Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues

Rahul K. Das and Rohit V. Pappu[1]

Department of Biomedical Engineering and Center for Biological Systems Engineering, Washington University in St. Louis, St. Louis, MO 63130

Edited by Robert L. Baldwin, Stanford University, Stanford, CA, and approved July 8, 2013 (received for review March 11, 2013)

The functions of intrinsically disordered proteins (IDPs) are governed by relationships between information encoded in their amino acid sequences and the ensembles of conformations that they sample as autonomous units. Most IDPs are polyampholytes, with sequences that include both positively and negatively charged residues. Accordingly, we focus here on the sequence–ensemble relationships of polyampholytic IDPs. The fraction of charged residues discriminates between weak and strong polyampholytes. Using atomistic simulations, we show that weak polyampholytes form globules, whereas the conformational preferences of strong polyampholytes are determined by a combination of fraction of charged residues values and the linear sequence distributions of oppositely charged residues. We quantify the latter using a patterning parameter κ that lies between zero and one. The value of κ is low for well-mixed sequences, and in these sequences, intrachain electrostatic repulsions and attractions are counterbalanced, leading to the unmasking of preferences for conformations that resemble either self-avoiding random walks or generic Flory random coils. Segregation of oppositely charged residues within linear sequences leads to high κ-values and preferences for hairpin-like conformations caused by long-range electrostatic attractions induced by conformational fluctuations. We propose a scaling theory to explain the sequence-encoded conformational properties of strong polyampholytes. We show that naturally occurring strong polyampholytes have low κ-values, and this feature implies a selection for random coil ensembles. The design of sequences with different κ-values demonstrably alters the conformational preferences of polyampholytic IDPs, and this ability could become a useful tool for enabling direct inquiries into connections between sequence–ensemble relationships and functions of IDPs.

Intrinsically disordered proteins (IDPs) feature prominently in proteins associated with transcriptional regulation and signal transduction (1, 2). IDPs fail to fold autonomously, their sequences are deficient in hydrophobic groups and enriched in polar and charged residues (3), and the thermodynamics and kinetics of coupled folding and binding are linked to the intrinsic conformational properties of IDPs (4–12).

IDP sequences include both types of charges, and at least 75% of known IDPs are polyampholytes (13). Coarse-grain parameters that are relevant for describing polyampholytes include the fraction of charged residues (FCR) and net charge per residue (NCPR), which are defined as FCR = $(f_+ + f_-)$ and NCPR = $|f_+ - f_-|$, where f_+ and f_- denote the fractions of positive and negatively charges, respectively. Polyampholytes are either strong (FCR $\geq$ 0.3) or weak (FCR $<$ 0.3) and can be neutral (NCPR $\sim$ 0) or have a net charge. Single molecule measurements have been used to measure the dimensions of three different polyampholytic systems (8), and a mean field model (14) that requires only FCR, NCPR, and the Debye length as inputs was successful in explaining the experimental data (8). NCPR also serves as an order parameter for predicting the distinction of polyelectrolytic IDPs into globules vs. swollen coils (7).

Can one predict the dimensions and internal structure of all polyampholytic IDPs using information regarding f_+ and f_- alone? Here, we answer this question by showing that NCPR is inadequate as a descriptor of sequence–ensemble relationships for a majority of IDPs, which are polyampholytes. Instead, FCR and sequence-specific distributions of oppositely charged residues are synergistic determinants of conformational properties of polyampholytic IDPs.

Quantitative studies of sequence–ensemble relationships of polyampholytic IDPs are important given the functions associated with them. Representative examples include the M domain of the yeast prion protein Sup35 (5), disordered stretches in RNA chaperones and splicing factors that undergo post-translational modifications (15), and Pro-Glu-Val-Lys (PEVK) stretches in the giant muscle protein titin (16). The outcomes of our investigations are germane to understanding the selection of specific patterns for linear sequence distributions of oppositely charged residues that are seen in polyampholytic IDPs. For example, is it important that the Glu and Lys residues essentially alternate within PEVK stretches of titin? Will changes to the linear sequence patterning of oppositely charged residues influence the passive elasticity of titin under physiologically relevant extensional forces? To be able to answer these types of questions, we present a framework for sequence–ensemble relationships of polyampholytic IDPs that is based on results from atomistic Metropolis Monte Carlo simulations. We use the ABSINTH (self-assembly of biomolecules studied by an implicit, novel, and tunable Hamiltonian) implicit solvation model and force field paradigm (17), a combination that has yielded verifiably accurate results for other IDPs (7, 18). We introduce a patterning parameter κ to distinguish between different sequence variants based on the linear sequence distributions of oppositely charged residues. We show that the types of conformations accessible to polyampholytes are governed by a combination of their κ- and FCR values. Finally, we introduce a scaling theory to explain the dependence of conformational properties on κ and show that de novo sequence design can be used to modulate sequence–ensemble relationships of polyampholytic IDPs.

Results

Parameter κ. A blob refers to the number of residues beyond which the balance of chain–chain, chain–solvent, and solvent–solvent energies is of order kT (19). Here, T denotes temperature, and k is Boltzmann's constant. The radius of gyration of a g residue blob scales as $g^{1/2}$, and for sequences lacking in proline residues, $g \sim 5$ (20). The overall charge asymmetry is defined as

Author contributions: R.K.D. and R.V.P. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1] To whom correspondence should be addressed. E-mail: pappu@wustl.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1304749110/-/DCSupplemental.

$\sigma = \frac{(f_+ - f_-)^2}{(f_+ + f_-)}$ (19). For each sequence variant, we calculate κ by partitioning the sequence into N_{blob} overlapping segments of size g. For each g residue segment, we calculate $\sigma_i = \frac{(f_+ - f_-)_i^2}{(f_+ + f_-)_i}$, which is the charge asymmetry for blob i in the sequence of interest. We quantify the squared deviation from σ as $\delta = \frac{\sum_{i=1}^{N_{\text{blob}}} (\sigma_i - \sigma)^2}{N_{\text{blob}}}$. Different sequence variants will have different values of δ, and the maximal value $\delta_{\max}$ for a given amino acid composition is used to define $\kappa = \left(\frac{\delta}{\delta_{\max}}\right)$, such that $0 \le \kappa \le 1$. We calculate κ using two values for the blob size: $g = 5$ and $g = 6$, and the final κ for a given sequence variant is an average of the two values.

Fig. 1 shows 30 sequence variants of the synthetic strong polyampholyte system (Glu-Lys)$_{25}$, for which $n = 50$, $f_+ = f_- = 0.5$, FCR = 1, and NCPR = σ = 0. The sequences in Fig. 1 span the range of κ-values, and *SI Appendix*, Table S1 summarizes predictions of their disorder tendencies. Low values of κ are realized for well-mixed sequence variants, and $\kappa \rightarrow 1$ if oppositely charged residues are segregated in the linear sequence. The number density of sequences $n(\kappa)$ with specific values of κ will be high for low κ-values and decrease as κ increases (*SI Appendix*, Fig. S1).

Conformational Properties for Sequence Variants of (Glu-Lys)$_{25}$ Vary Considerably with κ Despite Having Identical NCPR Values. Fig. 2 plots the ensemble averaged radii of gyration $\langle R_g \rangle$ for sequence variants of (Glu-Lys)$_{25}$ with different κ-values. In general, $\langle R_g \rangle$ decreases as κ increases. The linear Pearson correlation coefficient is $r = -0.83$ with a significance of $P = 6.1 \times 10^{-9}$. The $\langle R_g \rangle$ values exceed expectations for classical Flory random coils (~18 Å), and the smallest value of $\langle R_g \rangle$, obtained for $\kappa \rightarrow 1$, is greater by a factor of 1.6 than the value of 11 Å expected for a compact globule (21). Additionally, for well-mixed sequences, the $\langle R_g \rangle$ values are slightly larger than values expected for self-avoiding random walks (~28 Å).

Fig. 3 plots $\langle R_{ij} \rangle$, the ensemble-averaged interresidue distances against sequence separations $|j - i|$ for a subset of the sequence variants listed in Fig. 1 (*SI Appendix*, Fig. S2). These $\langle R_{ij} \rangle$ profiles quantify local concentrations of chain segments around each other and facilitate direct connections to measured pair distributions from small-angle X-ray scattering (22) and distance measurements from single molecule experiments (8). For $\kappa < 0.05$, $\langle R_{ij} \rangle$ increases monotonically with increasing $|j - i|$. For higher values of κ, the $\langle R_{ij} \rangle$ profiles show evidence of long-range electrostatic attractions between oppositely charged blocks. The conformational properties for sequences with low κ-values are, on average, similar to self-avoiding random walks, whereas sequences with high κ-values sample hairpin-like conformations. The effects of changes to solution conditions viz., salt concentration and temperature, are discussed in *SI Appendix*, Figs. S3–S5.

Label	Sequence	κ
sv1	EKEKEKEKEKEKEKEKEKEKEKEKEKEKEKEKEKEKEKEKEKEKEKEKEK	0.0009
sv2	EEEKKKEEEKKKEEEKKKEEEKKKEEEKKKEEEKKKEEEKKKEEEKKKEK	0.0025
sv3	KEKKKEKKEEKKEEKEKEKEKEEKKKEEKEKEKEKKKEEKEKEEKKEEEE	0.0139
sv4	KEKEKKEEKEKKEEEKKEKEKEKKKEEKKKEEKEEKKEEKKKEEKEEEKE	0.0140
sv5	KEKEEKEKKKEEEEKEKKKKEEKEKEKEKEEKKEEKKKKEEKEEKEKEKE	0.0245
sv6	EEEKKEKKEEKEEKKEKKEKEEEKKKEKEEKKEEEKKKEKEEEEKKKKEK	0.0273
sv7	EEEEKKKKEEEEKKKKEEEEKKKKEEEEKKKKEEEEKKKKEEEEKKKKEK	0.0450
sv8	KKKKEEEEKKKKEEEEKKKKEEEEKKKKEEEEKKKKEEEEKKKKEEEEKE	0.0450
sv9	EEKKEEEKEKEKEEEEEKKEKKEKKEKKKEEKEKEKKKEKKKKEKEEEKE	0.0624
sv10	EKKKKKKEEKKKEEEEEKKKEEEKKKEKKEEKEKEEKEKKEKKEEKEEEE	0.0834
sv11	EKEKKKKKEEEKKEKEEEEKEEEEKKKKKEKEEEKEEKKEEKEKKKEEKK	0.0841
sv12	EKKEEEEEEKEKKEEEEKEKEKKEKEEKEKKEKKKEKKEEEKEKKKKEKK	0.0864
sv13	KEKKKEKEKKEKKKEEEKKKEEEKEKKKEEKKEKKEKKEEEEEEEKEEKE	0.0951
sv14	EKKEKEEKEEEEKKKKKEEKEKKEKKKKEKKKKKEEEEEEKEEKEKEKEE	0.1311
sv15	KKEKKEKKKEKKEKKEEEKEKEKKEKKKKEKEKKEEEEEEEEKEEKKEEE	0.1354
sv16	EKEKEEKKKEEKKKKEKKEKEEKKEKEKEKKEEEEEEEEEKEKKEKKKKE	0.1458
sv17	EKEKKKKKKEKEKKKKEKEKKEKKEKEEEKEEKEKEKKEEKKEEEEEEEE	0.1643
sv18	KEEKKEEEEEEEKEEKKKKKEKKKEKKEEEKKKEEKKKEEEEEEKKKKEK	0.1677
sv19	EEEEEKKKKKEEEEEKKKKKEEEEEKKKKKEEEEEKKKKKEEEEEKKKKK	0.1941
sv20	EEKEEEEEEKEEEKEEKKEEEKEKKEKKEKEEKKEKKKKKKKKKKKKEEE	0.2721
sv21	EEEEEEEEEKEKKKKKEKEEKKKKKKEKKEKKKKEKKEEEEEEKEEEKKK	0.2737
sv22	KEEEEKEEKEEKKKKEKEEKEKKKKKKKKKKKKEKKEEEEEEEEKEKEEE	0.3218
sv23	EEEEEKEEEEEEEEEEEKEEKEKKKKKKEKKKKKKKEKEKKKKEKKEEKK	0.3545
sv24	EEEEKEEEEEKEEEEEEEEEEEEKKKEEKKKKKEKKKKKKKEKKKKKKKK	0.4456
sv25	EEEEEEEEEEEKEEEEKEEKEEKEKKKKKKKKKKKKKKKKKKEEKKEEKE	0.5283
sv26	KEEEEEEEKEEKEEEEEEEEEKEEEEKEEKKKKKKKKKKKKKKKKKKKKE	0.6101
sv27	KKEKKKEKKEEEEEEEEEEEEEEEEEEEEKEEKKKKKKKKKKKKKKKEKK	0.6729
sv28	EKKKKKKKKKKKKKKKKKKKKKEEEEEEEEEEEEEEEEEEKKEEEEEKEK	0.7666
sv29	KEEEEKEEEEEEEEEEEEEEEEEEEEEKKKKKKKKKKKKKKKKKKKKKKK	0.8764
sv30	EEEEEEEEEEEEEEEEEEEEEEEEEKKKKKKKKKKKKKKKKKKKKKKKKK	1.0000

Fig. 1. Thirty sequence variants for the (Glu-Lys)$_{25}$ system. Column 1 shows the label of each sequence variant. Column 2 shows the actual sequence, with Glu residues in red and Lys residues in blue. Column 3 shows the κ-values.

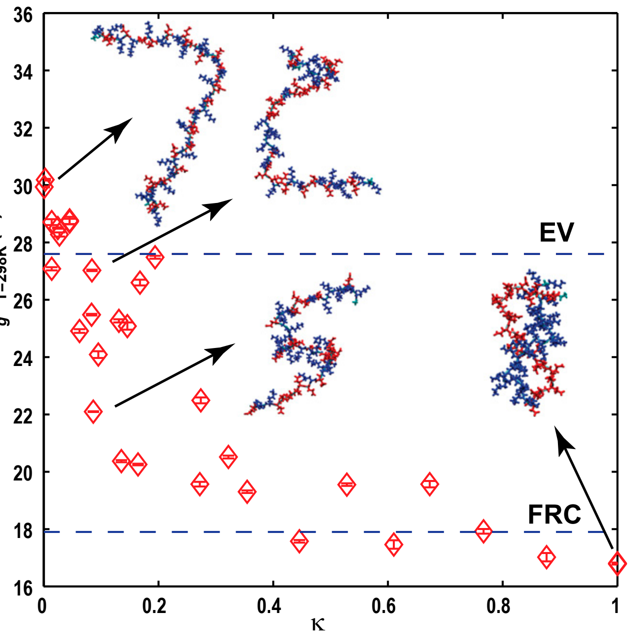


Fig. 2. Ensemble-averaged radii of gyration $<R_g>$ for sequence variants of the (Glu-Lys)$_{25}$ system. *Insets* show representative conformations for four sequence variants. Side chains of Glu are shown in red, and side chains of Lys are shown in blue. The two dashed lines intersect the ordinate at $<R_g>$ values expected for all sequence variants of the (Glu-Lys)$_{25}$ system modeled in the EV limit or as Flory random coils (FRCs).

SI Appendix, Fig. S6 plots the asphericity (δ^*) of each sequence variant against κ. For perfect spheres, $\delta^* \sim 0$ and $\delta^* \sim 1$ for rods (23). As κ increases, the asphericity values decrease from ~0.6 to ~0.2. This decrease in asphericity is consistent with a transition from elongated prolate ellipsoids to semicompact hairpins as illustrated in *SI Appendix*, Fig. S7, which shows representative conformations for different sequence variants of (Glu-Lys)$_{25}$.

Phenomenological Explanation for the κ-Dependence of Conformational Properties. In our atomistic simulations, the potential energy U_c associated with a specific conformation c is a sum of terms (i.e., $U_c = U_{\text{EV}} + U_{\text{Disp}} + U_{\text{tor}} + W_{\text{Solv}} + W_{\text{el}}$). Here, U_{tor} denotes torsional potentials; $U_{\text{EV}} + U_{\text{Disp}}$ models van der Waals interactions using the Lennard–Jones model, where U_{EV} and U_{Disp} refer to short-range repulsive and attractive dispersion terms, respectively. W_{Solv} quantifies the conformation-specific free energy of solvation using the ABSINTH model; W_{el} models the effects of changes to the degrees of solvation that lead to conformation-specific descreening of intrachain electrostatic interactions. This term captures the effects of solvent-mediated electrostatic interactions between all charged groups, including charged side chains, partial charges that lead to backbone and side chain hydrogen bonding, and electrostatic interactions involving mobile ions in solution.

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

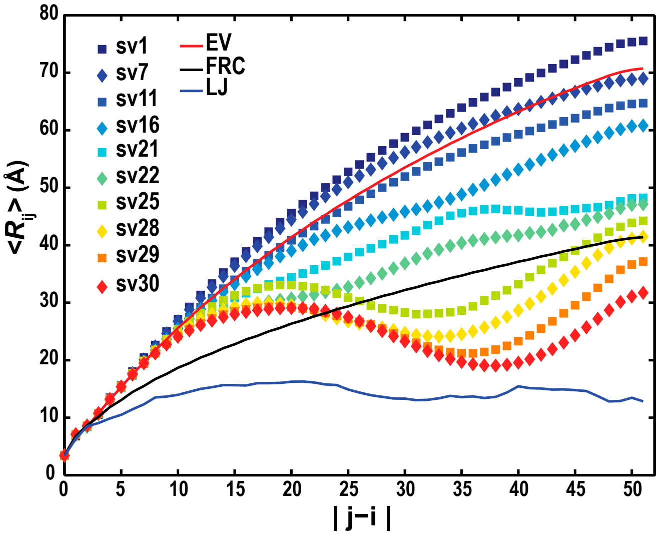


Fig. 3. $<R_{ij}>$ profiles for sequence variants of the (Glu-Lys)$_{25}$ system. The red curve denotes the profile expected for (Glu-Lys)$_{25}$ polymers in the EV limit. The black curve is expected for an FRC, and the solid blue curve is obtained when (Glu-Lys)$_{25}$ polymers form maximally compact globules. For compact globules, $<R_{ij}>$ plateaus to a value that is prescribed by their densities.

If all terms excepting U_{EV} are zeroed out, then self-avoiding random walk distributions result, because the polypeptide samples conformations from the excluded volume (EV) limit. When the ensemble-averaged effects of intrachain electrostatic attractions and repulsions are counterbalanced, the underlying EV limit behavior is unmasked, which is the case for low κ-variants of (Glu-Lys)$_{25}$ (Fig. 3). For short sequence separations ($|j-i| < 2g$), there are not enough intrachain electrostatic interactions to perturb chain statistics away from the EV limit. The $\langle R_{ij}\rangle$ profiles for short separations should, therefore, resemble the profiles of unperturbed self-avoiding random walks. For sequences with higher κ-values, there should be a range of intermediate sequence separations ($2g \le |j-i| \le l_c$), where oppositely charged blocks act as counterion clouds for each other, leading to electrostatic attractions induced by conformational fluctuations. Here, g is the blob length, and l_c is the length scale over which deviations from the EV limit occur. The resultant semicompact hairpin-like or partial hairpin-like conformations will cause the $\langle R_{ij}\rangle$ profiles to deviate from the profiles of chains in the EV limit. The degree of this deviation will depend on κ. Finally, for sequence separations greater than l_c, the strength of the ensemble-averaged electrostatic attractions is $\sim kT$, and the EV limit behavior is recovered.

Development of a Scaling Theory for $\langle R_{ij}\rangle$. Based on the preceding discussion, we propose that the variation of conformational properties for different κ-variants of (Glu-Lys)$_{25}$ can be modeled using a scaling theory akin to the theory in the work by Yamakov et al. (24). We use the EV limit distribution as the reference state as justified for (Glu-Lys)$_{25}$ in *SI Appendix*, Fig. S8. We write $\langle R_{ij}\rangle$ for all sequence separations of a given sequence as $\langle R_{ij}\rangle = R_0^{\mathrm{EV}} f(|j-i|)|j-i|^{\nu}$. Here, $R_0^{\mathrm{EV}} \approx 7.0\text{Å}$ is the nonuniversal prefactor that describes the scaling of $\langle R_{ij}\rangle$ for (Glu-Lys)$_{25}$ polymers as a function of $|j-i|$ in the EV limit. The exponent $\nu = 0.59$ is universal and prescribes the correlation length for polymers in the EV limit (25). The scaling function $f(|j-i|)$ describes deviations from the EV limit that result from unbalanced electrostatic interactions. The form for $f(|j-i|)$ derived from analysis of the $\langle R_{ij}\rangle$ profiles for (Glu-Lys)$_{25}$ variants is

$$\begin{aligned} f(|j-i|) &= 1 && \text{if} \quad |j-i| < 2g \\ f(|j-i|) &= p_0 + p_1|j-i| && \text{if} \quad 2g \le |j-i| \le l_c \\ f(|j-i|) &= \left(1-\frac{l_c}{|j-i|}\right)^{\nu} && \text{if} \quad |j-i| > l_c \end{aligned} . \qquad [1]$$

Results from numerical fits to the $\langle R_{ij}\rangle$ profile for sv30 of (Glu-Lys)$_{25}$ using the scaling theory are shown in Fig. 4, and results for all other sequence variants are shown in *SI Appendix*, Fig. S9. The coefficients p_0 and p_1 quantify the intercept and slope for the linear interpolation between the two regimes that show EV limit-like behavior. The values of p_1 quantify the deviations from the EV limit profiles and are either small ($p_1 \sim 0$ for low κ) or negative as κ increases (*SI Appendix*, Fig. S10). The intercept p_0 quantifies corrections to the excluded volume per residue that result from the effects of electrostatic interactions. The form for $f(|j-i|)$ for $|j-i| > l_c$ implies that sequence separations between distal segments that restore EV limit behavior are renormalized to smaller effective separations, thus giving rise to continuous transitions between the regime where deviations are caused by intrachain electrostatic interactions and the EV limit.

On the Choice of Reference State for the Scaling Theory. Our choice of the EV limit as the reference state for the scaling theory was based on the observation that counterbalancing of electrostatic repulsions and attractions unmasks EV limit behavior for well-mixed sequence variants of (Glu-Lys)$_{25}$. In systems with smaller values of FCR, the counterbalancing in well-mixed sequence variants might unmask a different reference state, such as the Flory random coil. The precise form of the reference state that is unmasked by counterbalancing of electrostatic repulsions and attractions in well-mixed sequences will depend on the preferences encoded by the collective contributions of the non-electrostatic terms of the potential function. Accordingly, we introduce an intrinsic solvation (IS) limit, whereby simulations to generate the reference state are performed using all terms of the potential function except W_{el}. Comparison of simulation results obtained using the full Hamiltonian with the results of the IS limit allows us to unmask the κ-specific contributions that arise because of competition between intrachain electrostatic attractions and repulsions. The free energies of solvation of charged

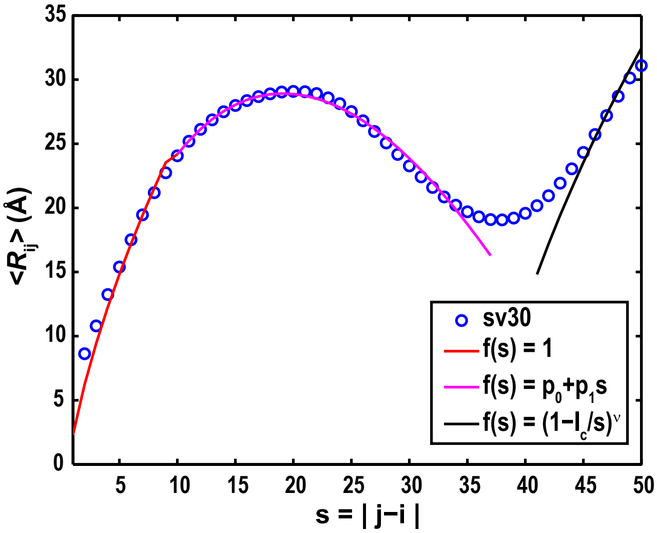


Fig. 4. Numerical fits to the $<R_{ij}>$ profile for sequence variant sv30 of the (Glu-Lys)$_{25}$ system. The red, magenta, and black curves correspond to three distinct regimes viz.: $|j-i| < 2g$, $2g \le |j-i| \le l_c$, and $|j-i| > l_c$, respectively.

side chains are highly favorable (∼−100 kcal/mol), and for high FCR, the IS limit ensembles are qualitatively similar to the ensembles of the EV limit, which are shown in *SI Appendix*, Fig. S11 for sequence variants of (Glu-Lys)$_{25}$. However, as FCR decreases, there is good reason to expect significant deviation of $\langle R_{\mathrm{ij}}\rangle$ profiles calculated in the IS limit from those profiles of the EV limit (which will be shown below). Therefore, for sequences with FCR < 1, the development of a general form of the scaling relation for $\langle R_{\mathrm{ij}}\rangle$ will require that we use the appropriate IS limit profiles as reference models.

Inferring Deviations from Limiting Behavior from Sequence. The presence of unbalanced intrachain electrostatic interactions can be assessed from sequence information if one computes the dimensionless Coulomb coupling parameter Γ_{ij} (26). For a pair of blobs i and j, $\Gamma_{\mathrm{ij}} = \frac{\langle z_i z_j\rangle}{4\pi\varepsilon_0 \varepsilon RT\xi}$; ε = 78 is the dielectric constant of water at 298 K, ε_0 is the permittivity of free space, ξ = 6 Å is the radius of a blob (*SI Appendix*, Fig. S12), R is the ideal gas constant, T is the temperature, and z_{i} and z_{j} denote the signed NCPR values of blobs i and j, respectively. The product $z_{\mathrm{i}}z_{\mathrm{j}}$ is positive or negative depending on whether the signed NCPR values for blobs i and j are of similar or opposite signs. For a given sequence variant, we calculate the product $z_{\mathrm{i}}z_{\mathrm{j}}$ for all pairs of blobs i and j that satisfy the constraint $|\mathrm{j} - \mathrm{i}| > g = 5$, and Γ_{ij} is computed by averaging over $z_{\mathrm{i}}z_{\mathrm{j}}$ values for all pairs of blobs corresponding to a linear separation of $|\mathrm{j} - \mathrm{i}|$.

SI Appendix, Fig. S13 plots the cumulative sum of $\Omega_k = \sum_{k=1}^{|j-i|} < \Gamma_k >$ against the linear separation between pairs of blobs. Of interest are the length scales for which Ω_k is negative with a magnitude larger than RT. *SI Appendix*, Fig. S14 in the *SI Appendix* quantifies the correlation between p and $\mathbf{min}(\Omega_k)$. This plot shows that the two parameters show significant positive correlation (Pearson r = 0.79). To a first approximation, if we neglect the small contributions of p_{o} and use the equation for the line of best fit that relates p_1 to $\mathbf{min}(\Omega_k)$, we can obtain qualitative assessments of the degree to which electrostatic attractions will lead to a deviation of the $\langle R_{\mathrm{ij}}\rangle$ profile from a reference state, such as the EV limit.

Results for Naturally Occurring Polyampholytic IDPs. *SI Appendix*, Table S2 summarizes information regarding 10 IDP sequences extracted from a combination of the DisProt database (13) and published experimental data. For these sequences, 0.14 ≤ FCR ≤ 0.73, and 0.0 ≤ NCPR ≤ 0.25. *SI Appendix*, Fig. S15 shows the $\langle R_{\mathrm{ij}}\rangle$ profiles for these sequences in the IS limit. These reference state profiles are between the profiles for the EV limit and the Flory random coil, with the general trend of converging on the latter as FCR decreases. The critical exponent quantifying the correlation length switches from ν = 0.59 in the EV limit to ν = 0.5 for the Flory random coil. Profiles bearing similarity to the latter are realized for polymers in θ-solvents, where the statistical effects of intrachain and chain-solvent interactions are counterbalanced (27, 28).

Fig. 5 shows $\langle R_{\mathrm{ij}}\rangle$ profiles from simulation results obtained using the full ABSINTH Hamiltonian for all 10 sequences. Comparisons of these profiles with their respective IS limit profiles are shown in *SI Appendix*, Fig. S16. The contributions of intrachain, solvent-mediated electrostatic interactions lead to either weak perturbations from the IS limit, which was seen for polyglutamine tract binding protein (PQBP-1), DP00166, DP00357, DP00503, and QSH22, or significant compaction vis-à-vis the IS limit, which was seen for the remaining five sequences. The extent of the perturbation with respect to the IS limit is clearly governed by FCR. Hofmann et al. (28) have recently shown that the degree of deviation of unfolded state dimensions from an effective θ-state as measured under folding conditions is also dependent on FCR.

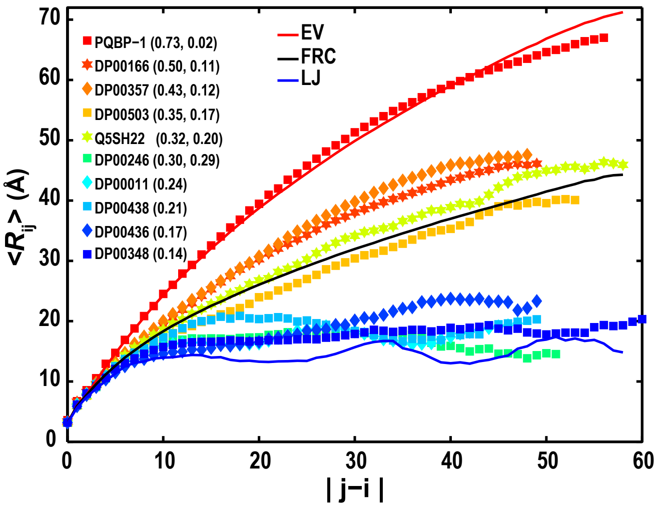


Fig. 5. <R_{ij}> profiles for 10 naturally occurring IDPs. The legend shows the DisProt or other identifier for each sequence. The solid curves are reference profiles that are similar to those profiles described in Fig. 3. The legend shows the sequence identifiers and the combination of FCR and κ-values in parentheses. For globule formers, the values of κ have no significance, and for these sequences, the legend shows only their FCR values.

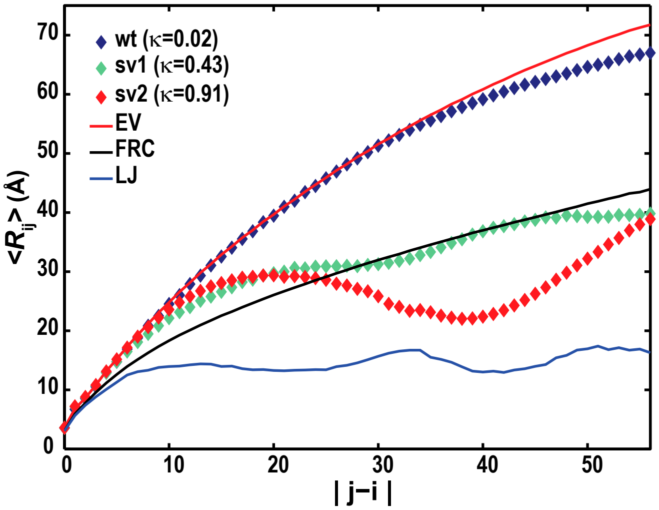
Sequences with FCR < 0.3 and NCPR values < 0.25 are weak polyampholytes, and compaction results from decreased FCR with charged residues on the surfaces of globules (*SI Appendix*, Fig. S17). *SI Appendix*, Fig. S18 shows the temperature dependence of $\langle R_{\mathrm{g}}^2\rangle$ values for the 10 naturally occurring polyampholytic IDPs from *SI Appendix*, Table S2. These results show that the conformational properties for polyampholytes with lower FCR values show more pronounced temperature dependencies compared with sequence variants of (Glu-Lys)$_{25}$.

Conformational Properties of Polyampholytic IDPs Can Be Modulated Through de Novo Sequence Design. The N-terminal end of the PQBP-1 includes a WW domain that binds RNA polymerase II and is connected to the C-terminal U5 15 kDa binding region (29) by a polyampholytic stretch. Multiple lines of experimental evidence suggest that this polyampholytic stretch is a flexible tether that adopts expanded conformations (29, 30). Fig. 6 shows the $\langle R_{\mathrm{ij}}\rangle$ profile for the 55-residue construct WPP-(PQBP-1)$_{132-183}$, for which FCR = 0.73, NCPR = 0, and κ = 0.024. We reasoned that high κ-variants of this sequence should have very different conformational properties. We tested this hypothesis by comparing the conformational properties of the WT sequence with the properties of two variants with higher κ-values (Fig. 6). The results show considerable differences between the $\langle R_{\mathrm{ij}}\rangle$ profile of the WT sequence and its higher κ-variants, such that changes of ∼28 Å in the end-to-end distance can be achieved by sequence permutations. For a fixed amino acid composition, systems with the designation of strong polyampholytes are likely to have higher designability than weak polyampholytes, because significant modulation of conformational properties is achievable by varying κ.

Discussion

Mao et al. (7) proposed a predictive diagram of states, whereby the ensemble type (namely globule or coil) can be inferred based on the NCPR value for a given sequence. We annotated this diagram of states using a subset of IDP sequences from the DisProt database (13). Approximately 95% of these sequences have amino acid compositions with NCPR < 0.25, which would imply that they form compact globules (*SI Appendix*, Fig. S19). However, this inference is questionable, because most of the

BIOPHYSICS AND COMPUTATIONAL BIOLOGY



wt WPPDRGHDKSDRDRERGYDKVDRERERDRERDRDRGYDKADREEGKERRHHRREE
sv1 WPPYDDRSRHERRHKYRRRRARKRHKGDREEGEEDVEDEDGDRRRRRKDDDDEGE
sv2 WPPGGEDDEDDDDEEDDEGEDEDEDEAHYYKSHGRRRRKKRRKRRRHRRRRRRVR

Fig. 6. <R_{ij}> profiles for the WT linker from PQBP-1 and two designed sequence variants. The sequences of the WT stretch and sequence permutants are shown. The solid red, black, and blue curves correspond to <R_{ij}> profiles for WPP-(PQBP-1)$_{132–183}$:wt simulated in the EV limit, FRC, and compact Lennard–Jones (LJ) globules, respectively.

sequences annotated as being globule formers are, in fact polyampholytes. If NCPR alone was a sufficient descriptor of conformational properties, then the results of Figs. 2, 3, and 5 would have been consistent with globule formation, irrespective of the κ- and FCR values, which is clearly not the case. We modified the original diagram of states to account for the findings from this work. In the modified diagram of states (Fig. 7), ∼70% of the IDPs that were classified as globules (*SI Appendix*, Fig. S19) are found to have compositions that place them in either the strong polyampholytic region or the boundary between globules and strong polyampholytes. Sequences within the boundary are distinct from globule formers and strong polyampholytes. Inferring their sequence–ensemble relationships requires additional considerations, such as the compositions of polar residues, the proline contents, and the presence of sequence stretches with preferences for specific secondary structures.

Assessing Polyampholyte Theories. Mean field theories for polyampholytes describe the dependence of R_g and internal structure on values of FCR, NCPR, and N (14, 19, 31, 32). These theories predict that neutral polyampholytes will form globules with liquid-like organization of opposite charges within the interior of globules that resembles globules of 1:1 electrolytes. Alternative predictions suggest more EV limit-like behavior (33). Our results contradict the predictions of typical mean field theories because of two weaknesses in the theories. First, they apply to an ensemble-averaged sequence, which is obtained by averaging over all possible sequence variants for a given FCR and NCPR (32). Therefore, they cannot work for individual sequence variants (34, 35). Second, all theories ignore the effects of highly favorable solvation free energies of charged groups, which clearly require fundamentally different reference states, such as the IS limit.

We have presented a preliminary scaling theory to account for the effects of κ-specific correlations in sequence variants of (Glu-Lys)$_{25}$. The theory is based on the observation that counterbalancing of electrostatic attractions and repulsions in well-mixed sequence variants of (Glu-Lys)$_{25}$ unmasks conformational preferences obtained in the EV limit. For well-mixed variants of weaker polyampholytes (0.3 ≤ FCR < 1), counterbalancing of electrostatic attractions and repulsions will unmask the IS limit as the relevant reference state. Consequently, for polyampholytes with 0.3 ≤ FCR < 1 that show κ-specific conformational properties, an extension of the scaling theory might simply require switching the reference critical exponent from $\nu = 0.59$ to $\nu = 0.5$. However, for globule-forming weak polyampholytes (FCR < 0.3), the collapse becomes weakly dependent or even independent of κ. Inasmuch as the IS limit resembles the Flory random coil or effective θ-state, a theoretical framework to describe the collapse of weak polyampholytes will likely resemble the framework of theories for coil-to-globule transitions (36). Large-scale simulations performed using different combinations of FCR, NCPR, and κ and integration of these results should yield a unifying theoretical framework for sequences that span the spectrum of FCR values. This task seems practicable and will be pursued in future work.

Broader Implications. *SI Appendix*, Fig. S20 shows the joint distribution of FCR and κ-values for strong polyampholytic IDPs extracted from the DisProt database. The distribution is peaked around κ ∼ 0.2, implying that naturally occurring sequences are reasonably well-mixed and likely to have conformational properties that are between the EV limit and Flory random coil models. If an IDP is a strong polyampholyte, then posttranslational

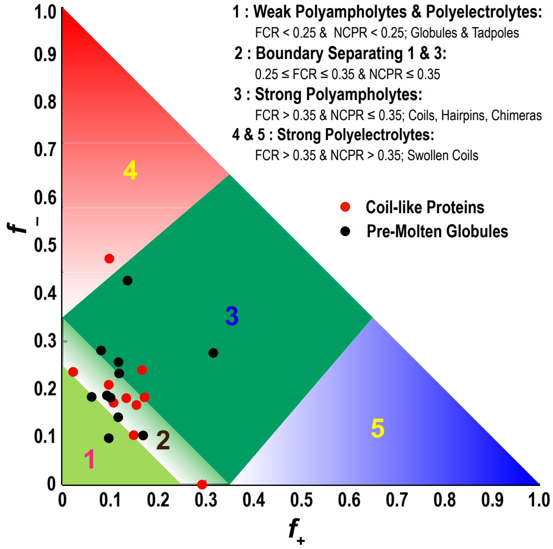


Fig. 7. Diagram of states for IDPs. We focus on sequences that fall below the parameterized line (NCPR = 2.785H − 1.151), developed by Uversky et al. (43) to separate IDPs from sequences that fold autonomously. Here, H refers to the hydropathy score. Region 1 corresponds to either weak polyampholytes or weak polyelectrolytes that form globule or tadpole-like conformations (*SI Appendix*, Fig. S17). Region 3 corresponds to strong polyampholytes that form distinctly nonglobular conformations that are coil-like, hairpin-like, or admixtures. A boundary region labeled 2 separates regions 1 and 3, and the conformations within this region are likely to represent a continuum of possibilities between the types of conformations adopted by sequences in regions 1 and 3. Sequences with compositions corresponding to regions 4 and 5 are strong polyelectrolytes with FCR > 0.35 and NCPR > 0.3. These sequences are expected to sample coil-like conformations that largely resemble EV limit ensembles. The legend summarizes statistics for different regions based on sequences drawn from the DisProt database. The figure includes annotation by properties of sequences that have been designated as being "coils" or "pre-molten-globules" by Uversky (3) based on measurements of hydrodynamic radii. These sequences (listed in *SI Appendix*, Tables S3 and S4) are expected to be expanded vis-à-vis folded proteins, and our annotation shows that, indeed, all but one of the sequences is outside the globule-forming region.

modification, such as Ser/Thr phosphorylation, can increase FCR and NCPR and lead to coil-like properties (37). If phosphorylation converts an IDP from a polyelectrolyte to a strong polyampholyte (38), then the conformational properties will be governed by the combination of FCR and κ for the modified sequence. The sequences of IDPs can also be altered by alternative splicing (39), and for polyampholytic IDPs, the effects of splicing will give rise to altered sequence–ensemble relationships on the protein level. Therefore, posttranscriptional and posttranslational regulations seem to afford tuning of sequence–ensemble relationships of IDPs (40)—a feature that is enabled by the predominantly polyampholytic nature of these proteins.

Materials and Methods

Simulations were performed using the CAMPARI package using the ABSINTH implicit solvation model and force-field paradigm (17) (http://campari.sourceforge.net/). Parameters were taken from the abs3.2_opls.prm file. Conformational space for each IDP was sampled using Markov Chain Metropolis Monte Carlo moves that were combined with thermal replica exchange (41) to enhance the quality of sampling. Neutralizing ions and excess Na^+ and Cl^- ions were modeled explicitly to mimic a concentration of 15 or 125 mM in spherical droplets of 75 Å radius. Details of the simulation setup, including move sets used, temperature schedules, choices for droplet size, treatment of long-range interactions, and analysis methods, are provided in *SI Appendix*, Section 2. We report results from simulations for 42 sequence variants; the shortest was 46 residues long, and the longest has 59 residues. This level of throughput is essential to unmask how FCR and κ determine sequence–ensemble relationships. We have documented the intractability of using explicit solvent models for large-scale simulations of highly charged systems (7), because we require robust statistics regarding excursions into and out of expanded/compact conformations without the confounding effects of finite-sized artifacts (42) and artificial confinement imposed by the use of small periodic systems.

ACKNOWLEDGMENTS. We thank Scott Crick, Alex Holehouse, Nicholas Lyle, Albert Mao, and Anuradha Mittal for helpful discussions. This work was supported by National Science Foundation Grant MCB-1121867 and the Center for High Performance Computing at Washington University in St. Louis.

1. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6(3):197–208.
2. Tantos A, Han KH, Tompa P (2012) Intrinsic disorder in cell signaling and gene transcription. *Mol Cell Endocrinol* 348(2):457–465.
3. Uversky VN (2002) What does it mean to be natively unfolded? *Eur J Biochem* 269(1): 2–12.
4. Bright JN, Woolf TB, Hoh JH (2001) Predicting properties of intrinsically unstructured proteins. *Prog Biophys Mol Biol* 76(3):131–173.
5. Mukhopadhyay S, Krishnan R, Lemke EA, Lindquist S, Deniz AA (2007) A natively unfolded yeast prion monomer adopts an ensemble of collapsed and rapidly fluctuating structures. *Proc Natl Acad Sci USA* 104(8):2649–2654.
6. Wells M, et al. (2008) Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc Natl Acad Sci USA* 105(15):5762–5767.
7. Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu RV (2010) Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc Natl Acad Sci USA* 107(18):8183–8188.
8. Müller-Späth S, et al. (2010) From the Cover: Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc Natl Acad Sci USA* 107(33): 14609–14614.
9. Marsh JA, Forman-Kay JD (2010) Sequence determinants of compaction in intrinsically disordered proteins. *Biophys J* 98(10):2383–2390.
10. Potoyan DA, Papoian GA (2011) Energy landscape analyses of disordered histone tails reveal special organization of their conformational dynamics. *J Am Chem Soc* 133(19): 7405–7415.
11. Zhang WH, Ganguly D, Chen JH (2012) Residual structures, conformational fluctuations, and electrostatic interactions in the synergistic folding of two intrinsically disordered proteins. *PLoS Comput Biol* 8(1):e1002353.
12. Mao AH, Lyle N, Pappu RV (2013) Describing sequence-ensemble relationships for intrinsically disordered proteins. *Biochem J* 449(2):307–318.
13. Sickmeier M, et al. (2007) DisProt: The database of disordered proteins. *Nucleic Acids Res* 35(Database issue):D786–D793.
14. Higgs PG, Joanny JF (1991) Theory of polyampholyte solutions. *J Chem Phys* 94(2): 1543–1554.
15. Fu XD (1995) The superfamily of arginine/serine-rich splicing factors. *RNA* 1(7): 663–680.
16. Forbes JG, et al. (2005) Titin PEVK segment: Charge-driven elasticity of the open and flexible polyampholyte. *J Muscle Res Cell Motil* 26(6–8):291–301.
17. Vitalis A, Pappu RV (2009) ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J Comput Chem* 30(5):673–699.
18. Das RK, Crick SL, Pappu RV (2012) N-terminal segments modulate the α-helical propensities of the intrinsically disordered basic regions of bZIP proteins. *J Mol Biol* 416(2):287–299.
19. Dobrynin AV, Rubinstein M (1995) Flory theory of a polyampholyte chain. *Journale de Physique II France* 5(5):677–695.
20. Pappu RV, Wang X, Vitalis A, Crick SL (2008) A polymer physics perspective on driving forces and mechanisms for protein aggregation. *Arch Biochem Biophys* 469(1): 132–141.
21. Dima RI, Thirumalai D (2004) Asymmetry in the shapes of folded and denatured states of proteins. *J Phys Chem B* 108(21):6564–6570.
22. Bernadó P, Svergun DI (2012) Analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Methods Mol Biol* 896:107–122.
23. Steinhauser MO (2005) A molecular dynamics study on universal properties of polymer chains in different solvent qualities. Part I. A review of linear chain properties. *J Chem Phys* 122(9):094901.
24. Yamakov V, et al. (2000) Conformations of random polyampholytes. *Phys Rev Lett* 85(20):4305–4308.
25. Schäfer L (1999) *Excluded Volume Effects in Polymer Solutions as Explained by the Renormalization Group* (Springer, Berlin).
26. Tanaka M, Tanaka T (2000) Clumps of randomly charged polymers: Molecular dynamics simulation of condensation, crystallization, and swelling. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 62(3 Pt B):3803–3816.
27. Nettels D, et al. (2009) Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proc Natl Acad Sci USA* 106(49):20740–20745.
28. Hofmann H, et al. (2012) Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc Natl Acad Sci USA* 109(40):16155–16160.
29. Takahashi M, et al. (2010) Polyglutamine tract-binding protein-1 binds to U5-15kD via a continuous 23-residue segment of the C-terminal domain. *Biochim Biophys Acta* 1804(7):1500–1507.
30. Rees M, et al. (2012) Solution model of the intrinsically disordered polyglutamine tract-binding protein-1. *Biophys J* 102(7):1608–1616.
31. Edwards SF, King PR, Pincus P (1980) Phase-changes in polyampholytes. *Ferroelectrics* 30(1–4):3–6.
32. Dobrynin AV, Colby RH, Rubinstein M (2004) Polyampholytes. *J Polym Sci B* 42(19): 3513–3538.
33. Kantor Y, Kardar M (1995) Randomly charged polymers: An exact enumeration study. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 52(1):835–846.
34. Gutin AM, Shakhnovich EI (1994) Effect of a net charge on the conformation of polyampholytes. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 50(5): R3322–R3325.
35. Srivastava D, Muthukumar M (1996) Sequence dependence of conformations of polyampholytes. *Macromolecules* 29(6):2324–2326.
36. Sanchez IC (1979) Phase transition behavior of the isolated polymer chain. *Macromolecules* 12(5):980–988.
37. Borg M, et al. (2007) Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proc Natl Acad Sci USA* 104(23):9650–9655.
38. Kumar S, Hoh JH (2004) Modulation of repulsive forces between neurofilaments by sidearm phosphorylation. *Biochem Biophys Res Commun* 324(2):489–496.
39. Buljan M, et al. (2012) Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell* 46(6):871–883.
40. Vuzman D, Levy Y (2012) Intrinsically disordered regions as affinity tuners in protein-DNA interactions. *Mol Biosyst* 8(1):47–57.
41. Mitsutake A, Sugita Y, Okamoto Y (2003) Replica-exchange multicanonical and multicanonical replica-exchange Monte Carlo simulations of peptides. II. Application to a more complex system. *J Chem Phys* 118(14):6676–6688.
42. Chen AA, Marucho M, Baker NA, Pappu RV (2009) Simulations of RNA interactions with monovalent ions. *Methods Enzymol* 469:411–432.
43. Uversky VN, Gillespie JR, Fink AL (2000) Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* 41(3):415–427.