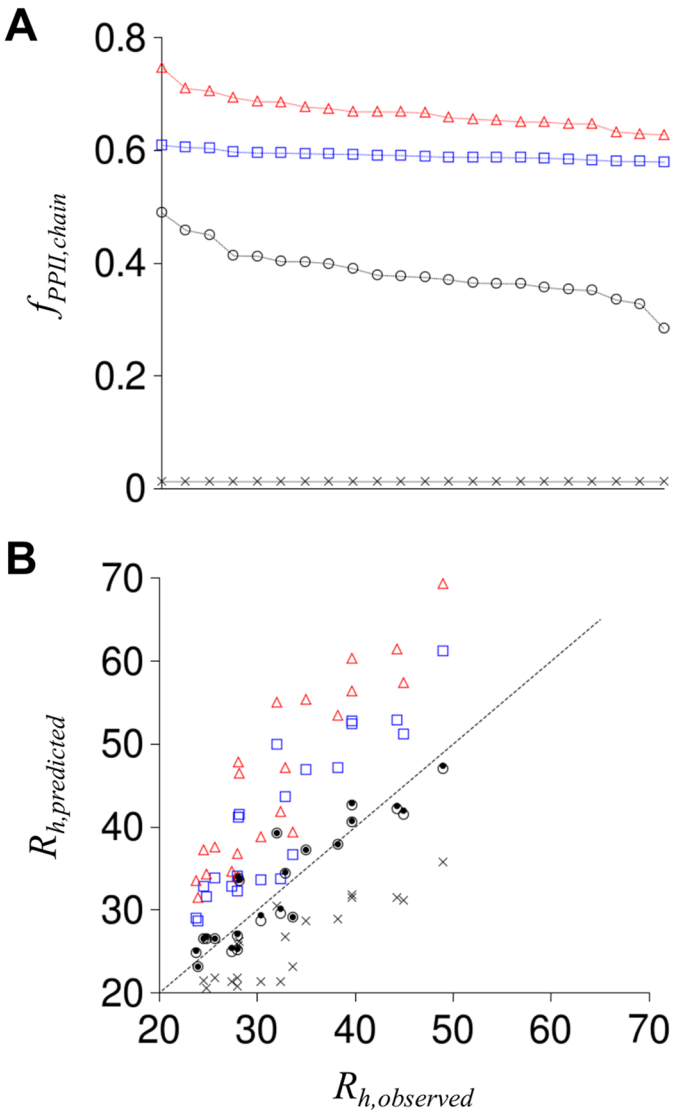

Fig 3. Chain propensity for PP_{II} from experimental scales and comparison of predicted and observed R_h. Panel **A** gives $f_{PPII,chain}$ for each IDP sequence, ordered left to right to show the range obtained with each scale, calculated using experimental PP_{II} propensities from Kallenbach (red triangles), Creamer (blue squares), and Hilser (open circles). X is $f_{PPII,chain}$ from the null model. Panel **B** shows R_h predicted for each IDP using Eq (6) and $f_{PPII,chain}$ from panel A. Symbols in panel B match panel A representations. Black dots show R_h predicted from the composite propensity scale. Stippled line is the identity line.

doi:10.1371/journal.pcbi.1004686.g003

scales gave $f_{PPII,chain}$ from 0.609 to 0.579 and 0.489 to 0.283, respectively. Eq (6) was then used to predict R_h from $f_{PPII,chain}$ for comparison to experimentally observed R_h, which is shown in Fig 3B. The average prediction error ($|R_{h,predicted}–R_{h,observed}|$) and the correlation between predicted and observed R_h is given in Table 2. To assess contributions from the amino acid scales for predicting R_h, a null model was included by assigning each amino acid the PP_{II} propensity of 0.012, the background f_{PPII} calculated from HSC simulations when no sampling bias for PP_{II} structure was applied (i.e., $S_{PPII} = 0$). Accordingly, the null model represents random coil values.

Different values of $f_{PPII,chain}$ predict different R_h for a given IDP sequence, as expected from Eq (6). For example, the null model, which used the smallest $f_{PPII,chain}$ values, predict R_h that

Table 2. Comparison of predicted and observed R_h.

Propensity Scale	Average Error (Å)[a]	R^2 [b]	Average Normalized Error[c]	R^2 [d]
Null (random coil)	7.1 ± 3.7	0.797	-0.28 ± 0.13	0.265
Kallenbach	13.4 ± 5.4	0.819	0.51 ± 0.15	0.301
Creamer	8.4 ± 4.3	0.817	0.32 ± 0.13	0.297
Hilser	2.5 ± 1.8	0.825	0.006 ± 0.12	0.407
Composite	2.4 ± 1.8	0.834	0.015 ± 0.12	0.423
Static	2.6 ± 2.0	0.799	-0.016 ± 0.13	0.291

[a]determined from |predicted R_h—observed R_h|

[b]coefficient of determination, correlation of predicted R_h and observed R_h

[c]determined from (predicted R_h—observed R_h)/(random coil R_h)

[d]coefficient of determination, correlation of normalized error and net charge density

doi:10.1371/journal.pcbi.1004686.t002

are smaller than observed for each IDP. In contrast, PP_{II} propensities from Kallenbach and Creamer, which report relatively large $f_{PPII,chain}$ values, predict R_h that are larger than observed for each IDP. Experimental propensities from Hilser predict R_h that trend with the identity line, showing good agreement, but also showing scatter relative to that line (average error was 2.5 Å). In an attempt to reduce prediction error, a composite PP_{II} propensity scale that used the Hilser values by default but the Kallenbach values for residues located between GLY (i.e., GLY-X-GLY) and Creamer values for residues located between PRO (i.e., PRO-X-PRO) was tested. This context-specific composite propensity scale (identified as "Composite" in Table 2 and Fig 3B) caused only small changes in predicted R_h, with no significant improvement in prediction capabilities relative to using only the Hilser reported PP_{II} propensities.

Since R_h increases with N (Fig 1), prediction error was normalized for peptide length by,

$$\text{normalized error} = (\text{predicted } R_h - \text{observed } R_h)/(\text{random coil } R_h) \cdot \quad (7)$$

Random coil R_h was calculated using Eq (6) with $f_{PPII,chain} = 0.012$, the null model value. Average normalized error is given in Table 2 for each propensity scale. Fig 4 shows trends in the normalized error with N and net charge density, determined as the absolute net charge normalized for peptide length,

$$\text{net charge density} = |Q|/(\text{random coil } R_h) \cdot \quad (8)$$

S1 Table gives net charge and N for each IDP. No obvious bias with peptide length (i.e., N) was observed in the normalized error for the Hilser and composite propensity scales. Normalized error clearly increased with N when using Kallenbach and Creamer values, indicating that these PP_{II} propensities may be over-estimated when applied to IDP sequences to predict R_h. Since the exponent in Eq (6) becomes larger with increasing $f_{PPII,chain}$, a set of propensity values that systematically are too large would cause normalized errors that increase with N.

It is interesting to note that normalized error correlated with net charge density for each experimental propensity scale (Fig 4B and Table 2), suggesting that prediction error was caused partially by charge effects on R_h that were not included in the model. This is not surprising since Marsh and Forman-Kay demonstrated that increases in net charge correlate with increases in IDP R_h [49] and the trend we observed of decreasing normalized error with increased net charge density is consistent with their conclusions. Extrapolating this trend to zero net charge density for the Hilser and composite propensity scales yields positive

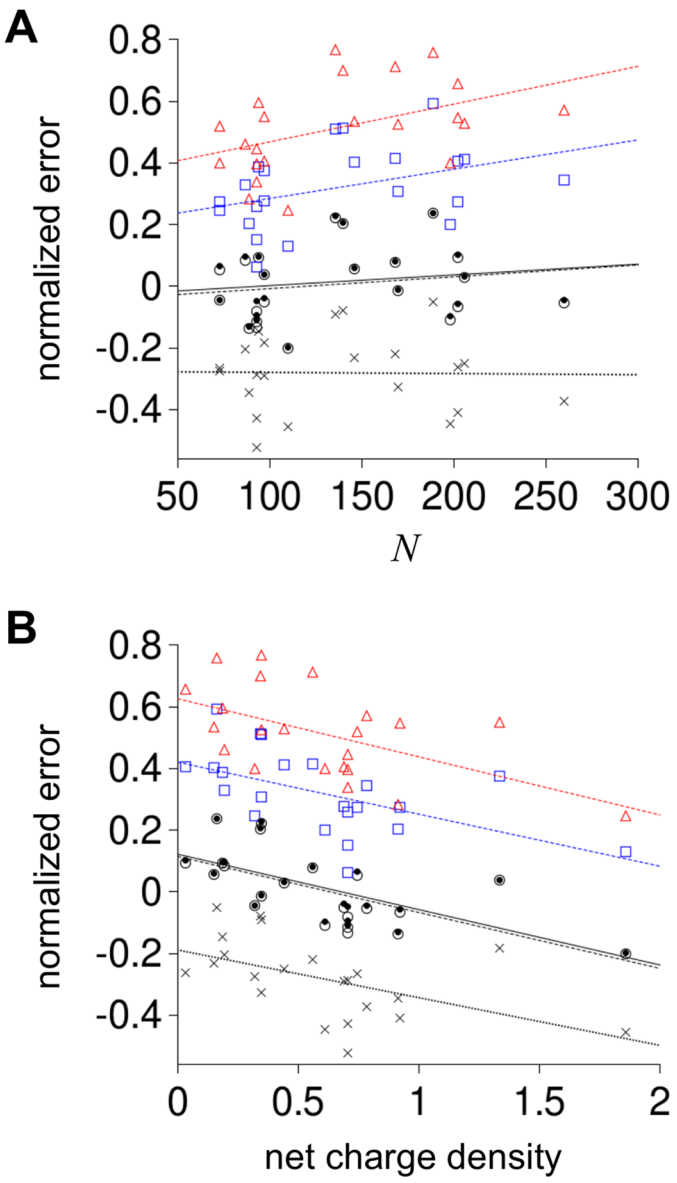

Fig 4. Correlation of normalized error in predicted R_h to N and net charge density. Normalized error and net charge density were calculated for each IDP using Eqs (7) and (8), respectively. In both panels, red triangles show normalized error from R_h predicted using the Kallenbach reported propensities, blue squares from Creamer reported propensities, open circles from Hilser reported propensities, black dots from the composite propensity scale, and X is the null model. Lines are linear fits to the five prediction sets colored as the symbols (Kallenbach scale was red; Creamer was blue, Hilser was stippled black, composite was solid black, and null was dotted black).

doi:10.1371/journal.pcbi.1004686.g004

normalized errors suggesting that, in the background of no net charge contributions to R_h, the PP_{II} propensities reported by Hilser may also be slightly too large when using Eq (6) to predict R_h.

While this analysis of experimental PP_{II} propensities indicated that one of the scales was capable of reproducing experimental R_h with good agreement for a set of IDPs, it is important to recognize that comparative tests based on Eq (6) may not be suitable for affirmation. Since R_h in this model depends only on N and chain averaged propensity for PP_{II} structure, contrived

scales that predict IDP R_h with similar agreement in terms of the average prediction error are simple to generate. For example, each IDP could be given a sequence-independent $f_{PPII,chain}$ value of 0.364, which was determined by converting experimental R_h to an apparent $f_{PPII,chain}$ using Eq (6) and then averaging over the IDP dataset. Using this static $f_{PPII,chain}$ to predict IDP R_h gives an average prediction error (identified as "Static" in Table 2) that is close to the error obtained when using the experimental scale from Hilser. Correlations between predicted and observed R_h and between normalized error and net charge density for the contrived static scale, however, decreased relative to the correlations that were observed with the experimental scales, suggesting that static representations of $f_{PPII,chain}$ may not fully capture some molecular dependencies that are inherent to IDP R_h.

To further investigate the capabilities of Eq (6) for relating IDP R_h and PP_{II} propensity, random sets of amino acid scales were generated following a two-step protocol and analyzed. First, a random number between 0 and 1 was used to target an average propensity for a scale. Then, random scales were generated, where each amino acid was assigned a different random value between 0 and 1, until a set was found whose average for the 20 amino acids matched the target determined in the first step (±0.05). The goal from using two steps to generate scales was to ensure that chain averaged propensities in the high, medium, and low range were evenly sampled. This sampling scheme was repeated until 100,000 random scales were generated. Each propensity scale was then used to predict R_h from Eq (6) and the results are summarized in Fig 5. It was observed that randomly generated scales gave average prediction errors for the IDP dataset ranging from 1.9 to 239.8 Å, correlations between predicted and observed R_h ranging from 0.02 to 0.88, and correlations between normalized error and net charge density from 0 to 0.81. Optimal values for these metrics (i.e., highest correlations coupled with lowest average error), seem to focus toward values of R^2 and average error that are obtained when using experimental PP_{II} propensities from Hilser. This result shows that experimental R_h of the IDP dataset are in good qualitative agreement with experimental PP_{II} propensities reported by Hilser, and vice versa, giving evidence that the molecular properties of IDPs that link R_h, N, and $f_{PPII,chain}$ are well-approximated by the simple power-law scaling relationship of Eq (6).

Effects of Coulombic interaction energies on R_h

In the HSC model used for this study, a computer algorithm generates polypeptide structures by random conformational search until R_h (Eq (3)) converges to a stable ensemble-averaged value [22]. A structure-based energy function parameterized to solvent-accessible surface areas that has been tested extensively [54–62] is used to population-weight each randomly generated structure. To approximate charge effects on ensemble populations, the energy function was modified to include Coulombic interaction energies by,

$$\Delta G_{Coulomb} = \frac{332}{D_{H2O} \cdot 2} \cdot \sum_i Z_i \cdot \left(\sum_j \frac{Z_j}{R_{ij} \cdot e^{\kappa \cdot R_{ij}}} \right), \quad (9)$$

where the constant 332 converts the energy into units of kilocalories per mole at 25°C, D_{H2O} is the dielectric of water, Z is the charge at site i or j, R_{ij} is the distance between two charged sites i and j (in Å), κ (the Debye parameter) accounts for screening from solution ionic strength, and the sums are over all charge-bearing sites. The Debye parameter was calculated as,

$$\kappa = 2.913 \cdot \sqrt{I/D_{H2O}}, \quad (10)$$

where I is ionic strength (in molarity, M). D_{H2O} used was 78.3 [63] and I was 0.1 M to represent normal conditions. Since the simulations used poly-ALA chains, charged residues were

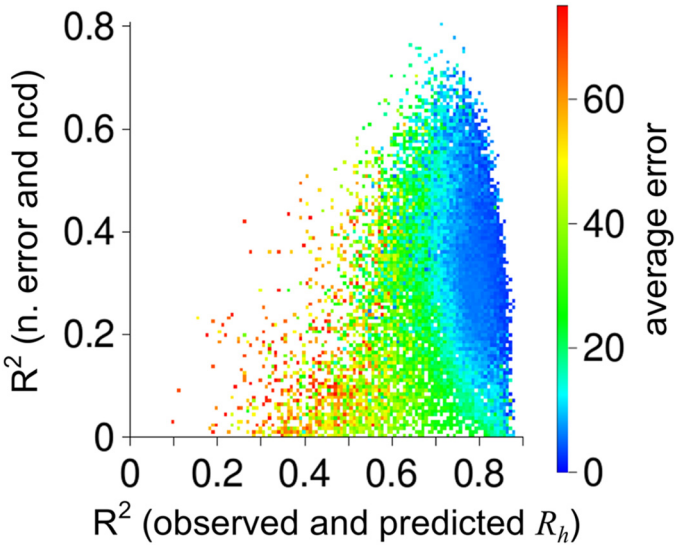

Fig 5. R_h prediction from random PP_{II} propensity scales. Random scales were generated as described in the text and used to predict R_h for each IDP by Eq (6). Shown is the correlation (R^2) obtained for each scale between observed and predicted R_h plotted against the correlation obtained between the normalized error (n. error) and the net charge density (ncd). Shown by color is the average prediction error of each scale. Random scales giving average prediction error larger than 75 Å were omitted to emphasize differences at lower error values.

doi:10.1371/journal.pcbi.1004686.g005

modeled with a positive or negative charge located at the coordinates of the Cβ atom to denote the approximate location for flexible and charged side chains. Coordinates for the backbone N and O atoms of the first and last residues were used to assign positive and negative charge, respectively, to N- and C-termini. Simulations were limited to 25 residue poly-ALA chains to establish trends for the effects of charge on R_h in this model. For each ensemble, an identical S_{PPII} was applied at each residue position. S_{PPII} was varied among the different simulations to target ensemble-averaged $f_{PPII,chain}$ ranging from 0.1 to 0.92.

Fig 6A shows that introducing charge at N- and C-termini had no effect on simulated R_h for poly-ALA chains. Modeling negative charge at the Cβ position of each residue, or positive charge (S3 Fig), caused large increases in R_h from repulsive electrostatic intramolecular interactions. Identical charge at every other residue position caused smaller increases in R_h, while identical charge at every third position gave R_h that were mostly similar to R_h of poly-ALA modeled with no charges. These data predict that the effects of charge on IDP R_h should weaken as charged residues separate in sequence, as expected. Fig 6B shows the ensemble-averaged distance between "charged" Cβ atoms that were closest in sequence for each ensemble in panel A, indicating repulsive charge-charge interactions at distances $\geq$9 Å had only minor effects on R_h. The Debye length for the modeled conditions (i.e., 1/κ) was 9.6 Å, which is the distance where interactions between charged groups become negligible at a given ionic strength. The simulation results thus trend with expected outcomes for fully solvated charges. It was also observed that, for polypeptides with each residue position charged, $f_{PPII,chain}$ calculated for an ensemble was larger than expected based upon the applied S_{PPII} (Fig 6A inset). This result predicts that repulsive charge-charge interactions between side chain groups preferentially select for the extended PP_{II} structure to minimize unfavorable interaction energies.

To test the effects of clusters of charge on R_h, polypeptides with patterns consisting of three consecutively charged residues were also simulated (Fig 7). Similar trends were observed,

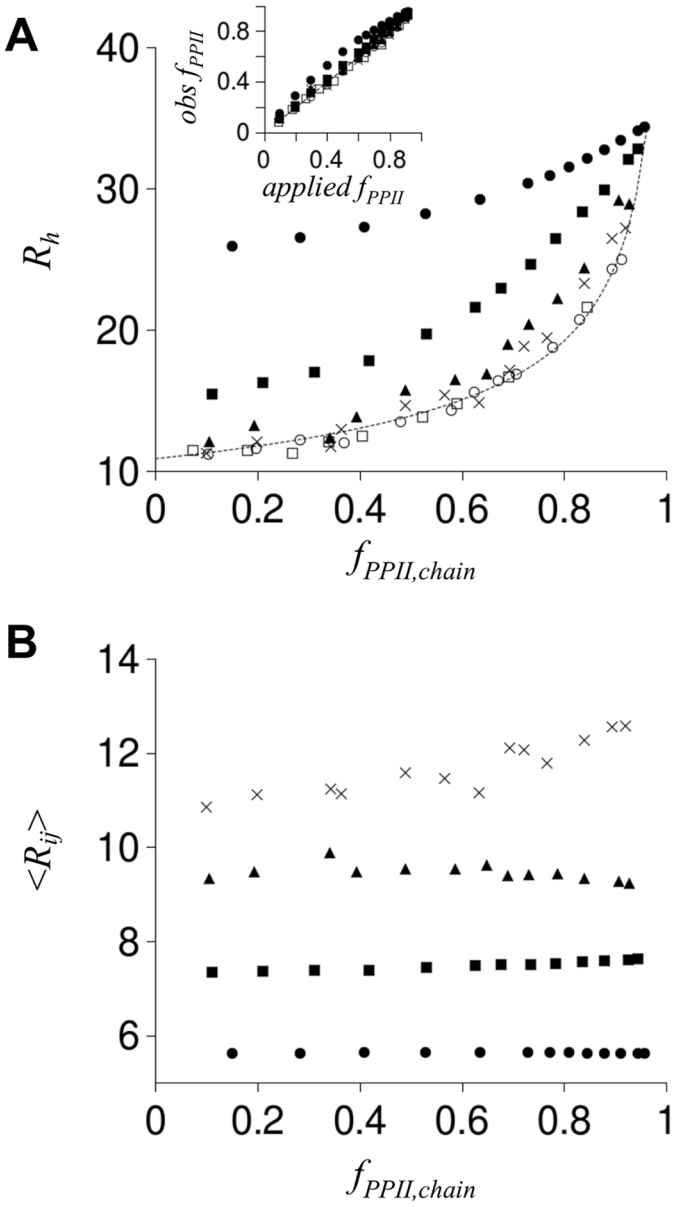

Fig 6. Simulated effect of charged residues on R_h. In panel **A**, the stippled line is R_h from Eq (6) with $N = 25$ and $f_{PPII,chain} = 0–0.98$. Plotted symbols are R_h from poly-ALA simulations ($N = 25$) calculated using Eq (3). Open squares are uncharged poly-ALA and open circles have charged termini. Filled circles have each residue modeled with negative charge at the Cβ atom. Filled squares have every other residue modeled with negative charge, filled triangles have every third residue with negative charge, and X is every fourth residue with negative charge. In panel **B**, $\langle R_{ij} \rangle$ is the ensemble averaged distance (in Å) between Cβ atoms from two charged residues, *i* and *j*, closest in sequence. Panel B symbols match panel A representations. **A inset:** comparison of observed $f_{PPII,chain}$ (shown as *obs* f_{PPII}) to $f_{PPII,chain}$ expected from the applied S_{PPII} (shown as *applied* f_{PPII}; calculated as $f_{PPII} = S_{PPII} - 0.062 \cdot \exp(-(S_{PPII}-0.63)^2/(2 \cdot 0.28^2))$ [22]. Note that filled circles trend higher than other plotted data. Inset symbols match panel representations.

doi:10.1371/journal.pcbi.1004686.g006

whereby the effects of charge on R_h weaken as charged groups (i.e., clusters) were separated in sequence. Charge clusters, however, affected R_h when modeled with 4 intervening non-charged residues, with weaker effects persisting at even larger separation distances between the clusters. This contrasts with the simulation results for non-clustered charged residues that exhibited

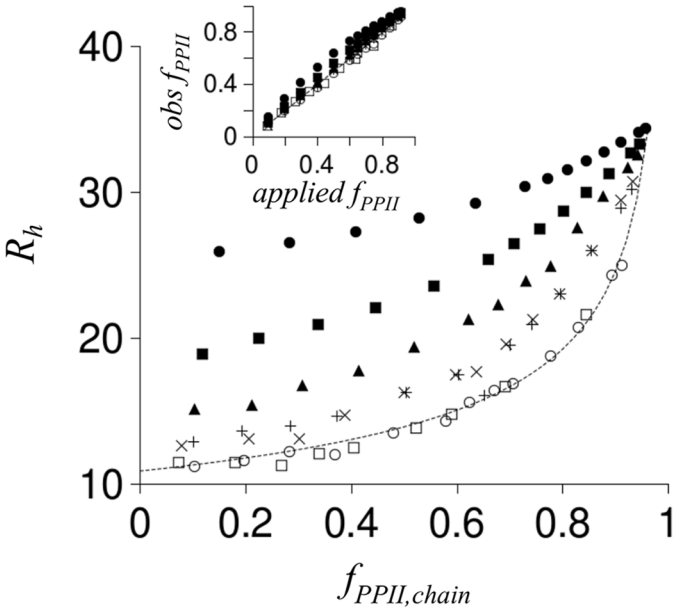

Fig 7. Simulated effect of clusters of charged residues on R_h. Filled circles, open circles, open squares, and the stippled line were reproduced from Fig 6A. As in Fig 6A, R_h was calculated from poly-ALA simulations with $N = 25$. A charge cluster was defined as three consecutive residues with negative charge modeled at the Cβ atoms. Charge clusters separated in sequence by two uncharged residues (no charge modeled at Cβ) are shown with filled squares whereas charge clusters separated by four uncharged residues are shown with filled triangles. X and + symbols represent charge clusters separated by six and eight uncharged residues, respectively. **Inset:** comparison of observed $f_{PPII,chain}$ to $f_{PPII,chain}$ expected from the applied S_{PPII} (following Fig 6A inset description). Inset symbols match panel representations.

doi:10.1371/journal.pcbi.1004686.g007

negligible effects on R_h when charges were separated by as little as 2 intervening uncharged residue positions (Fig 6A).

Since IDPs, in general, contain both positive and negative charges, simulations with opposite charge at adjacent residue positions were also performed. Fig 8A shows that repeating patterns of opposite charge had minimal effects on R_h in these simulations, even when each residue position was charged. This was mostly the case for charge clusters too (Fig 8B) with the exception that the simulation would sporadically generate ensembles with compacted R_h, whereby "compacted" is used to indicate R_h smaller than what was observed for non-charged poly-ALA coils of identical N. Overall, the amount of R_h compaction owing to favorable interactions between oppositely charged residues (or clusters) was small when compared to increases in R_h that were observed owing to unfavorable interactions between identically charged residues (or clusters).

The results in Figs 6–8 from modeling charge effects on R_h indicate that, in general, the strongest effects on R_h should occur owing to identical charges at sequentially-adjacent residue positions (Figs 6 and 7) and for polypeptides with the least amount of mixing of positive and negative charge types (Fig 8). To test these two general observations, the IDP dataset was analyzed to determine the net number of adjacent charges in each IDP sequence. This was calculated by first summing the number of ASP residues that had GLU or ASP immediately next or prior in sequence with the number of GLU residues that had GLU or ASP immediately next or prior in sequence to determine the total number of negative charges with an adjacent negatively charged neighbor. A similar calculation was performed using LYS and ARG to determine the number of positive charges with an adjacent positively charged neighbor. The net number of adjacent charges for an IDP was then the absolute value in the difference between the positive

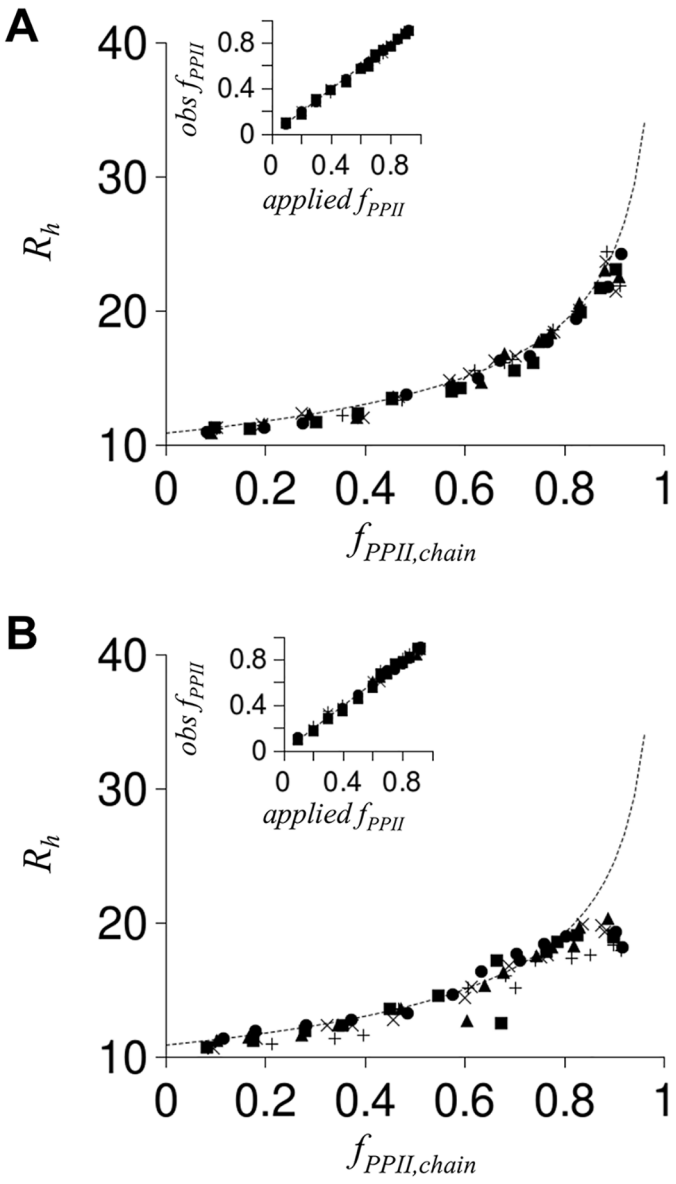

Fig 8. Simulated effect on R_h from oppositely charged residues. Stippled line in each panel was reproduced from Fig 6A. As in Fig 6A, R_h was calculated from poly-ALA simulations with $N = 25$. Charge was modeled with opposite charge at adjacent residue positions (panel A) or adjacent clusters (panel B). In panel **A**, filled circles have each residue modeled with charge at the Cβ atom (first residue negative, second residue positive, third residue negative, etc.). Filled squares have every other residue modeled with charge (first residue negative, third residue positive, etc.), filled triangles have every third residue modeled with charge, and X represents every fourth residue modeled with charge. In panel **B**, each residue in a cluster had identical charge while clusters adjacent in sequence had opposite charge. Filled circles are poly-ALA with every residue charged (i.e., residues 1–3 having negative charge, residues 4–6 with positive charge, residues 7–9 with negative charge, etc.). Charge clusters separated in sequence by two uncharged residues are shown with filled squares (i.e., residue 1–3 with negative charge, residues 4–5 uncharged, residues 6–8 with positive charge, etc.) whereas charge clusters separated by four uncharged residues are shown by filled triangles. X and + symbols represent charge clusters separated by six and eight uncharged residues, respectively. **Insets:** comparison of observed $f_{PPII,chain}$ to $f_{PPII,chain}$ expected from the applied S_{PPII} (following Fig 6A inset description). Inset symbols match panel representations.

doi:10.1371/journal.pcbi.1004686.g008

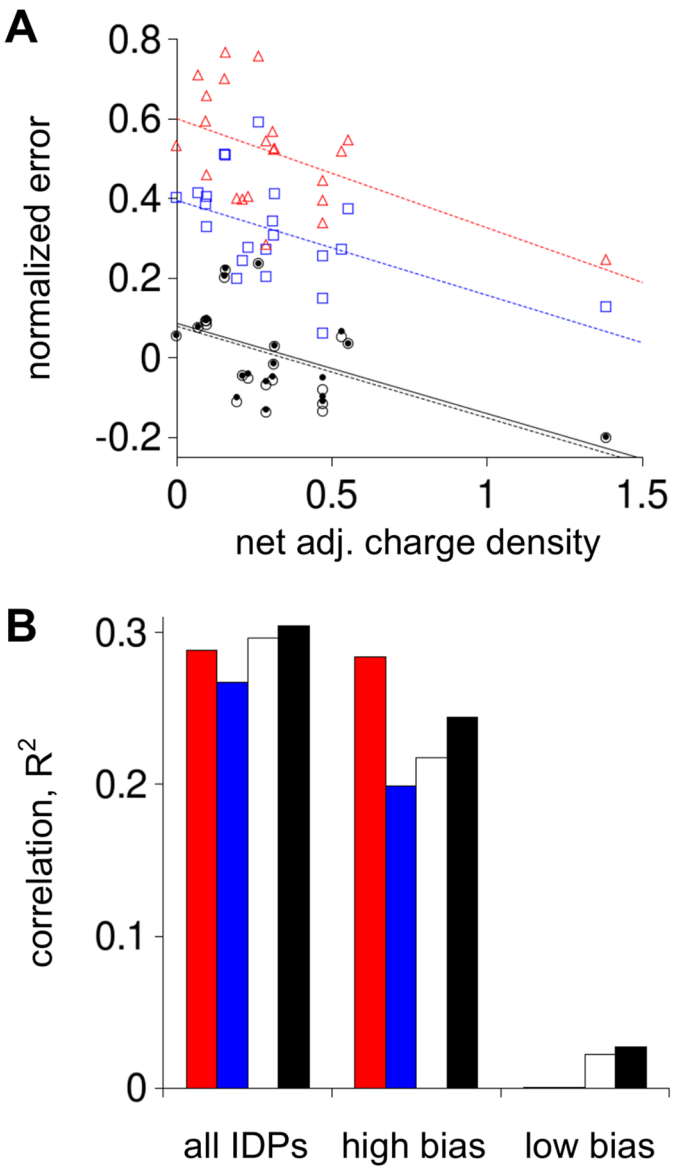

Fig 9. Correlation of normalized error in predicted R_h to net adjacent charge density. Panel **A** symbols and lines match their Fig 4 representations. Panel **B** shows correlations (R^2) between normalized error and net adjacent charge density for all IDPs, IDPs in the high charge bias group (labeled as "high bias"), and IDPs in the low charge bias group (labeled as "low bias"). Red columns are correlations from using the Kallenbach propensity scale to predict R_h, blue from using the Creamer propensities, white the Hilser propensities, and black the composite propensity scale.

doi:10.1371/journal.pcbi.1004686.g009

and negative adjacent charge numbers (provided in S1 Table). Fig 9A shows that normalized error in predicted R_h for the IDP dataset trends with the net adjacent charge density (i.e., net adjacent charge normalized for peptide length), similar to the correlation that was observed between normalized error and net charge density (Fig 4B). This should be expected since net charge and net adjacent charge correlate with $R^2 = 0.64$ in the dataset.

The set of IDPs was also split according to the amount of mixing of positive and negative charge types in a given sequence. To do this, a "charge bias" was calculated for each IDP as the simple ratio of total negative charges (sum of ASP and GLU residues) to total positive charges

(sum of LYS and ARG residues), or vice versa, depending on which ratio gave a value greater than 1. As a metric for separating IDPs with "high" and "low" charge bias, a "typical" charge bias was calculated for the entire dataset by the concatenated sequence and found to be 1.9. The average IDP charge bias, found to be 4.2, was not used to separate IDPs since: 1) ratio-based distributions are skewed, 2) only 7 IDPs would have been in the "high" charge bias set, and 3) 4 of these 7 were sequences derived from the p53 protein. Using the charge bias of the concatenated sequence gave 12 IDPs in the high charge bias set and 10 IDPs in the low charge bias set.

Fig 9B shows that correlations between net adjacent charge density and normalized error in predicted R_h persisted in the set of IDPs with high charge bias and mostly disappeared for IDPs with low charge bias, seeming to agree with the simulation prediction that significant mixing of positive and negative charge types in a sequence should reduce charge effects on R_h. Applying this analysis to net charge density gave different results (S4 Fig). Correlations between net charge density and normalized error in predicted R_h decreased for both the high and low charge bias sets. This could be owing to trends shown in Fig 6, whereby net charge effects on R_h depended strongly on the distance between the charged groups. Overall, these results seem to indicate that charge effects on IDP structures are highly dependent on sequence, however, charge effects on R_h can be weakened substantially by mixing negative and positive charge types or by slight increases in the distances between charged groups in sequence. The hypothesis that charge effects on R_h may be generally weak for IDPs is supported by data in Fig 3B showing that R_h could be predicted without specific consideration of charges when provided an appropriate amino acid scale for intrinsic PP_{II} propensities.

Discussion

Fig 1 shows that experimental R_h for IDPs are much larger than computational predictions based on random coil modeling of the R_h dependence on N. Numerous studies have demonstrated the importance of Coulombic effects for regulating IDP structural preferences [13–15]. Thus, it could be surprising to note that sequence effects on IDP R_h can be predicted with good agreement from sequence differences in PP_{II} propensity, even when other intramolecular factors are ignored. R_h predicted from IDP sequence and Eq (6) seemed to work best when using an experimental PP_{II} propensity scale from Hilser and colleagues [19], or a composite scale that combined the Hilser, Kallenbach [17], and Creamer [18] propensities, giving an average error of ~2.5 Å for an IDP dataset covering a wide range of residue lengths, net charge, and sequence composition. As examples of sequence differences in this dataset, the fractional number of PRO residues (f_{PRO} = (# PRO residues)/N) varied from 0 to 0.24, SER from 0.02 to 0.20, GLU from 0.06 to 0.31, and ALA from 0 to 0.16, indicating significant sequence diversity among the IDPs that were tested.

If it were established that molecular descriptions for R_h depend mostly on PP_{II} propensities for disordered proteins, this would have important implications. First, R_h well-above random coil estimates would indicate non-trivial preferences for PP_{II} structure. Fig 1 shows this to be the case for many IDPs. And second, large variations in R_h for IDPs with similar N would indicate large differences in propensity for PP_{II} structure among the biologically common amino acids. Observed differences in amino acid propensity for PP_{II} [17–19,53] are thus consistent with the observed differences in R_h for IDPs with similar N. For example, consider that R_h varied from 24.5 Å to 32.4 Å for IDPs with N = 87–97 in Fig 1. The average prediction error in R_h for these 8 IDPs from using Eq (6) and the composite propensity scale was only 1.7 ± 0.7 Å, though net charge ranged from 4 to 29 for these proteins. In contrast, predictions using random coil values give R_h from 20.5 to 21.7 Å with an average error of 6.4 ± 2.7 Å.

The simulation-derived relationship between R_h, N, and $f_{PPII,chain}$ appears to be surprisingly simple for disordered proteins. As noted above, Eq (6) should be interpreted as an ideal relationship that excludes many molecular factors known to regulate structural preferences in proteins (e.g., electrostatic effects, *cis-trans* isomerization rates). Observed deviations from this "ideal" behavior can then be interpreted in terms of factors that were not modeled, as shown (Fig 4B). We recognize that exclusive use of poly-ALA for computational modeling may prove to be unjustified with further studies. Poly-ALA was used as a simplifying step since the effects of N on R_h were mostly independent of amino acid sequence in previous HSC-based simulations and agreed with general IDP trends determined from a literature survey [22,49]. As shown here, this simulation-derived relationship provides a straight-forward molecular explanation for R_h variations among IDPs. The R_h dependence on $f_{PPII,chain}$ also predicts heat-induced compaction of IDP R_h since the enthalpy of unfolding PP_{II} structure is positive [16,64]. Many studies have demonstrated R_h compaction caused by elevated temperatures for IDPs [22,43,44].

As noted above, the simulation results presented here could be interpreted as indicating that charge effects on R_h are generally weak for IDPs, relative to the effects of intrinsic PP_{II} propensities. These data demonstrate, however, that certain sequence patterns of charge can modulate R_h substantially (see Fig 6). For charged groups, this would be those that are separated at distances averaging less than the solution Debye length, involving identical charge type (i.e., positive or negative), and within a region showing higher than typical charge bias. These general rules are in qualitative agreement with results from Pappu and colleagues showing that simulated hydrodynamic sizes for highly charged and disordered polypeptides, with every residue modeled as GLU or LYS, depend strongly on the mixing of negative and positive charge types [15]. In that study, mixing of charge types in a sequence caused structural compaction relative to biased charge distributions, similar to our own conclusions. The observation that unfavorable charge-charge interactions between side chain groups can promote PP_{II} structure (Figs 6A and 7 insets) has also been noticed in computational studies from other researchers [14,65]. This result predicts multiple mechanisms for charge-mediated regulation of IDP structure; possibly owing to both the accumulation of charge and local modulation of PP_{II} propensities. Overall, these data demonstrate the importance of sequence context for understanding the structural properties of IDPs and for describing quantitatively how disordered protein structures respond to discrete perturbations such as changes in charge state and amino acid substitutions.

Methods

Computer generation of polypeptide structures

Detailed description of the computer algorithm that was used is provided elsewhere [22,24]. Briefly, simulations of disordered protein structures were limited to poly-ALA polypeptides. Main chain atoms of poly-ALA were generated using the standard bond angles and bond lengths [66] and a random sampling of the dihedral angles Φ, Ψ, and ω. The dihedral angle ω was given a Gaussian fluctuation of ±5° around the *trans* value of 180°. To sample conformational space efficiently, (Φ,Ψ) values were restricted to the allowed Ramachandran regions [67]. Of the two possible positions of the side chain Cβ atom, the one corresponding to L-alanine was used throughout the studies. To calculate state distributions typical of protein ensembles, a structure-based energy function parameterized to solvent-accessible surface areas was used to population-weight the generated structures [54–62].

Supporting Information

S1 Fig. Comparison of f_{PPII} and S_{PPII}. In this figure, S_{PPII} is the average applied sampling rate for PP_{II} for residues with $S_{PPII} \neq 0$ in a simulation, while f_{PPII} was the observed per-position average PP_{II} rate, also excluding residues with $S_{PPII} = 0$. Open circles are from ensembles where position-specific S_{PPII} followed the pattern specified in the text (i.e., different simulations had different S_{PPII} ranging from 0.1 to 0.9 in 0.1 increments applied to each residue, every other residue, every third residue, etc.) which is why circles align at $S_{PPII} = 0.1–0.9$ in 0.1 increments. Blue circles give the average f_{PPII} for each applied S_{PPII}. Open squares represent this calculation performed on simulations using randomly assigned position-specific S_{PPII}. Stippled line is the identity; solid line is the relationship between f_{PPII} and S_{PPII} established previously for S_{PPII} applied at constant values across all residues [22]. In general, f_{PPII} trends with S_{PPII} by: $f_{PPII} = S_{PPII}\text{-}0.062\cdot\exp(-(S_{PPII}\text{-}0.63)^2/(2\cdot0.28^2))$. This gives the algorithm the ability to target specific f_{PPII} from the applied value of S_{PPII}.
(TIF)

S2 Fig. Correlation of experimental PP_{II} propensities for the common amino acids. Panel **A**, correlation of Kallenbach [17] and Creamer reported values [18]. Panel **B**, correlation of Kallenbach and Hilser reported values [19]. Panel **C**, correlation of Creamer and Hilser reported values. Panel **D**, correlation of Creamer and Zondlo reported values [53].
(TIF)

S3 Fig. Simulated effect of positive charged residues on R_h. Stippled line is R_h from Eq (6) with $N = 25$ and $f_{PPII,chain}$ from 0 to 0.98. Symbols are simulated R_h from ensembles of poly-ALA ($N = 25$) using Eq (3) ($R_h = <L>/2$). Filled circles have each residue modeled with positive charge at the Cβ atom. Filled squares have every other residue modeled with positive charge, filled triangles have every third residue modeled with positive charge, and X represents every fourth residue modeled with positive charge. **Inset:** comparison of observed $f_{PPII,chain}$ to $f_{PPII,chain}$ expected from the applied S_{PPII} (following Fig 6A inset description). Inset symbols match panel representations.
(TIF)

S4 Fig. Correlation of normalized error in predicted R_h to net charge density. Shown are correlations (R^2) between normalized error and net charge density for all IDPs, IDPs in the high charge bias group (labeled as "high bias"), and IDPs in the low charge bias group (labeled as "low bias"). Red columns are correlations from using the Kallenbach propensity scale to predict R_h, blue from using the Creamer propensities, white the Hilser propensities, and black the composite propensity scale.
(TIF)

S1 Table. IDP dataset.
(DOCX)

S2 Table. Sequence of each IDP in dataset.
(DOCX)

Author Contributions

Conceived and designed the experiments: STW. Performed the experiments: STW. Analyzed the data: MET MJT DD STW. Wrote the paper: STW.

References

1. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inf Ser. 2000; 11: 161–171.

2. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. Intrinsic disorder and functional proteomics. Biophys J. 2007; 92: 1439–1456. PMID: 17158572
3. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol. 2004; 337: 635–645. PMID: 15019783
4. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of intrinsically disordered regions and proteins. Chem Rev. 2014; 114: 6589–6631. doi: 10.1021/cr400525m PMID: 24773235
5. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. 2004; 32: 1037–1049. PMID: 14960716
6. Uversky VN, Davé V, Iakoucheva LM, Malaney P, Metallo SJ, Pathak RR, Joerger AC. Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. Chem Rev. 2014; 114: 6844–6879. doi: 10.1021/cr400713r PMID: 24830552
7. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signaling and regulation. Nat Rev Mol Cell Biol. 2015; 16: 18–29. doi: 10.1038/nrm3920 PMID: 25531225
8. Babu MM, van der Lee R, de Groot NS, Gsponer J. Intrinsically disordered proteins: regulation and disease. Curr Opin Struct Biol. 2011; 21: 432–440. doi: 10.1016/j.sbi.2011.03.011 PMID: 21514144
9. Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol. 2002; 323: 573–584. PMID: 12381310
10. Shi Z, Chen K, Liu Z, Kallenbach NR. Conformation of the backbone in unfolded proteins. Chem Rev. 2006; 106: 1877–1897. PMID: 16683759
11. Shi Z, Olson CA, Rose GD, Baldwin RL, Kallenbach NR. Polyproline II structure in a sequence of seven alanine residues. Proc Natl Acad Sci USA. 2002; 99: 9190–9195. PMID: 12091708
12. Schweitzer-Stenner R. Conformational propensities and residual structures in unfolded peptides and proteins. Mol BioSyst. 2012; 8: 122–133. doi: 10.1039/c1mb05225j PMID: 21879108
13. Müller-Späth S, Soranno A, Hirschfeld V, Hofmann H, Rüegger S, Reymond L, Nettels D, Schuler B. Charge interactions can dominate the dimensions of intrinsically disordered proteins. Proc Natl Acad Sci USA. 2010; 107: 14609–14614. doi: 10.1073/pnas.1001743107 PMID: 20639465
14. Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu RV. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. Proc Natl Acad Sci USA. 2010; 107: 8183–8188. doi: 10.1073/pnas.0911107107 PMID: 20404210
15. Das RK, Pappu RV. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc Natl Acad Sci USA. 2013; 110: 13392–13397. doi: 10.1073/pnas.1304749110 PMID: 23901099
16. Chen K, Liu Z, Kallenbach NR. The polyproline II conformation in short alanine peptides is noncooperative. Proc Natl Acad Sci USA. 2004; 101: 15352–15357. PMID: 15489268
17. Shi Z, Chen K, Liu Z, Ng A, Bracken WC, Kallenbach NR. Polyproline II propensities from GGXGG peptides reveal an anticorrelation with beta-sheet scales. Proc Natl Acad Sci USA. 2005; 102: 17964–17968. PMID: 16330763
18. Rucker AL, Pager CT, Campbell MN, Qualls JE, Creamer TP. Host-guest scale of left-handed polyproline II helix formation. Proteins 2003; 53: 68–75. PMID: 12945050
19. Elam WA, Schrank TP, Campagnolo AJ, Hilser VJ. Evolutionary conservation of the polyproline II conformation surrounding intrinsically disordered phosphorylation sites. Protein Sci. 2013; 22: 405–417. doi: 10.1002/pro.2217 PMID: 23341186
20. Cho JH, Sato S, Horng JC, Anil B, Raleigh DP. Electrostatic interactions in the denatured state ensemble: their effect upon protein folding and protein stability. Arch Biochem Biophys. 2008; 469:20–28. PMID: 17900519
21. Cho JH, Raleigh DP. Mutational analysis demonstrates that specific electrostatic interactions can play a key role in the denatured state ensemble of proteins. J Mol Biol. 2005; 353:174–185. PMID: 16165156
22. Langridge TD, Tarver MJ, Whitten ST. Temperature effects on the hydrodynamic radius of the intrinsically disordered N-terminal region of the p53 protein. Proteins 2014; 82: 668–678. doi: 10.1002/prot.24449 PMID: 24150971
23. Perez RB, Tischer A, Auton M, Whitten ST. Alanine and proline content modulate global sensitivity to discrete perturbations in disordered proteins. Proteins 2014; 82: 3373–3384. doi: 10.1002/prot.24692 PMID: 25244701
24. Whitten ST, Yang HW, Fox RO, Hilser VJ. Exploring the impact of polyproline II (PII) conformational bias on the binding of peptides to the SEM-5 SH3 domain. Protein Sci. 2008; 17: 1200–1211. doi: 10.1110/ps.033647.107 PMID: 18577755

25. Richards FM. Areas, volumes, packing, and protein structure. Annu Rev Biophys Bioeng. 1977; 6: 151–176. PMID: 326146

26. Lowry DF, Stancik A, Shrestha RM, Daughdrill GW. Modeling the accessible conformations of the intrinsically unstructured transactivation domain of p53. Proteins 2008; 71: 587–598. PMID: 17972286

27. Donaldson L, Capone JP. Purification and characterization of the carboxyl-terminal transactivation domain of Vmw65 from herpes simplex virus type 1. J Biol Chem. 1992; 267: 1411–1414. PMID: 1309782

28. Sivakolundu SG, Nourse A, Moshiach S, Bothner B, Ashley C, Satumba J, Lahti J, Kriwacki RW. Intrinsically unstructured domains of Arf and Hdm2 form bimolecular oligomeric structures in vitro and in vivo. J Mol Biol. 2008; 384: 240–254. doi: 10.1016/j.jmb.2008.09.019 PMID: 18809412

29. Yi S, Boys BL, Brickenden A, Konermann L, Choy WY. Effects of zinc binding on the structure and dynamics of the intrinsically disordered protein prothymosin alpha: evidence for metalation as an entropic switch. Biochemistry 2007; 46: 13120–13130. PMID: 17929838

30. Sanchez-Puig N, Veprintsev DB, Fersht AR. Binding of natively unfolded HIF-1alpha ODD domain to p53. Mol Cell. 2005; 17: 11–21. PMID: 15629713

31. Campbell KM, Terrell AR, Laybourn PJ, Lumb KJ. Intrinsic structural disorder of the C-terminal activation domain from the bZIP transcription factor Fos. Biochemistry 2000; 39: 2708–2713. PMID: 10704222

32. Geething NC, Spudich JA. Identification of a minimal myosin Va binding site within an intrinsically unstructured domain of melanophilin. J Biol Chem. 2007; 282: 21518–21528. PMID: 17513864

33. Soragni A, Zambelli B, Mukrasch MD, Biernat J, Jeganathan S, Griesinger C, Ciurli S, Mandelkow E, Zweckstetter M. Structural characterization of binding of Cu(II) to tau protein. Biochemistry 2008; 47: 10841–10851. doi: 10.1021/bi8008856 PMID: 18803399

34. Adkins JN, Lumb KJ. Intrinsic structural disorder and sequence features of the cell cycle inhibitor p57Kip2. Proteins 2002; 46: 1–7. PMID: 11746698

35. Uversky VN, Permyakov SE, Zagranichny VE, Rodionov IL, Fink AL, Cherskaya AM, Wasserman LA, Permyakov EA. Effect of zinc and temperature on the conformation of the gamma subunit of retinal phosphodiesterase: a natively unfolded protein. J Proteome Res. 2002; 1: 149–159. PMID: 12643535

36. Haaning S, Radutoiu S, Hoffmann SV, Dittmer J, Giehm L, Otzen DE, Stougaard J. An unusual intrinsically disordered protein from the model legume Lotus japonicus stabilizes proteins in vitro. J Biol Chem. 2008; 283: 31142–31152. doi: 10.1074/jbc.M805024200 PMID: 18779323

37. Permyakov SE, Millett IS, Doniach S, Permyakov EA, Uversky VN. Natively unfolded C-terminal domain of caldesmon remains substantially unstructured after the effective binding to calmodulin. Proteins 2003; 53: 855–862. PMID: 14635127

38. Paleologou KE, Schmid AW, Rospigliosi CC, Kim HY, Lamberto GR, Fredenburg RA, Lansbury PT Jr, Fernandez CO, Eliezer D, Zweckstetter M, Lashuel HA. Phosphorylation at Ser-129 but not the phosphomimics S129E/D inhibits the fibrillation of alpha-synuclein. J Biol Chem. 2008; 283: 16895–16905. doi: 10.1074/jbc.M800747200 PMID: 18343814

39. Baker JMR. Structural characterization and interactions of the CFTR regulatory region. Ph.D. Dissertation. University of Toronto. 2009.

40. Choi UB, McCann JJ, Weninger KR, Bowen ME. Beyond the random coil: stochastic conformational switching in intrinsically disordered proteins. Structure 2011; 19: 566–576. doi: 10.1016/j.str.2011.01.011 PMID: 21481779

41. Magidovich E, Orr I, Fass D, Abdu U, Yifrach O. Intrinsic disorder in the C-terminal domain of the Shaker voltage-activated K1 channel modulates its interaction with scaffold proteins. Proc Natl Acad Sci USA. 2007; 104: 13022–13027. PMID: 17666528

42. Sanchez-Puig N, Veprintsev DB, Fersht AR. Human full-length securin is a natively unfolded protein. Protein Sci. 2005; 14: 1410–1418. PMID: 15929994

43. Kjaergaard M, Nørholm AB, Hendus-Altenburger R, Pedersen SF, Poulsen FM, Kragelund BB. Temperature-dependent structural changes in intrinsically disordered proteins: formation of alpha-helices or loss of polyproline II. Protein Sci. 2010; 19: 1555–1564. doi: 10.1002/pro.435 PMID: 20556825

44. Wuttke R, Hofmann H, Nettels D, Borgia MB, Mittal J, Best RB, Schuler B. Temperature-dependent solvation modulates the dimensions of disordered proteins. Proc Natl Acad Sci USA. 2014; 111: 5213–5218. doi: 10.1073/pnas.1313006111 PMID: 24706910

45. Wilkins DK, Grimshaw SB, Receveur V, Dobson CM, Jones JA, Smith LJ. Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. Biochemistry 1999; 38: 16424–16431. PMID: 10600103

46. Tcherkasskaya O, Davidson EA, Uversky VN. Biophysical constraints for protein structure prediction. J Proteome Res. 2003; 2: 37–42. PMID: 12643541

47. Fitzkee NC, Rose GD. Reassessing random-coil statistics in unfolded proteins. Proc Natl Acad Sci USA. 2004; 101: 12497–12502. PMID: 15314216

48. Auton M, Ferreon ACM, Bolen DW. Metrics that differentiate the origins of osmolyte effects on protein stability: a test of the surface tension proposal. J Mol Biol. 2006; 361:983–992. PMID: 16889793

49. Marsh JA, Forman-Kay JD. Sequence determinants of compaction in intrinsically disordered proteins. Biophys J. 2010; 98: 2383–2390. doi: 10.1016/j.bpj.2010.02.006 PMID: 20483348

50. Cowan PM, McGavin S. Structure of poly-L-proline. Nature 1955; 176:501–503.

51. Levenberg K. A method for the solution of certain non-linear problems in least squares. Quart Appl Math. 1944; 2: 164–168.

52. Marquardt DW. An algorithm for least-squares estimation of nonlinear parameters. SIAM J Appl Math. 1963; 11: 431–441.

53. Brown AM, Zondlo NJ. A propensity scale for type II polyproline helices (PPII): aromatic amino acids in proline-rich sequences strongly disfavor PPII due to proline-aromatic interactions. Biochemistry 2012; 51: 5041–5051. doi: 10.1021/bi3002924 PMID: 22667692

54. Baldwin RL. Temperature dependence of the hydrophobic interaction in protein folding. Proc Natl Acad Sci USA. 1986; 83: 8069–8072. PMID: 3464944

55. Murphy KP, Freire E. Thermodynamics of structural stability and cooperative folding behavior in proteins. Adv Protein Chem. 1992; 43: 313–361. PMID: 1442323

56. Murphy KP, Bhakuni V, Xie D, Freire E. Molecular basis of cooperativity in protein folding. III. Structural identification of cooperative folding units and folding intermediates. J Mol Biol. 1992; 227: 293–306. PMID: 1522594

57. Lee KH, Xie D, Freire E, Amzel LM. Estimation of changes in side chain configurational entropy in binding and folding: General methods and application to helix formation. Proteins 1994; 20: 68–84. PMID: 7824524

58. Xie D, Freire E. Structure based prediction of protein folding intermediates. J Mol Biol. 1994; 242: 62–80. PMID: 8078072

59. Gómez J, Hilser VJ, Xie D, Freire E. The heat capacity of proteins. Proteins 1995; 22: 404–412. PMID: 7479713

60. D'Aquino JA, Gómez J, Hilser VJ, Lee KH, Amzel LM, Freire E. The magnitude of the backbone conformational entropy change in protein folding. Proteins 1996; 25: 143–156. PMID: 8811731

61. Habermann SM, Murphy KP. Energetics of hydrogen bonding in proteins: A model compound study. Protein Sci. 1996; 5: 1229–1239. PMID: 8819156

62. Luque I, Mayorga OL, Freire E. Structure-based thermodynamic scale of alpha-helix propensities in amino acids. Biochemistry 1996; 35: 13681–13688. PMID: 8885848

63. Malmberg CG, Maryott AA. Dielectric constant of water from 0 to 100C. J Res Natl Bur Stand. 1956; 56: 1–8.

64. Hamburger JB, Ferreon JC, Whitten ST, Hilser VJ. Thermodynamic mechanism and consequences of the polyproline II (PII) structural bias in the denatured states of proteins. Biochemistry 2004; 43: 9790–9799. PMID: 15274633

65. Krimm S, Mark JE. Conformations of polypeptides with ionized side chains of equal length. Proc Natl Acad Sci USA. 1968; 60: 1122–1129. PMID: 16591670

66. Momany FA, McGuire RF, Burgess AW, Scheraga HA. Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. J Phys Chem. 1975; 79: 2361–2381.

67. Mandel N, Mandel G, Trus BL, Rosenberg J, Carlson G, Dickerson RE. Tuna cytochrome c at 2.0 Å resolution. III. Coordinate optimization and comparison of structures. J Biol Chem. 1977; 252: 4619–4636. PMID: 194885