Sequence Determinants of Compaction in Intrinsically Disordered Proteins

Joseph A. Marsh and Julie D. Forman-Kay*
Molecular Structure and Function, Hospital for Sick Children, and Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada

ABSTRACT Intrinsically disordered proteins (IDPs), which lack folded structure and are disordered under nondenaturing conditions, have been shown to perform important functions in a large number of cellular processes. These proteins have interesting structural properties that deviate from the random-coil-like behavior exhibited by chemically denatured proteins. In particular, IDPs are often observed to exhibit significant compaction. In this study, we have analyzed the hydrodynamic radii of a number of IDPs to investigate the sequence determinants of this compaction. Net charge and proline content are observed to be strongly correlated with increased hydrodynamic radii, suggesting that these are the dominant contributors to compaction. Hydrophobicity and secondary structure, on the other hand, appear to have negligible effects on compaction, which implies that the determinants of structure in folded and intrinsically disordered proteins are profoundly different. Finally, we observe that polyhistidine tags seem to increase IDP compaction, which suggests that these tags have significant perturbing effects and thus should be removed before any structural characterizations of IDPs. Using the relationships observed in this analysis, we have developed a sequence-based predictor of hydrodynamic radius for IDPs that shows substantial improvement over a simple model based upon chain length alone.

INTRODUCTION

The focus of structural biology has been dramatically expanded in recent years with the widespread interest in intrinsically disordered proteins (IDPs). These proteins have been associated with a large number of important cellular processes and, in particular, are often involved in regulatory macromolecular interactions (1). Whereas chemically denatured proteins typically exhibit behavior consistent with simple random or statistical coil models (2,3), IDPs can possess significant nonrandom structure. In particular, they are often more compact than is expected or observed for chemically denatured proteins of the same length (4). Although numerous recent studies have attempted to characterize the structural properties of disordered states of proteins (5–7), the precise nature of structure in IDPs is still an important question. Although transient secondary structure is common and can be studied in detail with NMR techniques (8,9), the prevalence and importance of tertiary contacts is less certain. In addition, the origin of the variation in compaction seen in IDPs is unclear: does it arise from differences in secondary and/or tertiary structure? If so, what is the nature of this structure?

The sequence differences between folded and intrinsically disordered proteins have been studied extensively (10). IDPs tend to be deficient in the hydrophobic residues necessary for folding and rich in charged residues (11). Various methods have been developed that can quite effectively predict intrinsically disordered regions of proteins from primary amino acid sequences alone (12–15). However, despite the observation that there is significant structural variation between different IDPs (4), there has been little investigation into the sequence determinants of this structure.

A simple way to assess structure in a disordered protein is to measure its hydrodynamic radius (R_h). The R_h is the radius of an idealized sphere that would diffuse at the same rate as the molecule of interest, and is based on the Stokes-Einstein relation in Eq. 1, where k_B is the Boltzmann constant, T is the temperature, η is the viscosity, and D is the translational diffusion coefficient. Thus, although the R_h is not a true measure of the radius of a nonglobular protein, as its diffusion is related to its nonspherical shape, it is very useful as a simple measure of compaction in disordered proteins.

$$R_h = \frac{k_B T}{6\pi\eta D}. \tag{1}$$

Commonly used methods for measuring R_h in IDPs include size exclusion chromatography (SEC) and pulsed-field-gradient (PFG) NMR. In SEC experiments, the size exclusion column is calibrated using folded protein standards of known molecular weight, which allows the apparent molecular weight of the protein of interest to be measured. The R_h is then simply determined as the R_h expected for a folded protein of that molecular weight, for which simple relations exist (16). In PFG NMR experiments, the translational diffusion coefficient of the protein can be directly measured and compared to a standard of known R_h, allowing simple determination of the R_h (17).

In this study, we have compiled R_h measurements and amino acid sequences for a sizeable set of IDPs, which has allowed us to investigate the sequence determinants of compaction in these proteins. We show that the number of proline residues and the net charge seem to be the primary natural determinants of compaction in IDPs. This has important implications for understanding the nature of

Submitted January 11, 2010, and accepted for publication February 4, 2010.
*Correspondence: forman@sickkids.ca
Editor: Patrick Loria.

© 2010 by the Biophysical Society
0006-3495/10/05/2383/8 $2.00

doi: 10.1016/j.bpj.2010.02.006

disordered-state structure and suggests a limited role for hydrophobic contacts. In addition, we also show that polyhistidine tags appear to have a large effect on compaction, suggesting that they should be removed before structural studies of IDPs are performed. Finally, we have used our findings to develop a new method for predicting the R_h of an IDP from the amino acid sequence that provides a substantial improvement over methods based on chain length alone.

METHODS

Compilation of hydrodynamic radius measurements and protein sequences

A search of the literature was conducted to identify R_h measurements of IDPs and their associated amino acid sequences. To avoid sequence bias from homologous proteins, the program Needle from the EMBOSS suite (18) was used to identify sequences with >50% similarity to each other; only the longer of two homologous sequences was retained. In addition, only proteins with <300 residues were chosen, because we expected that larger proteins would be more likely to contain a mixture of folded and disordered regions. Table S1 in the Supporting Material presents all of the proteins used in this study, along with their R_h values and amino acid sequences. In total, we used 32 R_h measurements, 12 determined by PFG NMR and 20 by SEC. We did not use any dynamic light-scattering measurements of R_h, because the number of IDP measurements we found was too small to adequately assess their similarity to PFG NMR and SEC measurements. For some proteins, the precise amino acid sequence could not be obtained due to insufficient details regarding the protein expression construct represented by X in Table S1. pH values for each R_h measurement were obtained for the purpose of determining the charge state of histidine residues by assuming a side-chain pKa of 6.8 (19). We also repeated all relevant calculations assuming a histidine side-chain pKa of 6.0; in this case, our results were nearly identical (not shown). Finally, we compiled R_h measurements for a number of folded and chemically denatured proteins; these are listed in Table S2 and Table S3.

Calculation of r_{aa} values and associated error bars

To determine error bars for the calculated r_{aa} values, which describe the correlation between different amino acid residues and increased compaction or expansion (defined later), we employed a simple bootstrapping analysis whereby random sets of 32 proteins were selected from the full data set (importantly, allowing for multiples of each protein so that each set is different). The r_{aa} values were calculated from each random set and the procedure was performed 10,000 times; error bars represent the standard deviations of r_{aa} from these replicates. The source code for these calculations is provided in the Supporting Material.

RESULTS

Intrinsically disordered proteins have a greater range of compaction than chemically denatured proteins

Previous studies have shown that there is a strong correspondence between the number of residues in a folded or disordered protein and its molecular size as measured by R_h or radius of gyration (R_g). The simple power-law scaling relationship in Eq. 2 has been used to provide remarkably good predictions; R is the R_g or R_h, N is the number of residues in the protein, and R_0 and v are constants (2,20). In an excluded-volume random coil, v is predicted to be 0.588 (21). This is very close to the empirically determined value of 0.598 based upon small-angle x-ray scattering measurements of R_g for a number of chemically denatured proteins, thus providing one of the strongest arguments for the random-coil-like behavior of chemically denatured proteins (2).

$$R = R_0 N^{\nu}. \tag{2}$$

In Fig. 1, we plot the R_h values versus the number of residues for a large number of folded, chemically denatured, and intrinsically disordered proteins. Each class of protein has been fit to Eq. 2, providing the following relationships between R_h and number of residues:

$$R_h^{\text{folded}} = 4.92N^{0.285}; \tag{3}$$

$$R_h^{\text{denatured}} = 2.33N^{0.549}; \tag{4}$$

and

$$R_h^{\text{IDP}} = 2.49N^{0.509}. \tag{5}$$

These relations are in very good agreement with the commonly used values determined by Wilkins et al. ($R_0 = 4.75$ and $v = 0.29$ for folded; $R_0 = 2.21$ and $v = 0.57$ for denatured) (20). R_h values predicted from the above relations agree very well with their experimentally determined values for folded and chemically denatured proteins, with root-mean-squared deviations (RMSDs) of 0.56 and 1.20 Å, respectively. However, for IDPs, a much greater range in compaction is observed, and the agreement is much worse (RMSD = 3.85 Å). Overall, the IDPs tend to be more compact than chemically denatured proteins, as has been recognized previously (4), although they can occasionally be even more expanded. Clearly, then, there must be some heterogeneity in the structural properties of IDPs to account for this divergence from the simple power-law relationship that describes folded and chemically denatured proteins so well. The main goal of this study is to relate this variation in compaction to primary amino acid sequences.

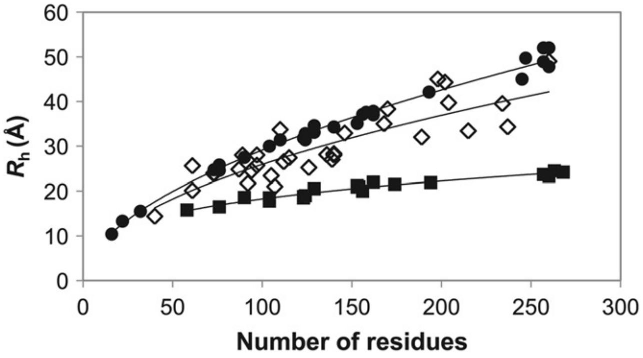

FIGURE 1 Number of residues versus R_h for 20 folded (*solid squares*), 27 chemically denatured (*solid circles*), and 32 intrinsically disordered (*open diamonds*) proteins.

Relating sequence to compaction in intrinsically disordered proteins

To express compaction, we define the term relative R_h as $R_{rel} = R_h/R_h^{IDP}$, where R_h^{IDP} comes from Eq. 5. It is important to note that R_{rel} shows no significant correlation with the number of amino acids for the proteins in the data set ($r = 0.03$), so it is independent of chain length. To assess the sequence dependence of compaction, we calculate the Pearson correlation coefficient, r_{aa}, between the fractional content of each type of amino acid (e.g., if alanines constituted 10% of the residues in a protein, this would be 0.1) and R_{rel} for each protein. Amino acids that tend to be associated with more expanded proteins will have positive values of r_{aa}, whereas those associated with increased compaction will have negative values.

In Fig. 2 *A*, we plot the r_{aa} for each type of amino acid. The first notable feature of this figure is the residue histidine, which has the lowest r_{aa} value. Given that 9 of the 32 proteins in the data set contain polyhistidine tags used for the affinity purification, we wondered whether the presence of these tags might be associated with increased compaction. In Fig. 2 *B*, we have treated histidine residues present in a polyhistidine tag separately from other histidine residues; they are identified by an asterisk. In this plot, we see that polyhistidine residues have an even stronger association with compaction, whereas other histidine residues are not significantly associated with compaction, suggesting that polyhistidine tags can have a major effect on the compaction of IDPs in vitro. This can be seen in more detail in Fig. 3 *A*, where the R_{rel} for each protein is plotted versus the fraction of polyhistidine tag residues in each protein (zero for most proteins). There is a remarkable tendency for the nine polyhistidine-tagged proteins to be more compact on average than nontagged proteins.

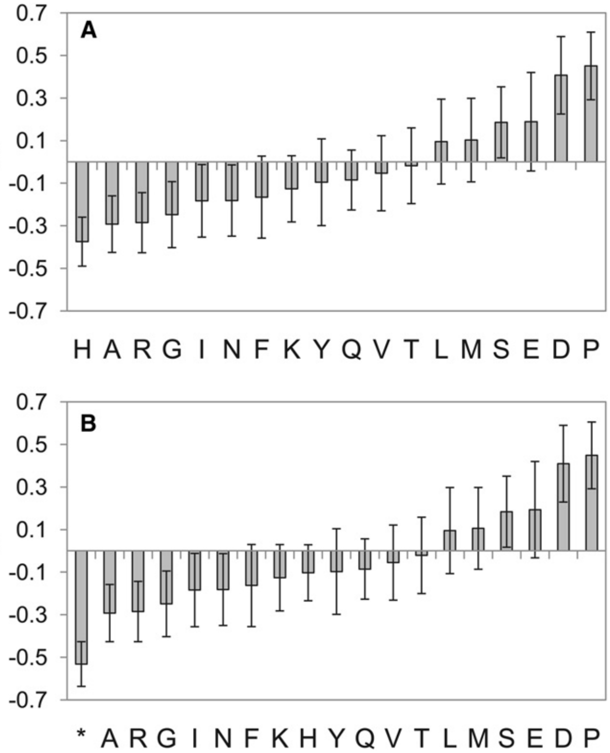

FIGURE 2 (*A*) Correlation (r_{aa}) between the fractional content of each amino acid from each protein and R_{rel}. (*B*) Same as *A*, except that histidine residues present in a polyhistidine tag are considered separately (asterisk). Error bars were calculated with a bootstrapping procedure (see Methods). W and C are not shown because the number of these residues in the data set was very low.

Proline has the highest r_{aa} value in Fig. 2, *A* and *B*, and thus has the strongest association with more highly expanded proteins. This is further demonstrated in Fig. 3 *B*, where we plot R_{rel} for all proteins in the data set versus the fraction of proline residues in each protein. There is a fairly strong, statistically significant correlation ($r = 0.46$, $p = 0.004$), suggesting that proline residues play an important role in determining the compaction of IDPs.

Another very interesting aspect of Fig. 2, *A* and *B*, is the distribution of charged residues. Aspartate and glutamate residues have the second- and third-highest r_{aa} values, suggesting that negatively charged side chains are associated with more expanded states. It appears that positively charged residues may have a slight tendency to associate with more compact states, with arginine having one of the lowest r_{aa} values and lysine and histidine also having negative r_{aa} values. This apparent discrepancy between positively and negatively charged residues can be explained if one considers that the IDPs in our data set have a greater tendency to be negatively charged than positively charged (19 of 32 IDPs in our data set have a net negative charge). Thus, positively charged residues will tend to reduce the overall net charge of the protein. In Fig. 3 *C*, we plot R_{rel} versus the absolute net charge for all proteins in the data set and observe a significant correlation ($r = 0.56$, $p = 0.0004$). This correlation is retained even when all histidine residues are ignored ($r = 0.58$, not shown), demonstrating that this association with net charge is not related to the presence of polyhistidine tags. We also compared R_{rel} to the average number of charged residues (treating histidine as its fractional charge state). In this case, very little correlation was observed ($r = 0.15$, $p = 0.21$) (not shown). These results strongly suggest that the overall net charge, but not simply the number of charged residues alone, is a key determinant of IDP compaction.

Given the importance of hydrophobic interactions in the structure of folded proteins, hydrophobicity might be expected to show a strong association with IDP compaction. However, in Fig. 2, *A* and *B*, there is no obvious relation that can be discerned for the hydrophobic residues. In Fig. 3 *D*, we plot the average Kyte-Doolitle hydrophobicity (22) versus R_{rel} for the IDPs in the data set. We observe only a very weak correlation between increased hydrophobicity and compaction that is not statistically significant ($r = -0.10$, $p = 0.29$). Closely related to hydrophobicity is the

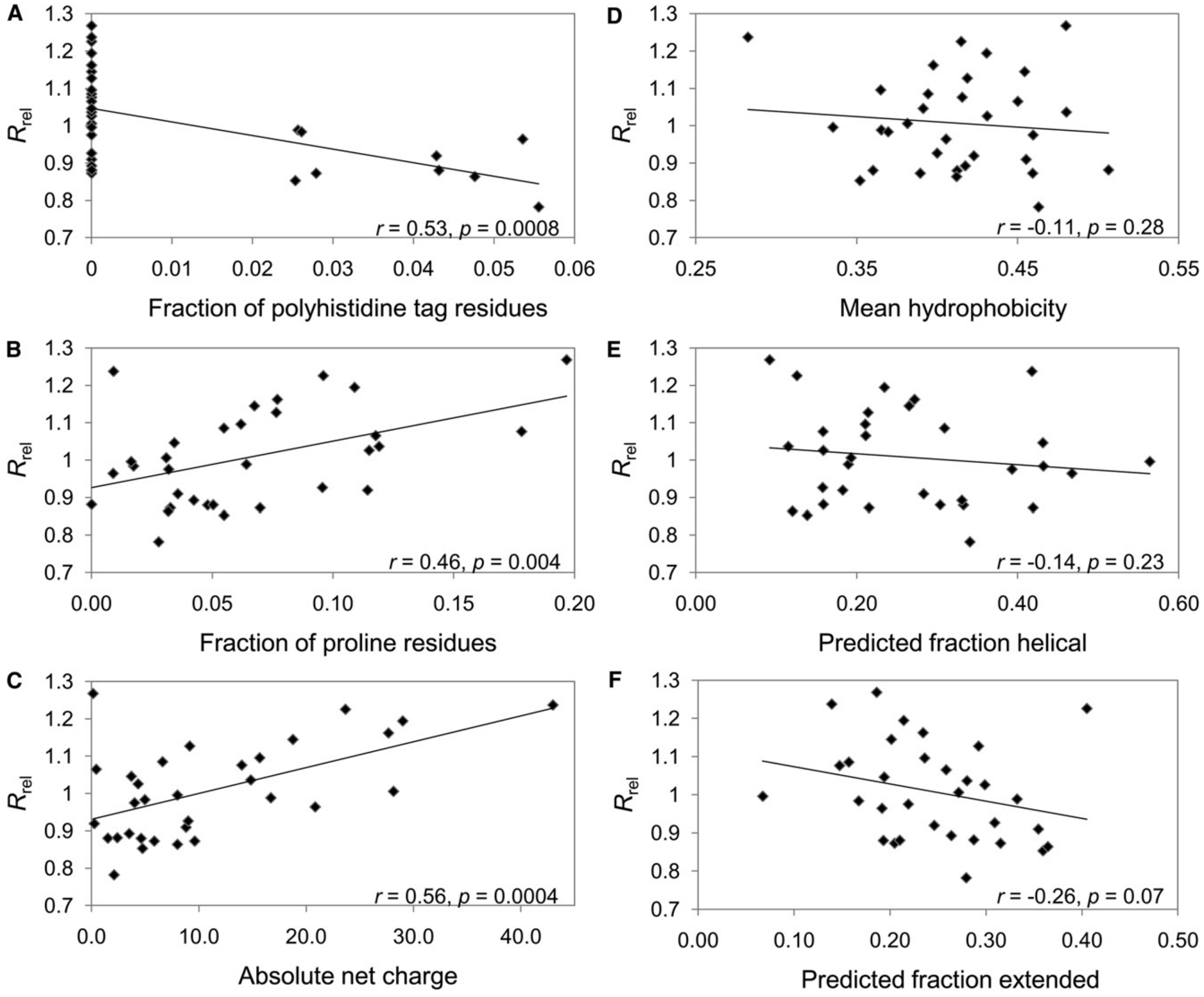

FIGURE 3 Comparisons of R_{rel} for all proteins in the IDP data set to the fraction of polyhistidine tag residues (*A*), fraction of proline residues (*B*), absolute net charge (*C*), mean hydrophobicity (*D*), predicted fraction of helical residues (*E*), and predicted fraction of extended residues (*F*).

average area buried upon folding (AABUF) (23), which was recently used to predict transient collapsed structure in acid-denatured apomyoglobin (24). AABUF shows a stronger correlation with R_{rel} ($r = -0.25$, $p = 0.08$) (not shown). However, the correlations for hydrophobicity and AABUF reduce to -0.02 and -0.01, respectively, if charged residues and prolines are ignored. Although these results do not rule out some role for hydrophobic contacts in the compaction of IDPs, they suggest that any contribution must be quite limited.

We also investigated whether intrinsic propensities for forming secondary structure might be correlated with compaction. In Fig. 3, *E* and *F*, we plot the predicted GOR3 (25) helical and extended secondary structure versus R_{rel} for each protein. There are slight negative correlations for both helical ($r = -0.14$, $p = 0.23$) and extended ($r = -0.26$, $p = 0.07$) secondary structure, suggesting that there may be a relationship between increased secondary structure and reduced R_h. However, the evidence is not strong enough to say with certainty that there is a statistically significant relationship between secondary structure and compaction. In addition, when charged residues and prolines are ignored in the GOR3 predictions, the correlations reduce to 0.049 and -0.17 for helical and extended structures, respectively (not shown).

Comparison of PFG NMR and SEC measurements

In this study, we have combined measurements of R_h that were made using very different experimental methods: PFG NMR and SEC. Therefore, it is important to address whether these different methods give similar measurements. In Fig. 4, we show the R_h values versus number of residues for PFG NMR and SEC measurements, as well as their best-fit lines (Eq. 1). There are no obvious divergences between the two data sets, and the best-fit lines are very similar,

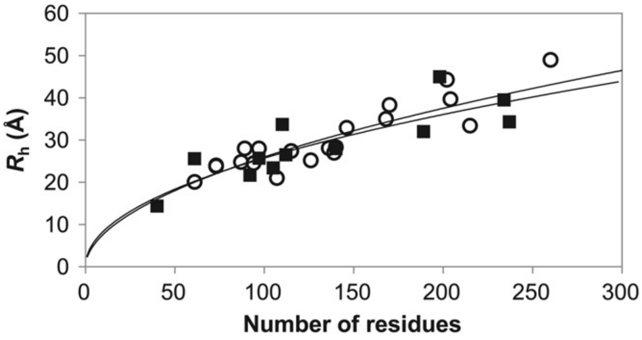

FIGURE 4 Number of residues versus R_h values measured with SEC (*open circles*) and PFG NMR (*solid squares*). The best-fit power-law scaling lines (Eq. 2) are shown for SEC measurements (*upper line*) and PFG NMR measurements (*lower line*).

supporting the similarity of the results obtained by these methods. In addition, we have also compared the correlations identified earlier using only proteins from the separate data sets. The PFG NMR and SEC data sets have correlations of 0.56 and 0.35, respectively, between the fraction of proline residues and R_{rel}, 0.52 and 0.66 between absolute net charge and R_{rel}, and −0.18 and −0.78 between the fraction of polyhistidine residues and R_{rel} (there are only three polyhistidine-tagged proteins in the PFG NMR data set, so there is little significance to this weak correlation). The fact that relatively similar correlations exist in these independent data sets (given the limited numbers of measurements available) both confirms the significance of these correlations and supports the validity of combining PFG NMR and SEC measurements in our full data set.

Improved prediction of hydrodynamic radius using sequence information

As discussed above, R_h values for folded and chemically denatured proteins can be predicted extremely well from the number of residues alone using a simple power-law relation (Eq. 1). IDPs, on the on the other hand, show much greater variation in compaction, which we observe to be correlated with their proline content, net charge, and the presence or absence of a polyhistidine tag. Therefore, we decided to investigate whether this sequence information could be used to improve R_h predictions of IDPs.

In Eq. 6, we have extended the simple form of Eq. 2 to account for the fraction of proline residues (P_{pro}), the absolute net charge ($|Q|$) and the presence of a polyhistidine tag, where *A–D* are constants that are fit from the slopes (*A* and *C*) and offsets (*B* and *D*) of the linear fits between R_{rel} and P_{pro} or $|Q|$, and S_{his*} is a scaling factor applied if the protein has a polyhistidine tag (S_{his*} is 1 if no tag is present). These constants are then optimized to maximize the agreement between the predicted and experimental R_h using a simple Monte Carlo procedure. The best-fit parameters for Eq. 6 from all 32 proteins in the data set are $A = 1.24$, $B = 0.904$, $C = 0.00759$, $D = 0.963$, $S_{his*} = 0.901$, $R_0 = 2.49$, and $v = 0.509$.

$$\mathrm{R_h} = \left(AP_{pro} + B\right)\left(C|Q| + D\right)S_{his*}R_0N^{\nu}. \quad (6)$$

To evaluate the ability of this new expression to predict the R_h of IDPs, an all-but-one procedure was performed in which, for each protein, all of the other proteins were used to fit the parameters in both Eqs. 2 and 6. These were then used to predict the R_h of that protein in an unbiased manner. Fig. 5 shows the differences between experimental R_h measurements and values predicted using both methods. We see that the new, sequence-based method (Eq. 6) shows a substantial improvement over the simple power-law relation (Eq. 2), with an RMSD of 2.37 Å between experimental and predicted R_h, compared to 4.13 Å for Eq. 2 (note that this is different from the value of 3.85 Å given earlier, because it comes from the unbiased all-but-one fitting instead of from all data points). In Table 1, we present the comparisons of R_h predictions using varying subsets of the sequence information (i.e., proline content, net charge, and polyhistidine tags). These results demonstrate quite clearly that net charge is the most useful information for predicting R_h, whereas proline content and the presence of a polyhistidine tag are of roughly similar utility. However, the best results are obtained when all sources of information are combined.

DISCUSSION

Our analysis of R_h values for a number of IDPs demonstrates that net charge and proline content are the primary natural determinants of R_h. In addition, the presence of polyhistidine tags in the recombinant protein samples also leads to significantly increased compaction. Hydrophobicity and secondary structure, on the other hand, contribute very little.

The finding that net charge is the most dominant factor contributing to the variation in IDP compaction is interesting but not surprising, as increasing repulsive electrostatic forces would be expected to cause conformational expansion. It is also very interesting when we consider the high fraction of

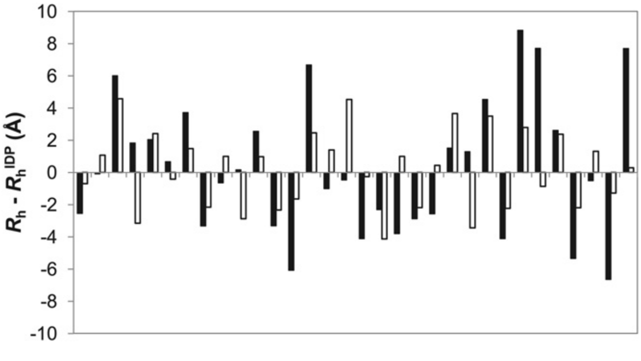

FIGURE 5 Differences between experimentally determined R_h and R_h^{IDP} predicted with the simple power-law model (*solid bars*) and the new sequence-based model (*open bars*) for all 32 IDPs in the data set.

TABLE 1 Comparison of RMSDs between experimental R_h values and values predicted using varying subsets of sequence information

Pro	Q	His*	RMSD (Å)
			4.13
X			3.82
		X	3.74
X		X	3.29
	X		3.25
X	X		2.78
	X	X	2.71
X	X	X	2.37

Pro, proline content; Q, net charge; His*, presence of a polyhistidine tag; X indicates that the applicable term from Eq. 6 was utilized for the calculation of R_h values.

charged residues (10,11) and the high frequency of residues that undergo phosphorylation in IDPs (26). Therefore, phosphorylation or other posttranslational modifications that affect net charge (e.g., ubiquitination or methylation) could represent a simple method for modulating compaction in IDPs. This could be utilized in various regulatory functions. Increased expansion might make binding sites more accessible or modulate the spatial separation between two domains separated by an intrinsically disordered linker. Conversely, increased compaction might bring separate binding elements closer together, thus facilitating an interaction, or sterically block access to another part of the protein.

Despite the strong relationship between compaction and net charge, there is very little correlation with the number of charged residues alone, suggesting that only the net effect of all charged residues is important. Our results are therefore consistent with the recently proposed polyelectrostatic model, in which independent point charges in a disordered protein are averaged in a mean-field interpretation and the overall net charge of the protein is found to be most significant for its interaction properties (27). Although this is obviously a highly simplified model, it has been very effective for describing the phosphorylation dependence of binding a polyvalent IDP to a single ligand (27,28). Of course, at a detailed molecular level, there are likely to be local structural effects due to sequence-dependent charge variations. However, from the perspective of overall molecular compaction, with the resolution available in our data set, only the net charge appears to be significant.

Other examples from the literature provide further support for the dependence of IDP compaction on net charge. For example, intrinsically disordered α-synuclein, which is best known for its aggregating role in Parkinson's disease, has been shown to become more compact at low pH, presumably due to neutralization of charged carboxylate side chains in the acidic C-terminal region (29–31). Another recent study of 14-residue peptides of varying sequence showed that increasing negative charge led to significant conformational expansion (32).

In addition to net charge, the proline content of IDPs also seems to be important for determining their compaction. This is easy to rationalize: proline residues have a strong tendency to adopt backbone dihedral angles in the extended regions of Ramachandran space and, in particular, the highly extended righthand side of the broad β-region, commonly referred to as the polyproline II region (33). This tendency toward extended backbone conformations thus leads to an increased R_h for proline-rich sequences. In addition, one could hypothesize that the tendency of proline residues to undergo *cis-trans* isomerization might inhibit the formation of more compact elements of structure. It seems very likely that the high proline content associated with IDPs is related to their effect on compaction, i.e., in addition to its role in promoting disorder (10), proline content may be evolutionarily related to the functional requirements for varying IDP compaction.

The lack of a contribution from hydrophobic residues in IDP compaction is a very important result of our analysis. Given that burial of hydrophobic residues is the dominant force driving protein folding, much of the speculation regarding transient structure in IDPs has focused upon the importance of hydrophobic side-chain interactions. However, given the deficiency of hydrophobic residues and abundance of charged residues in IDPs (11), it is perhaps not surprising that the effect of hydrophobic clustering is minimal. Of course, there could still be some contribution to compaction from hydrophobicity that is undetectable, given the limited size and resolution of our data set. In a recent study, Krishnan et al. made 10 phenylalanine-to-alanine mutations in the intrinsically disordered Nup116 FG domain (34). This resulted in a slight increase in R_h, from 25.2 to 27.1 Å, suggesting that hydrophobic interactions between phenylalanine side chains likely play some role in the compaction of this protein. In addition, there is significant evidence for hydrophobic clustering in the unfolded states of some folded proteins, including the drkN SH3 domain (35–37), reduced unfolded lysozyme (38), and acid-denatured apomyoglobin (24). The fact that collapsed structure in acid-denatured apomyoglobin could be predicted quite well using the AABUF parameter (24) although AABUF shows little correlation with our IDP data set suggests that there may be fundamental structural differences between IDPs and the unfolded states of normally folded proteins that are related to their very different sequence characteristics.

Fractionally populated secondary structure is known to be common in IDPs (8,9). Therefore, that we do not observe a statistically significant correlation between predicted secondary structure and compaction does not mean it is not an important aspect of IDP structure. Residues that preferentially sample different regions of Ramachandran space could be expected to have some effect on compaction (as is likely the case for proline), and sequences favoring β-turns that reverse the direction of the polypeptide chain should also lead to more compact states. However, these effects cannot be conclusively identified in our limited data set, as their contributions are probably quite small compared to electrostatic interactions.

The observation that polyhistidine tags used for affinity purification of recombinant proteins appear to cause a significant increase in IDP compaction is very interesting but somewhat disturbing, given the frequency with which these tags are used in in vitro studies. The effect is likely due to interactions between the polyhistidine tag and other residues in the protein, possibly to the partially charged (depending on the pH) nature of the histidine side chains. Due to the intrinsic conformational entropy of a disordered polypeptide chain, interactions involving the N- or C-termini are more favorable than internal regions, as was previously shown by Chan and Dill (39) and noted for the unfolded state of the drkN SH3 domain (35). Thus, the N- or C-terminus is the worst possible position for a tag that has a tendency to interact with other regions. Although highly useful for the protein purification procedure, clearly it is important to remove polyhistidine tags before performing any structural or functional studies on IDPs.

A significant limitation of the data set used in this study results from the different conditions under which each measurement was made. For example, although most measurements were made at room temperature, some were made around 4°C which may have a significant effect on any hydrophobic interactions. In addition, widely varying buffer conditions were used. The strong association we identified between electrostatics and compaction suggests that salt concentration should have a substantial effect on IDP structure. Thus, it is likely that the role of charge in IDP compaction is even greater than suggested by the correlations we observe, given the large variations in salt concentration between different experimental measurements. Additional studies of different proteins under uniform experimental conditions or of individual proteins under varying conditions will be extremely useful for more precise assessment of the contributions of different factors to IDP compaction.

The method for prediction of IDP R_h presented in this study should be valuable for future experimental studies of IDPs, as it provides an improved reference to which IDP compaction can be compared. Nevertheless, we still observe significant deviations between predicted and experimentally determined values. Assuming we had a perfect method for R_h prediction, the expected variation arising from experimental error alone can be estimated by looking at the deviation from simple power-law scaling in chemically denatured proteins, where an RMSD of 1.20 Å was observed. Although the chemically denatured proteins have more uniform solvent conditions, the RMSD of 2.37 Å for our predictions of IDP R_h suggests that significant improvement could be made to our prediction method. Increasing the size of the data set and decreasing the variation in sample conditions would likely allow more statistically significant correlations to be observed that could lead to improved predictions. In addition, it is highly likely that IDP structure is not encoded by the fractional residue content alone, but instead is significantly dependent on the order of residues in the primary sequence. For example, local clusters of charged or hydrophobic residues might be expected to have a cooperative effect on compaction compared to a uniform sequence distribution (as is likely the case for the observed effect of polyhistidine tags). Comparing predicted to experimental R_h will be useful for identifying IDPs with such nonrandom or cooperative structural properties.

Our results emphasize the growing recognition that a dynamic continuum from ordered to disordered states exists within proteins, with varying amounts of flexibility and structure being possible within folded and disordered proteins. IDPs do not have homogenous structural properties. Rather, simple sequence properties like charge and proline content can have a large effect on their compaction. The sequence determinants of structure in IDPs are potentially much less complex than in folded proteins, where precise three-dimensional structure must be encoded in the primary amino acid sequence, but nevertheless, they are still likely very important for determining the numerous important biological functions identified for IDPs. As our structural understanding of IDPs improves, so should our ability to relate their structural properties to biological functions.

SUPPORTING MATERIAL

Three tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(10)00234-1.

This work was supported by grants from the Canadian Institutes for Health Research.

REFERENCES

1. Dyson, H. J., and P. E. Wright. 2002. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* 12:54–60.
2. Kohn, J. E., I. S. Millett, …, K. W. Plaxco. 2004. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. USA.* 101:12491–12496.
3. Tanford, C., K. Kawahara, and S. Lapanje. 1966. Proteins in 6-M guanidine hydrochloride. Demonstration of random coil behavior. *J. Biol. Chem.* 241:1921–1923.
4. Uversky, V. N. 2002. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* 11:739–756.
5. Mittag, T., and J. D. Forman-Kay. 2007. Atomic-level characterization of disordered protein ensembles. *Curr. Opin. Struct. Biol.* 17:3–14.
6. Vendruscolo, M. 2007. Determination of conformationally heterogeneous states of proteins. *Curr. Opin. Struct. Biol.* 17:15–20.
7. Meier, S., M. Blackledge, and S. Grzesiek. 2008. Conformational distributions of unfolded polypeptides from novel NMR techniques. *J. Chem. Phys.* 128:052204.
8. Jensen, M. R., K. Houben, …, M. Blackledge. 2008. Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of Sendai virus nucleoprotein. *J. Am. Chem. Soc.* 130:8055–8061.

9. Marsh, J. A., V. K. Singh, …, J. D. Forman-Kay. 2006. Sensitivity of secondary structure propensities to sequence differences between α- and γ-synuclein: implications for fibrillation. *Protein Sci.* 15:2795–2804.
10. Dunker, A. K., J. D. Lawson, …, Z. Obradovic. 2001. Intrinsically disordered protein. *J. Mol. Graph. Model.* 19:26–59.
11. Uversky, V. N., J. R. Gillespie, and A. L. Fink. 2000. Why are ''natively unfolded'' proteins unstructured under physiologic conditions? *Proteins.* 41:415–427.
12. Dosztányi, Z., V. Csizmok, …, I. Simon. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics.* 21:3433–3434.
13. Linding, R., L. J. Jensen, …, R. B. Russell. 2003. Protein disorder prediction: implications for structural proteomics. *Structure.* 11: 1453–1459.
14. Romero, P., Z. Obradovic, …, A. K. Dunker. 2001. Sequence complexity of disordered protein. *Proteins.* 42:38–48.
15. Ward, J. J., J. S. Sodhi, …, D. T. Jones. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 337:635–645.
16. Uversky, V. N. 1993. Use of fast protein size-exclusion liquid chromatography to study the unfolding of proteins which denature through the molten globule. *Biochemistry.* 32:13288–13298.
17. Jones, J. A., D. K. Wilkins, …, C. M. Dobson. 1997. Characterisation of protein unfolding by NMR diffusion measurements. *J. Biomol. NMR.* 10:199–203.
18. Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16:276–277.
19. Cantor, C. R., and P. R. Schimmel. 1980. Biophysical Chemistry, Part I. W. H. Freeman, San Francisco.
20. Wilkins, D. K., S. B. Grimshaw, …, L. J. Smith. 1999. Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. *Biochemistry.* 38:16424–16431.
21. LeGuillou, J. C., and J. Zinn-Justin. 1977. Critical exponents for the n-vector model in three dimensions from field theory. *Phys. Rev. Lett.* 39:95–98.
22. Kyte, J., and R. F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105–132.
23. Rose, G. D., A. R. Geselowitz, …, M. H. Zehfus. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science.* 229:834–838.
24. Felitsky, D. J., M. A. Lietzow, …, P. E. Wright. 2008. Modeling transient collapsed states of an unfolded protein to provide insights into early folding events. *Proc. Natl. Acad. Sci. USA.* 105:6278–6283.
25. Garnier, J., J. F. Gibrat, and B. Robson. 1996. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* 266:540–553.
26. Iakoucheva, L. M., P. Radivojac, …, A. K. Dunker. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* 32:1037–1049.
27. Borg, M., T. Mittag, …, H. S. Chan. 2007. Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proc. Natl. Acad. Sci. USA.* 104:9650–9655.
28. Mittag, T., S. Orlicky, …, J. D. Forman-Kay. 2008. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc. Natl. Acad. Sci. USA.* 105:17772–17777.
29. Uversky, V. N., J. Li, and A. L. Fink. 2001. Evidence for a partially folded intermediate in α-synuclein fibril formation. *J. Biol. Chem.* 276:10737–10744.
30. Wu, K. P., D. S. Weinstock, …, J. Baum. 2009. Structural reorganization of α-synuclein at low pH observed by NMR and REMD simulations. *J. Mol. Biol.* 391:784–796.
31. McClendon, S., C. C. Rospigliosi, and D. Eliezer. 2009. Charge neutralization and collapse of the C-terminal tail of α-synuclein at low pH. *Protein Sci.* 18:1531–1540.
32. Soranno, A., R. Longhi, …, M. Buscaglia. 2009. Kinetics of contact formation and end-to-end distance distributions of swollen disordered peptides. *Biophys. J.* 96:1515–1528.
33. Adzhubei, A. A., and M. J. Sternberg. 1993. Left-handed polyproline II helices commonly occur in globular proteins. *J. Mol. Biol.* 229: 472–493.
34. Krishnan, V. V., E. Y. Lau, …, M. F. Rexach. 2008. Intramolecular cohesion of coils mediated by phenylalanine—glycine motifs in the natively unfolded domain of a nucleoporin. *PLOS Comput. Biol.* 4:e1000145.
35. Marsh, J. A., and J. D. Forman-Kay. 2009. Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints. *J. Mol. Biol.* 391:359–374.
36. Crowhurst, K. A., and J. D. Forman-Kay. 2003. Aromatic and methyl NOEs highlight hydrophobic clustering in the unfolded state of an SH3 domain. *Biochemistry.* 42:8687–8695.
37. Crowhurst, K. A., M. Tollinger, and J. D. Forman-Kay. 2002. Cooperative interactions and a non-native buried Trp in the unfolded state of an SH3 domain. *J. Mol. Biol.* 322:163–178.
38. Klein-Seetharaman, J., M. Oikawa, …, H. Schwalbe. 2002. Long-range interactions within a nonnative protein. *Science.* 295:1719–1722.
39. Chan, H. S., and K. A. Dill. 1989. Interchain loops in polymers: effects of excluded volume. *J. Chem. Phys.* 90:492–509.