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 $\delta = \frac{\sum_{i=1}^{N_{\mathrm{blob}}} (\sigma_i - \sigma)^2}{N_{\mathrm{blob}}}$. 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 $\kappa \to 1$ 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 $\kappa \to 1$, 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	EEKKEEEKEKEKEEEEEKKEKKEKKEKKKEEKEKEKKKEKKKKEKEEEEKE	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	KEEEEKEEEEEEEEEEEEEEEEEEEEKKKKKKKKKKKKKKKKKKKKKKKK	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.

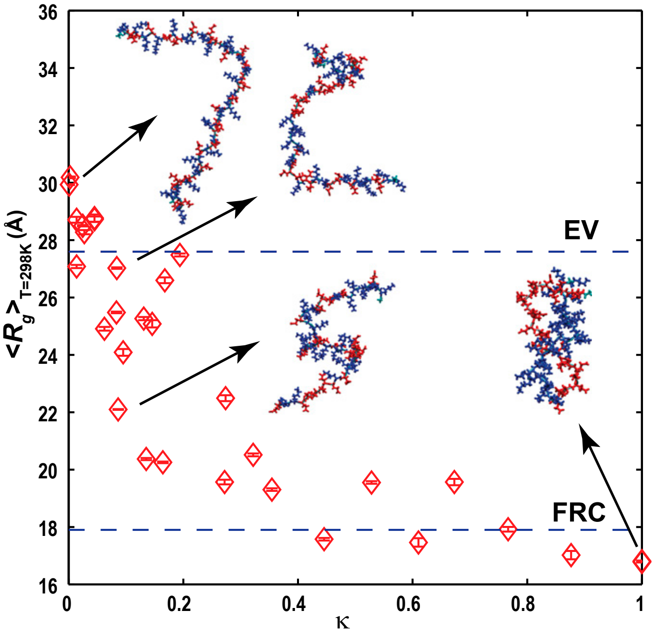

Fig. 2. Ensemble-averaged radii of gyration $<R_g>$ for sequence variants of the (Glu-Lys)$_{25}$ system. *Insets* show representative conformations for four sequence variants. Side chains of Glu are shown in red, and side chains of Lys are shown in blue. The two dashed lines intersect the ordinate at $<R_g>$ 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., $U_c = U_{EV} + U_{Disp} + U_{tor} + W_{Solv} + W_{el}$). Here, U_{tor} denotes torsional potentials; $U_{EV} + U_{Disp}$ 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.

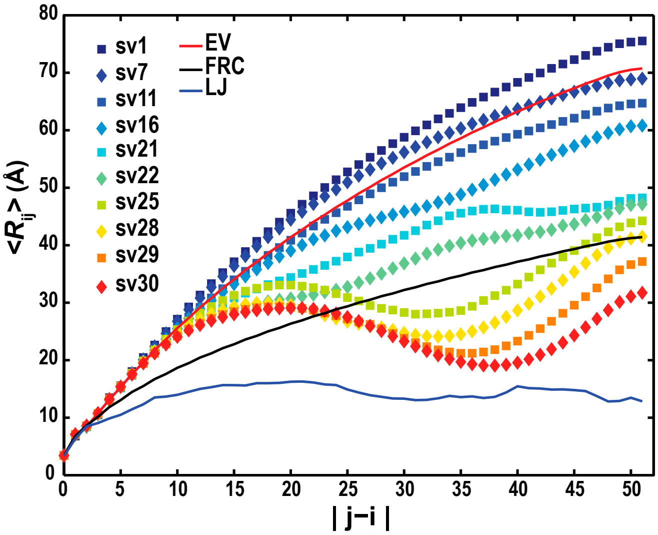

Fig. 3. <R_{ij}> profiles for sequence variants of the (Glu-Lys)$_{25}$ system. The red curve denotes the profile expected for (Glu-Lys)$_{25}$ polymers in the EV limit. The black curve is expected for an FRC, and the solid blue curve is obtained when (Glu-Lys)$_{25}$ polymers form maximally compact 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 ~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 $\langle R_{ij}\rangle = R_0^{EV} f(|j-i|)|j-i|^{\nu}$. Here, $R_0^{EV} \approx 7.0$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

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

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 ~ 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 nonelectrostatic 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

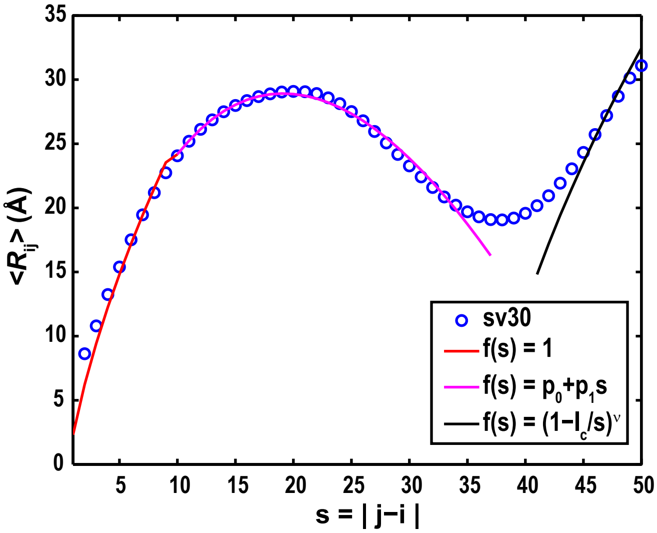

Fig. 4. Numerical fits to the <R_{ij}> profile for sequence variant sv30 of the (Glu-Lys)$_{25}$ system. The red, magenta, and black curves correspond to three distinct regimes viz.: $|j - i| < 2g$, $2g \le |j - i| \le l_c$, and $|j - i| > l_c$, respectively.

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, $\Gamma_{\mathrm{ij}} = \frac{\langle z_{\mathrm{i}} z_{\mathrm{j}} \rangle}{4\pi\varepsilon_0 \varepsilon RT\xi}$; $\varepsilon = 78$ is the dielectric constant of water at 298 K, ε_0 is the permittivity of free space, $\xi = 6$ 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 $\nu = 0.59$ in the EV limit to $\nu = 0.5$ 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.

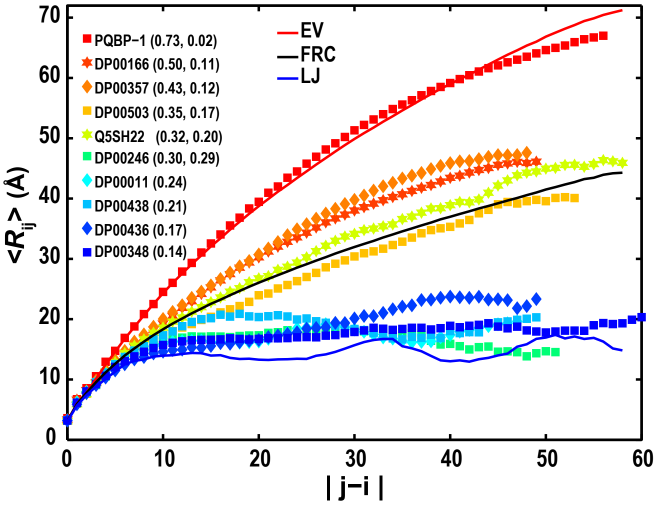

Fig. 5. $<R_{\mathrm{ij}}>$ profiles for 10 naturally occurring IDPs. The legend shows the DisProt or other identifier for each sequence. The solid curves are reference profiles that are similar to those profiles described in Fig. 3. The legend shows the sequence identifiers and the combination of FCR 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 $\kappa = 0.024$. 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

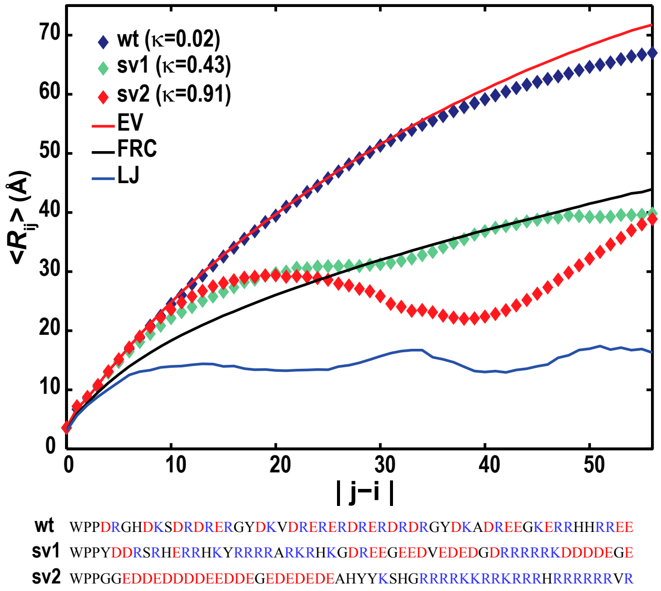

Fig. 6. <R_{ij}> profiles for the WT linker from PQBP-1 and two designed sequence variants. The sequences of the WT stretch and sequence permutants are shown. The solid red, black, and blue curves correspond to <R_{ij}> profiles for WPP-(PQBP-1)$_{132–183}$:wt simulated in the EV limit, FRC, and compact 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 ($0.3 \leq$ FCR < 1), counterbalancing of electrostatic attractions and repulsions will unmask the IS limit as the relevant reference state. Consequently, for polyampholytes with $0.3 \leq$ 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 $\kappa \sim 0.2$, 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

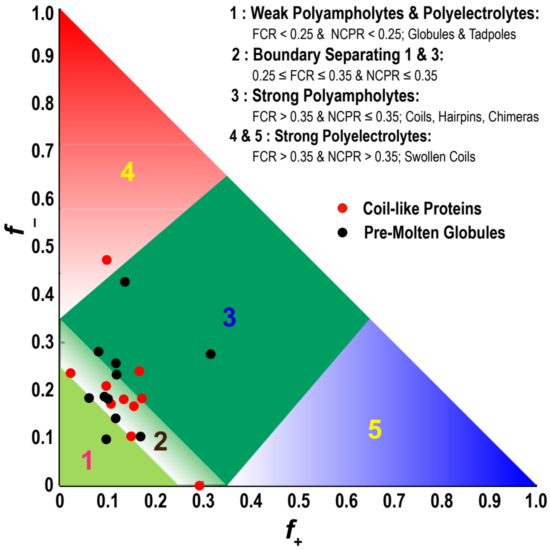

Fig. 7. Diagram of states for IDPs. We focus on sequences that fall below the parameterized line (NCPR = 2.785H – 1.151), developed by Uversky et al. (43) to separate IDPs from sequences that fold autonomously. Here, H refers to the hydropathy score. Region 1 corresponds to either weak polyampholytes 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.

SUPPORTING INFORMATION APPENDIX

Section 1 – Figures showing results that supplement analysis in the main text

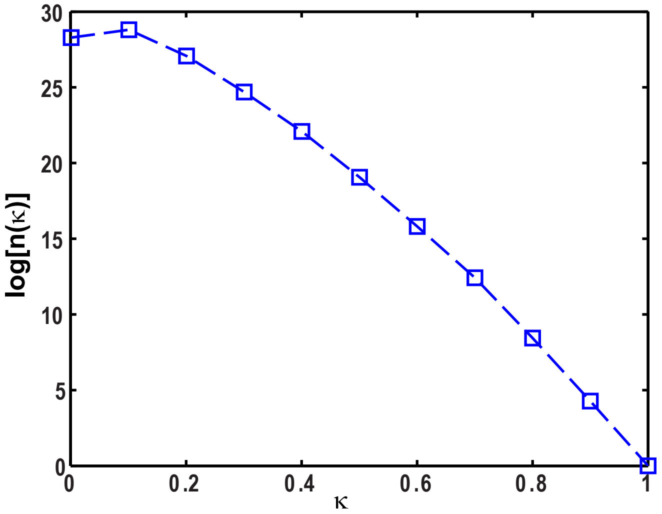

Figure S1: Semi-log plot of the number density of sequence variants of (Glu-Lys)$_{25}$, n(κ), that have similar κ values.

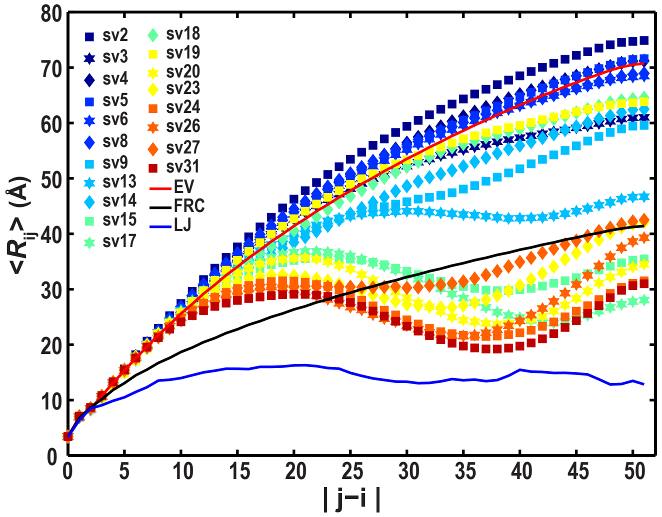

Figure S2: $\langle R_{\text{ij}} \rangle$ profiles for nineteen of the thirty sequence variants from Figure 1 of the main text. Profiles for the remaining sequence variants are shown in panels (A) and (B) of Figure 3 of the main text. The profiles are colored according to their κ values; the cooler colors correspond to low κ values, and colors become hotter as κ increases.

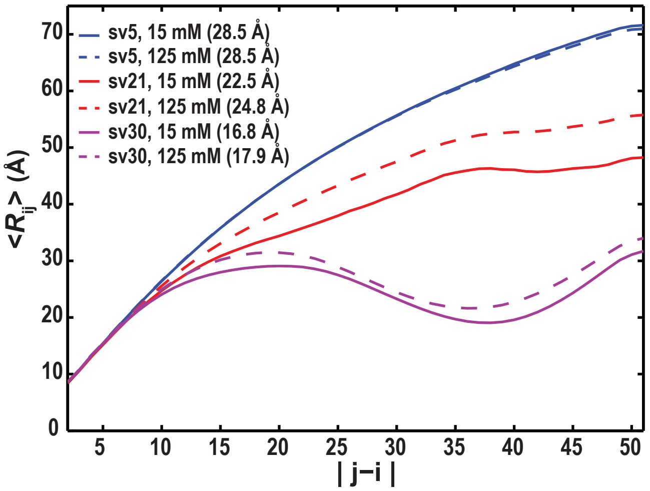

Figure S3: Quantification of the salt dependence of conformational properties for three of the sequence variants shown in Figure 1. The colors identify the sequence variants and the line types identify the profiles in 15 mM NaCl (solid), and 125 mM NaCl (dashed). In each case, the error bars (not shown here) are smaller than the differences between the profiles for two different solution conditions. The ensemble-averaged radii of gyration are as follows: For sv5, $\langle R_g \rangle$ = 28.5 ± 0.03 Å in 15 mM NaCl and 28.4 ± 0.12 Å in 125 mM NaCl. For sv21 $\langle R_g \rangle$ = 22.5 ± 0.08 Å in 15 mM NaCl and 24.8 ± 0.02 Å in 125 mM NaCl and for sv30, $\langle R_g \rangle$ = 16.8 ± 0.02 Å in 15 mM NaCl and 17.9 ± 0.02 Å in 125 mM NaCl. The figure shows that the $\langle R_{ij} \rangle$ profiles of variants sv21 (κ=0.2737) and sv30 (κ=1) are consistent with chain expansion in 125 mM NaCl due to weakened long-range attractions vis-à-vis 15 mM NaCl. In contrast, well-mixed sequences such as sv5 (κ=0.0245) show negligible salt dependence, which is explained by the *a priori* counterbalancing of electrostatic repulsions and attractions.

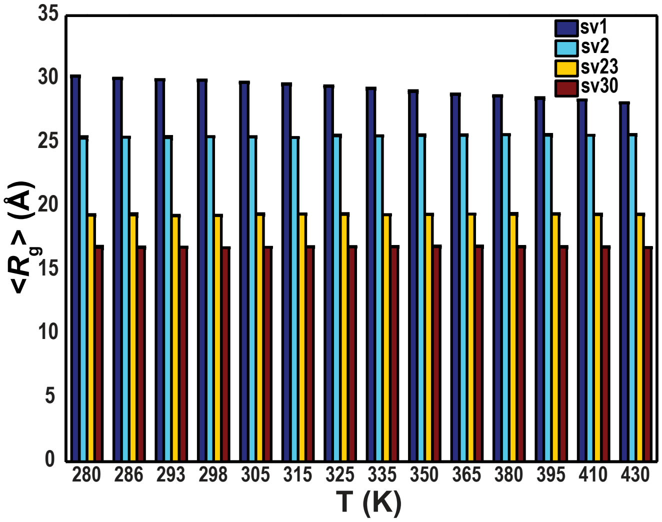

Figure S4: Temperature dependence of $\langle R_g^2 \rangle$ for different sequence variants of (Glu-Lys)$_{25}$ (15 mM NaCl). Each stack of bars corresponds to a single temperature and the colors pertain to different sequence variants. Figures S4 and S5 summarize the temperature dependence of conformational properties for sequence variants of (Glu-Lys)$_{25}$. For low κ-variants, the $\langle R_{ij} \rangle$ profiles shift toward that of the EV limit through a systematic, albeit weak chain contraction. Increased conformational entropy offsets the additional expansion vis-à-vis the EV limit that results from favorable solvation of charged sidechains. The contraction seen with increased temperature is considerably weaker than the collapse observed in single molecule measurements for unfolded proteins (1). For sequences with higher κ-values such as sv16, sv21, and sv22, the increase in entropy partially screens the attractions, leading to weak chain expansion, which is the canonical expectation

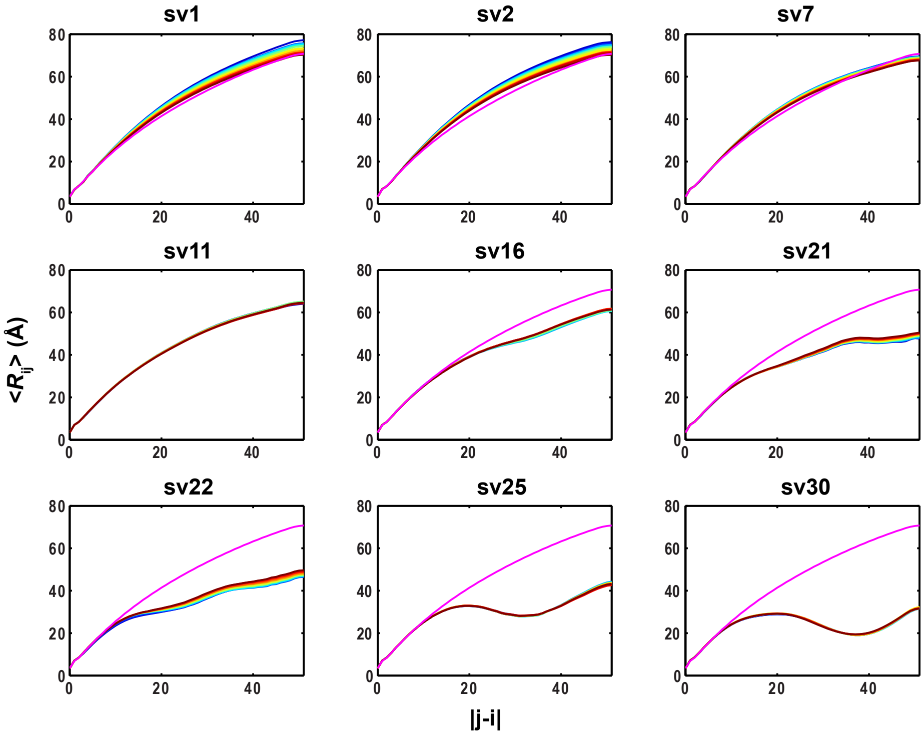

Figure S5: Temperature dependence of the $\langle R_{\mathrm{ij}} \rangle$ profiles for selected sequence variants of (Glu-Lys)$_{25}$ to illustrate the weak contraction / expansion of variants with low versus intermediate values of κ. This contraction / expansion results from increased conformational fluctuations that improves the convergence on EV limit profiles. In each panel, the EV limit panel is shown in magenta. As temperature increases, the color of the $\langle R_{\mathrm{ij}} \rangle$ profile switches from cooler colors to hotter ones. There are noteworthy complexities that have been documented with respect to the temperature dependence of conformational properties in atomistic simulations that can be traced to the temperature dependence of specific water models (2). In these studies, the effects of solvation are captured using explicit representations of water molecules. The weak expansion / contraction that is depicted here can be rationalized in terms of the increase in conformational entropy associated with increased thermal fluctuations as temperature increases. It is interesting that even in this simple framework one sees non-canonical results, *i.e.*, the weak contraction sequences with high FCR and / or low κ values. We propose that this represents a convergence toward the higher entropy EV limit distribution either via contraction or expansion. The degree of compaction / expansion seen in our simulations is weaker than that those observed in experiments (1). It is challenging to model such temperature dependent effects accurately because it requires an appropriate balancing of the temperature dependence of the hydrophobicity of non-polar groups and the favorable solvation of charged / polar groups. A generalization of the ABSINTH framework should make this feasible especially if we include the desired flexibility through the temperature dependence of the reference free energies of solvation for model compounds that make up the solvation groups in ABSINTH and the temperature dependence of the bulk dielectric constant of water.

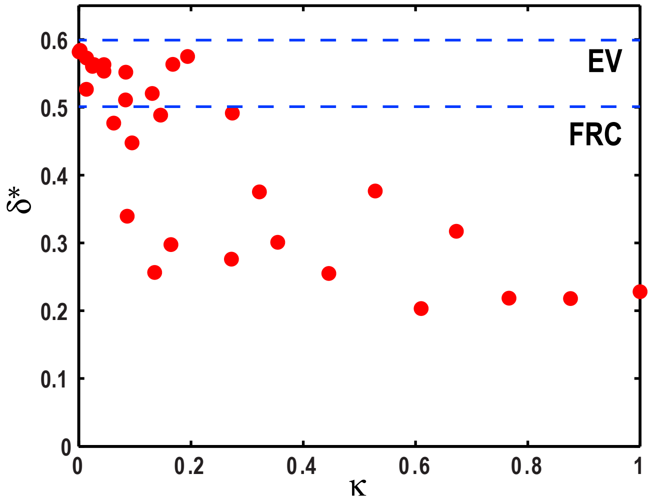

Figure S6: Plot of ensemble-averaged asphericity values δ^* against κ for sequence variants of (Glu-Lys)$_{25}$. For each sequence, δ^* is calculated from the ensemble of eigenvalues for gyration tensors that describe individual conformations as described in previous work (3). The two dashed lines intersect the ordinate at values corresponding to the average asphericity values obtained for a representative sequence variant of (Glu-Lys)$_{25}$ in the EV limit and as Flory random coils (FRC), respectively. The results show a transition between prolate ellipsoids for low κ and semi-compact, low asphericity hairpins for higher κ values. The κ dependence of asphericity values can be tested experimentally using measurements of rotational diffusion and changes to structure factors in small angle X-ray scattering.

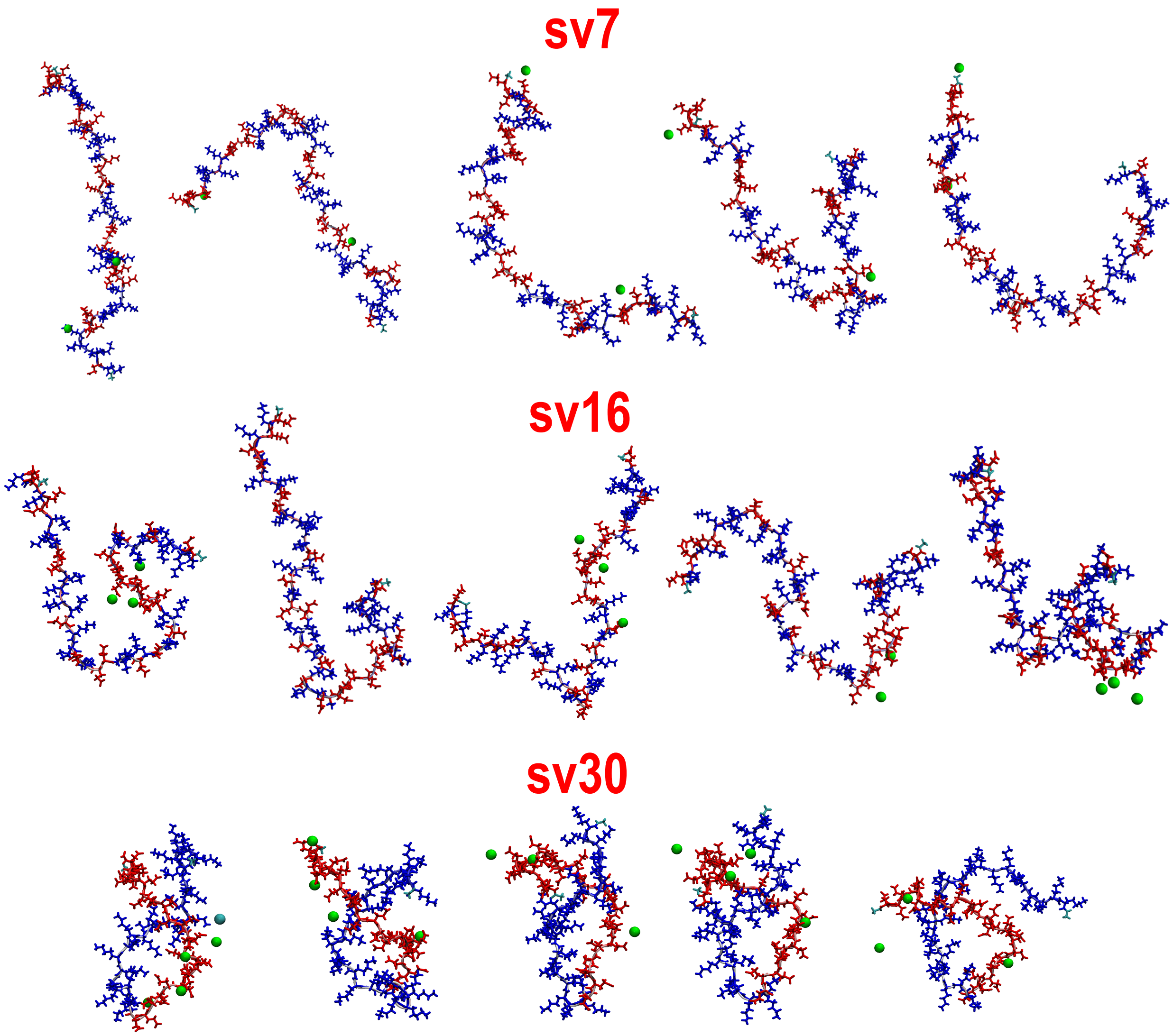

Figure S7: Montage of representative conformations drawn at random from individual Markov chain Metropolis Monte Carlo trajectories (T=298 K and [NaCl] = 15 mM) for different sequence variants of $(Glu\text{-}Lys)_{25}$. In this montage, the lysine and glutamate residues are shown in blue and red, respectively. For each of the snapshots shown, we identified the mobile solution ions that were within 7Å (the Bjerrum length) from the center of mass of the chain. Most of the mobile ions lie beyond this length scale because the overall charge neutrality of the sequences and the conformational fluctuations ensure that the oppositely charged sidechains act as counterion clouds for each other. For sequence variants of higher κ values or conformations that lead to higher density of similar charges in a local region, there is quantifiable presence of Na^+ ions around the higher charge density carboxylate moieties in glutamate sidechains and these are shown as green spheres; Cl^- ions are shown as light blue spheres. The conformations were rendered using the VMD package (4).

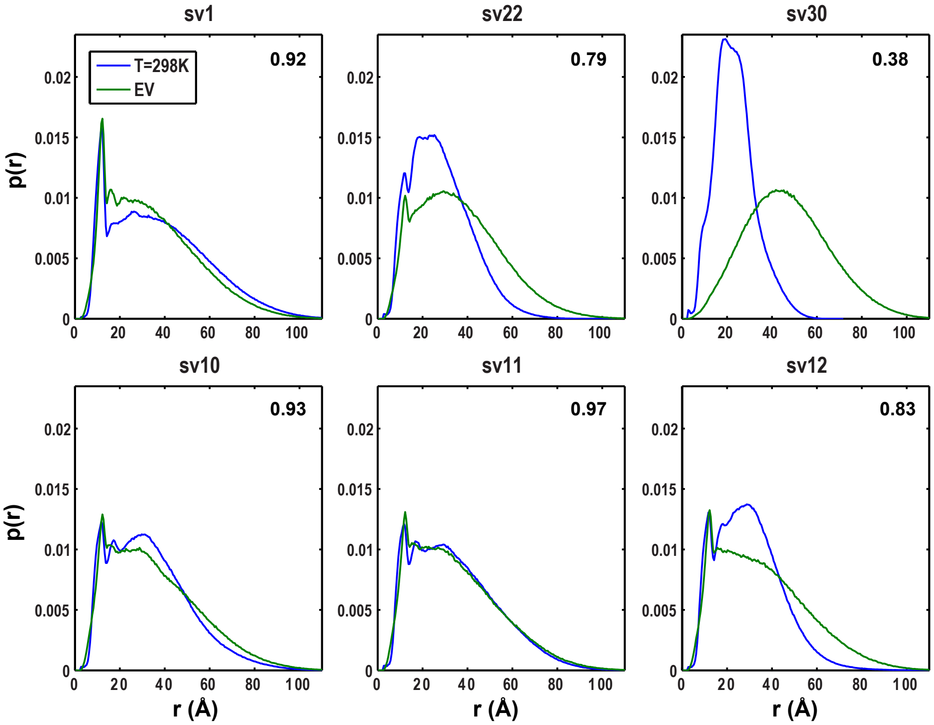

Figure S8: Comparison of histograms p(r) of distances r between the amino and carboxyl sidechain tips of Lys and Glu residues in different sequence variants of the (Glu-Lys)$_{25}$ system. In each panel, the number on the top right corner quantifies the overlap between p(r) histograms calculated in the EV limit (green curves) and histograms from simulations using the full ABSINTH model at *T*=298 K in 15 mM NaCl (blue curves). The overlap decreases with increasing κ (top row) and as attractions become more pronounced for similar values of κ, bottom row. Overlaps between pairs of distributions are calculated as described in section 2 of this *SI Appendix*. The short-range peaks in all of the pair distribution profiles are a consequence of the constraints imposed by chain connectivity and the validity of the blob concept, which ensure that some set of sidechains are guaranteed to be close to each irrespective of the electrostatic interactions. We calculate distance histograms between the epsilon nitrogen atom of the amine and the OE1 atom of carboxylate groups. It is worth noting that there are pronounced peaks the distance distributions even for the EV limit and these correspond to the "quenched disorder" due to amino acid sequence. In well-mixed sequences, the blobs are admixtures of amine and carboxylate groups and the differences between the sidechain structures is manifest in many of the inter-blob distances. Conversely, in strongly segregated sequences, the blobs are either entirely amines or carboxylates and hence all blobs have one or the other composition, giving rise to smooth inter-residue distance distributions. The structures in these profiles reflect the contributions of amino acid sequence. As for the peaks at larger separations that are seen for sequences with low κ, these correspond to the attractions between oppositely charged counterion clouds and the length scale reflects the fact that these attractions are the result of conformational fluctuations as opposed to salt bridges.

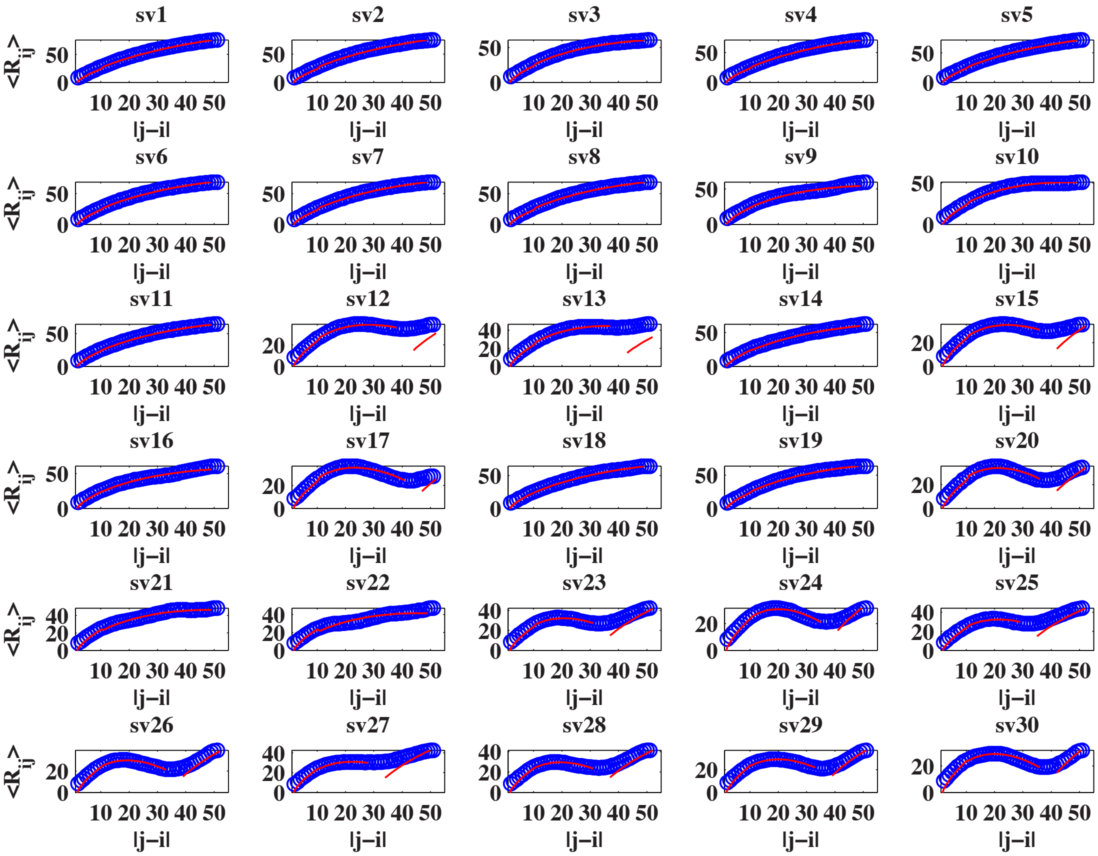

Figure S9: Fits to $\langle R_{\mathrm{ij}} \rangle$ profiles for thirty of the sequence variants of (Glu-Lys)$_{25}$. Details regarding the κ values for each variant are shown in Figure 1 of the main text. The blue circles are the raw data for $\langle R_{\mathrm{ij}} \rangle$ profiles and the red curves denote numerical fits to the data obtained from the scaling theory described in the main text. In plotting the fitting procedure, we ignore the crossover region between the two intervals $2g \leq |\mathrm{j}–\mathrm{i}| \leq l_c$ and $|\mathrm{j}–\mathrm{i}| > l_c$. The parameters p_0, p_1, and l_{c} for each of the sequence variants are shown in Figure S7.

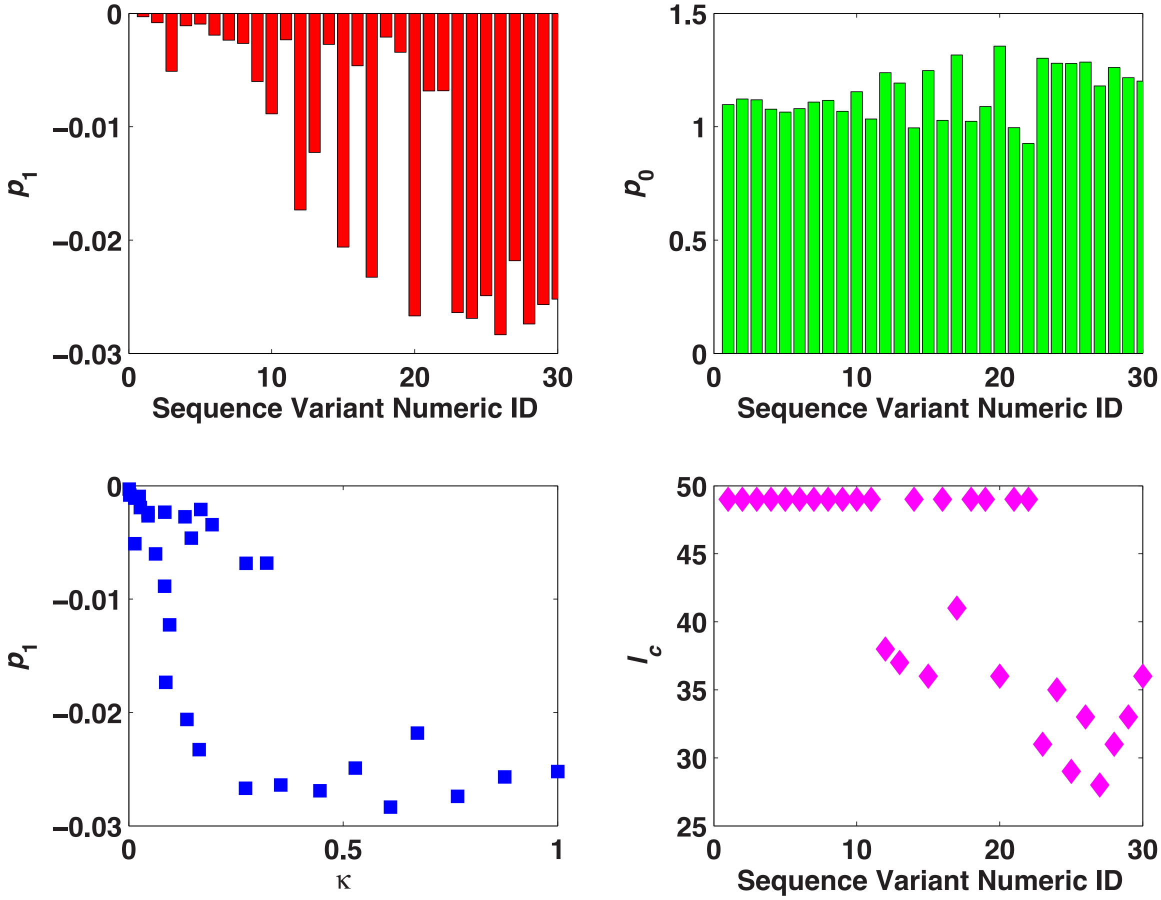

Figure S10: Parameters obtained from numerical fitting of scaling form to results for $\langle R_{ij} \rangle$ profiles for all sequence variants of (Glu-Lys)$_{25}$ shown in Figure 1 of the main text. The top left and top right panels show bar plots of the parameters p_1 and p_0, respectively. In these plots, the height of each bar corresponds to the value of the corresponding parameter along the ordinate and the numeric identifier of the sequence variant is shown along the abscissa. The bottom left panel plots the correlation between p_1 and κ. The bottom right panel shows the values for the crossover length l_c. In general, for well-mixed sequences, *i.e.*, low / intermediate κ values, the value of l_c spans the chain length and $p_1 \approx 0$ implying minimal deviations from the EV limit profiles.

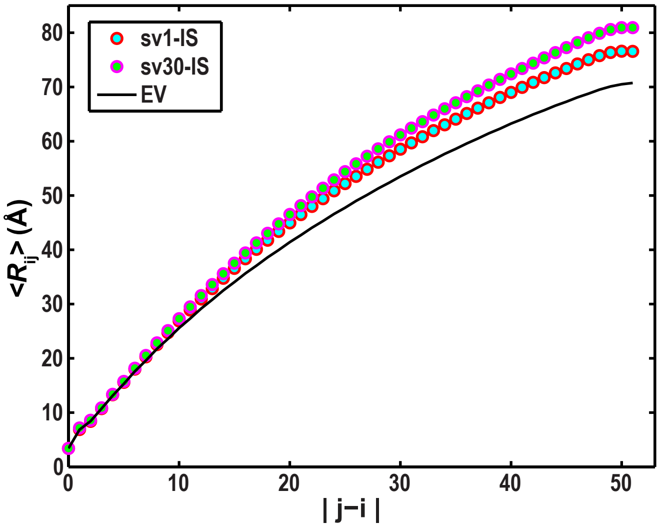

Figure S11: IS limit $\langle R_{ij} \rangle$ profiles for two of the sequence variants of (Glu-Lys)$_{25}$. These profiles are compared to that of the EV limit

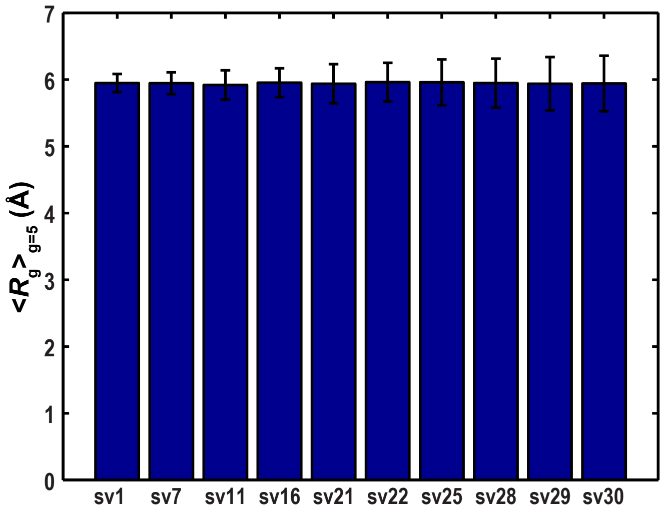

Figure S12: Demonstration of the validity of the blob concept. The plot shows average radii of gyration calculated for segments of length g=5 extracted from different sequence variants of (Glu-Lys)$_{25}$. The average R_g for blobs, denoted as ξ in the main text, is ≈6Å irrespective of the sequence from which the blobs are extracted.

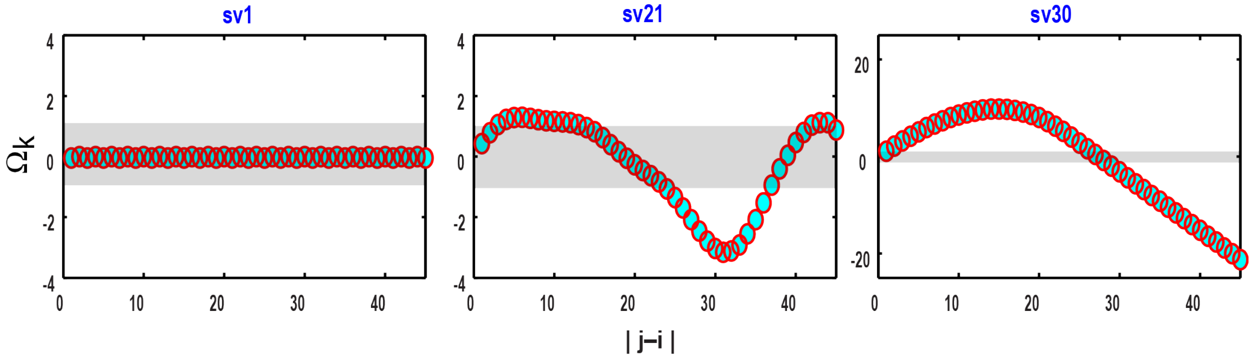

Figure S13: Plots of cumulative sums of the length scale specific Coulomb coupling parameters, $\Omega_k = \sum_{k=1}^{|j-i|} < \Gamma_k >$ for three different sequence variants of (Glu-Lys)$_{25}$. The ordinate corresponds to the cumulative sum of $\mathbf{\Gamma}_{\mathbf{ij}}$ and the height of the gray bars is ~ *kT*. Note that the scales along the ordinate are different for different panels.

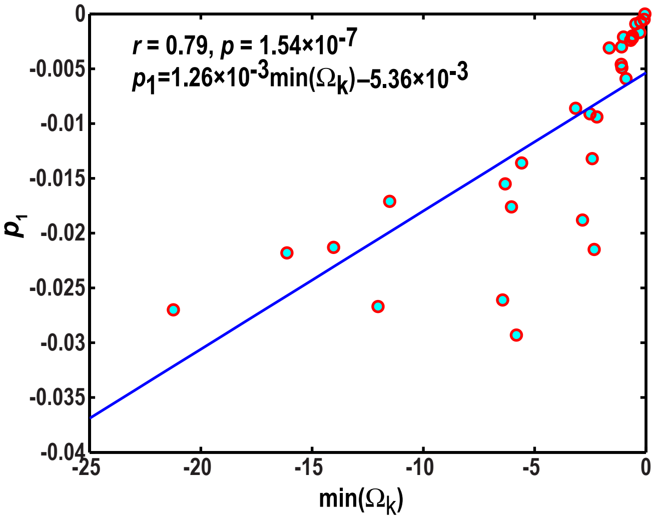

Figure S14: Plot of p_1 against min(Ω_k) for sequence variants of (Glu-Lys)$_{25}$. See caption to Figure S13 for definition of Ω_k. The Pearson *r* value, the *p*-value that quantifies the probability of realizing this correlation at random, and the equation for the line of best fit (shown in blue) are also shown in the legend to the plot.

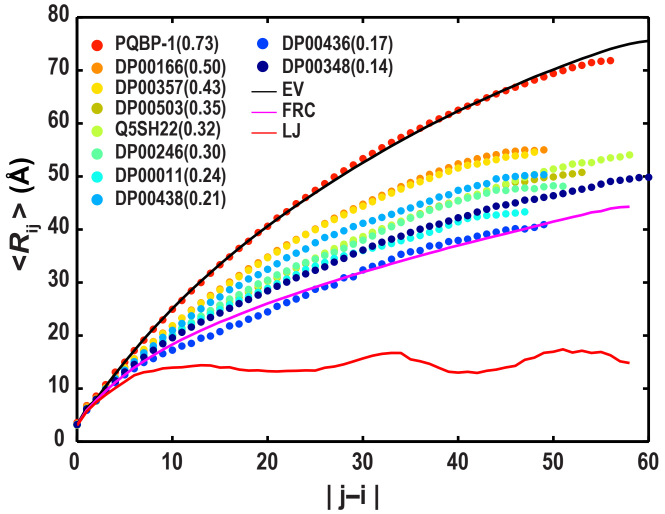

Figure S15: ⟨R_{ij}⟩ profiles for the ten naturally occurring IDPs (see Table S2) simulated in the IS limit. For comparison, the plot includes profiles for the EV limit, Flory random coil, and a maximally compact LJ globule. The legend shows the sequence identifiers and the FCR values for each sequence.

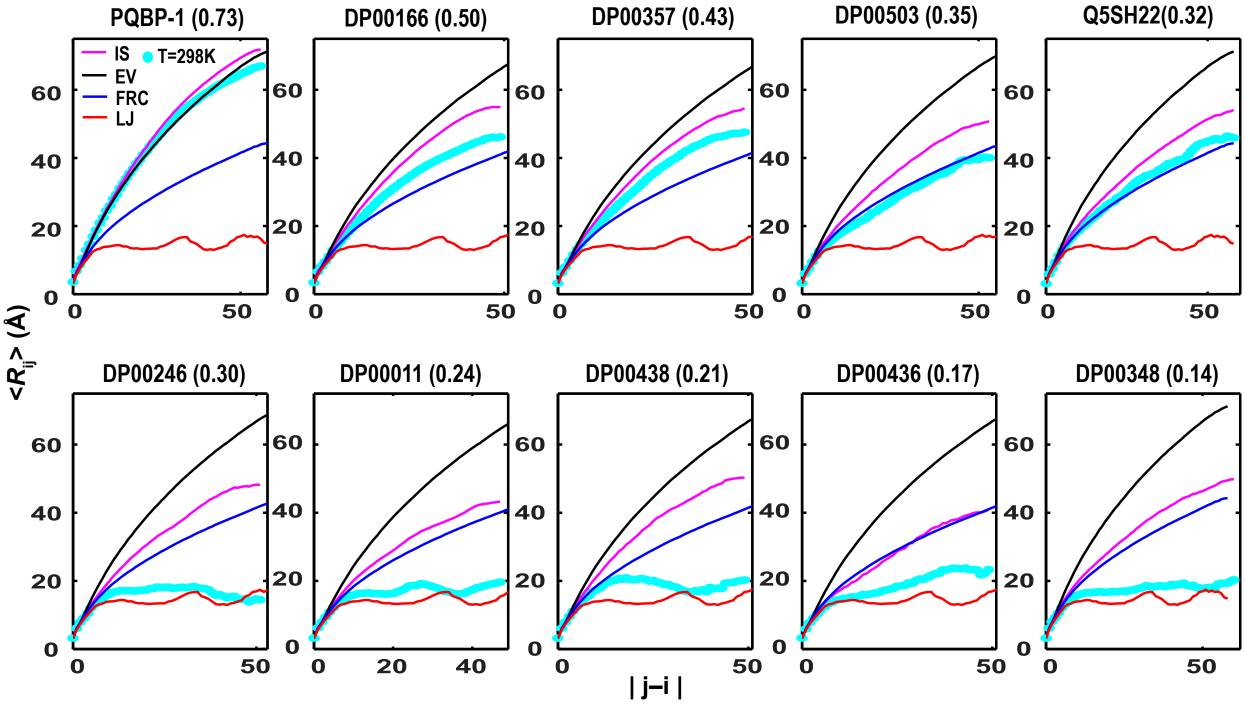

Figure S16: Comparison of IS limit profiles to those obtained using simulations based on the full ABSINTH Hamiltonian for all ten naturally occurring IDPs studied in this work. In each panel, the IS limit profile is shown in magenta, the EV limit profile in black, the profile for Flory random coils in blue, the LJ globule limit in red, and the actual profile in cyan circles. The title for each panel shows the sequence identifier and the FCR value in parentheses. The coil formers are shown along the top row and the globule formers along the bottom row.

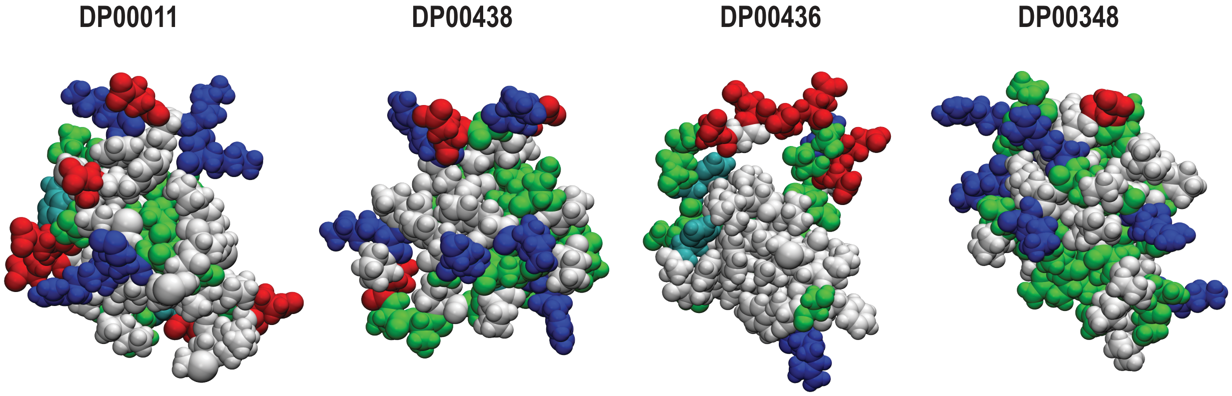

Figure S17: Space filling pictures of representative conformations drawn from the simulated ensembles for weak polyampholytic IDPs. The DisProt identifiers for each sequence are shown in the figure and sequence details are available from Table S2 and section 2 of this *SI Appendix*. The color coding is as follows: hydrophobic residues are in whitish gray, uncharged polar residues in green, proline residues in cyan, negatively charged residues in red, and positively charged residues in blue. In general, the compaction we observe represents an optimal trade off between the driving forces for chain compaction due to decreased FCR and the favorable solvation of charged sidechains. Conformations of DP00436 are "tadpole-like" in that they combine compact domains with extended regions – a result of the chimeric nature of the underlying sequence which includes an acidic C-terminal stretch in an otherwise neutral chain (see Table S2). The models were generated using version 1.9.1 of the Visual Molecular Dynamics software package (4).

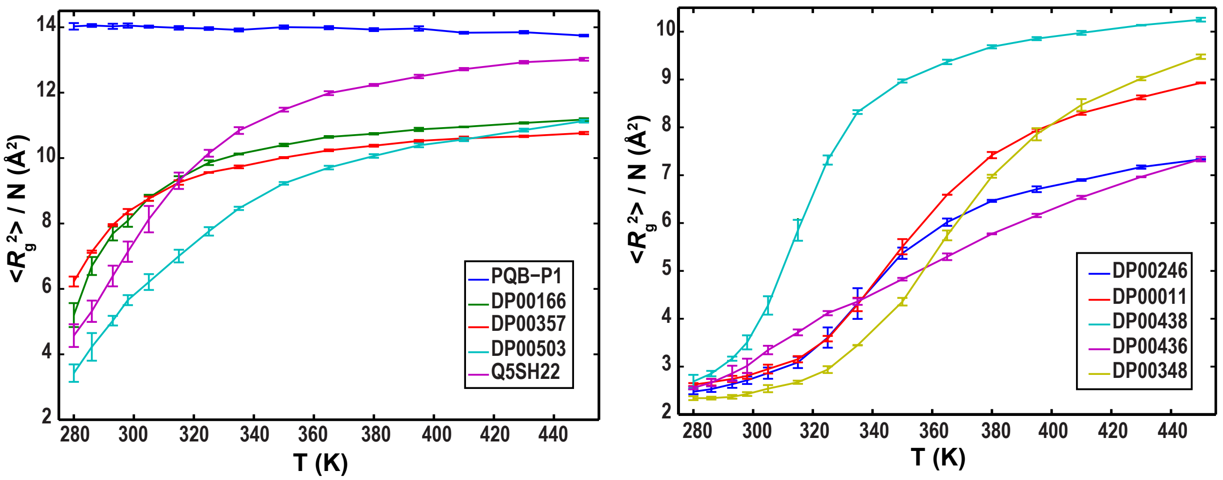

Figure S18: Temperature dependence of $\langle R_g{}^2 \rangle/N$ for the ten naturally occurring IDPs. In the interest of clarity we set the scales for the ordinates to be different between the two panels. These results show increased sensitivity of $\langle R_g{}^2 \rangle/N$ to changes in temperature for sequences that are weak polyampholytes (right panel) when compared to those that are strong polyampholytes (left panel).

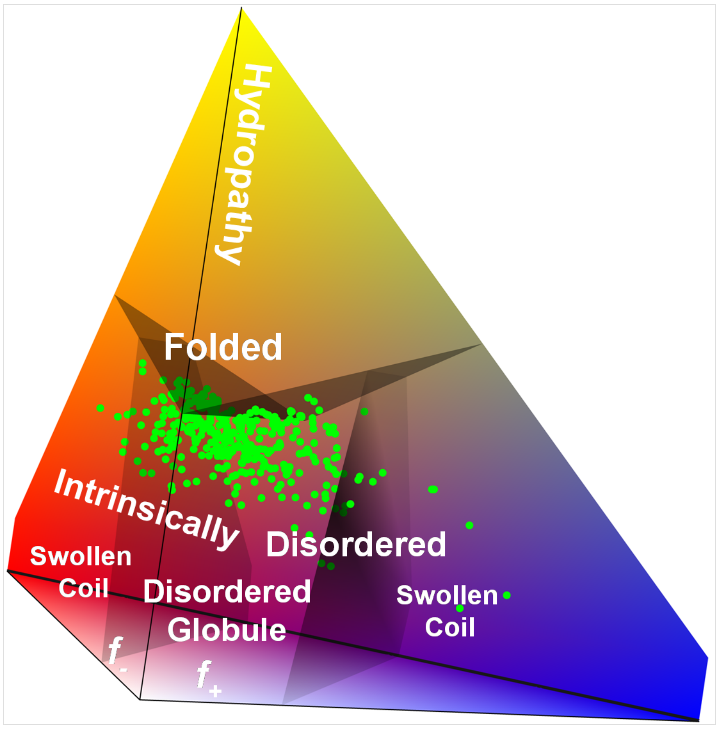

Figure S19: Diagram of states of Mao et al. (3) annotated by a subset of sequences drawn from the DisProt database (5). The three axes denote f_+, f_-, and the hydropathy and all three parameters are calculated from the amino acid composition. The diagram of states of Mao et al. and the designations of different regions were a generalization of the work of Uversky et al. (6) Their plane was turned into a pyramid by unfolding the parameterized line, NCPR = 2.785H – 1.151 where H denotes hydropathy, which divides the pyramid into a top and bottom portion. For sequences with low hydropathy values the work of Mao et al. distinguishes the bottom portion of the pyramid into globule versus coil formers based on the NCPR values. We annotated the Mao et al. diagram of states using a subset of sequences – 364 in all – drawn from the DisProt database. These sequences have hydropathy values that designate them as being disordered, *i.e.*, they lie in the bottom portion of the pyramid. Additional filters were used for chain length ($N > 30$) and the fraction of proline residues (f_{pro}) such that $f_{\text{pro}} < 0.3$. Ninety seven percent of sequences used in this annotation have NCPR < 0.26 and are predicted to be globule formers. Of these sequences, the NCPR values for 81.3% and 71.4% are less than 0.1 and 0.05, respectively.

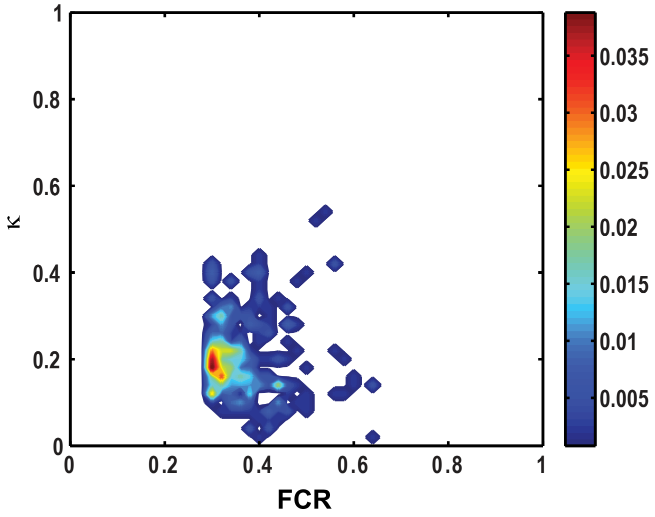

Figure S20: Joint density distribution ρ(FCR,κ) for a subset of the 364 sequences drawn from the DisProt database. As noted in the main text, this analysis focuses on sequences that satisfy the constraints $N > 30$, FCR ≥ 0.3, NCPR < 0.25, and fraction of proline residues < 0.3.

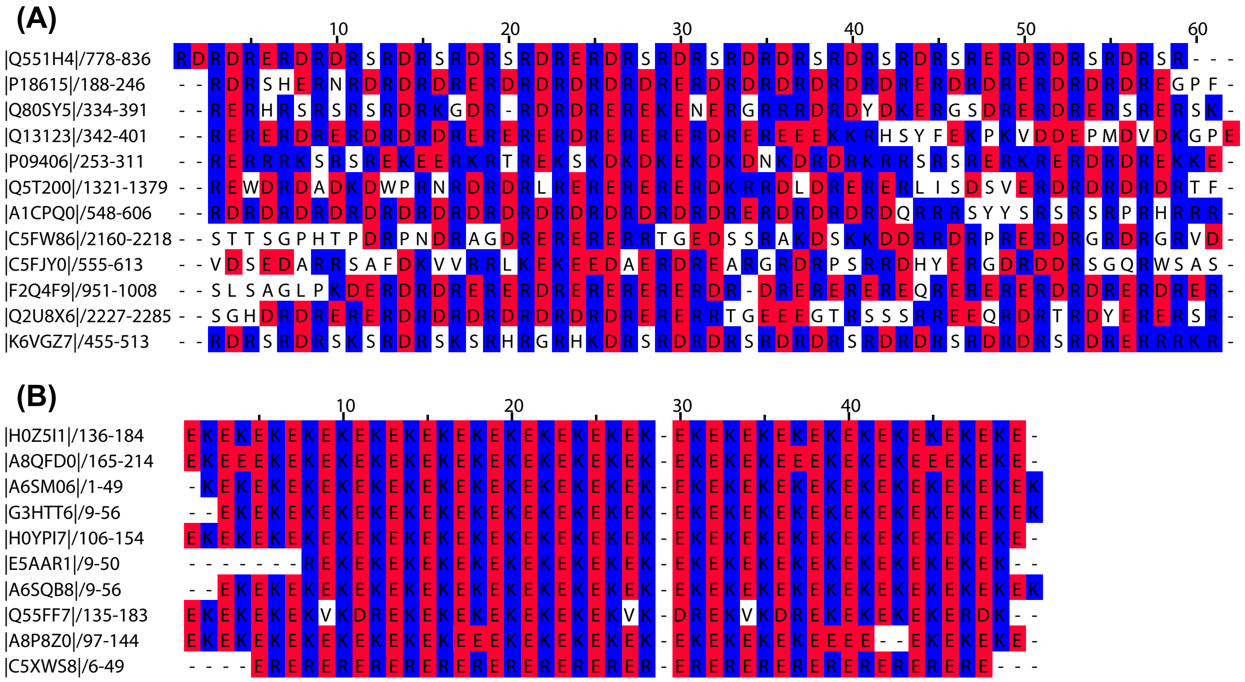
(A)
10
20
30
40
50
60
|Q551H4|/778-836 RDRDRERDRDRSRDRSRDRSRDRERDRSRDRSRDRDRSRDRSRDRSRERDRDRSRDRSR---
|P18615|/188-246 --RDRSHERNRDRDRDRERDRDRDRDRDRERDRDRDRDRDRDRERDRDRERDRDRDREGPF-
|Q80SY5|/334-391 --RERHRSRSRSRDRKGDR-RDRDRERERKENERGRRRDRDYDKERGSDRERDRERSRERSK-
|Q13123|/342-401 --REREDRERDRDRDRERERERDRERERERDREREEEKKRHSYFEKPKVDDEPMDVDKGPE
|P09406|/253-311 --RERRRKSRSREKEERKRTREKSKDKDKEKDKDNKDRDRKRRSRSRERKRERDRDREKKE-
|Q5T200|/1321-1379 --REWDRDADKDWPRNRDRDRLRERERERERDKRRDLDRERERLISDSVERDRDRDRDRTF-
|A1CPQ0|/548-606 --RDRDRDRDRDRDRDRDRDRDRDRDRDRDRDRERDRDRDRDQRRRSYYSRSRSRPRHRRR-
|C5FW86|/2160-2218 --STTSGPHTPDRPNDRAGDRERERERRTGEDSSRAKDSKKDDRRDRPRERDRGRDRGRVD-
|C5FJY0|/555-613 --VDSEDARRSAFDKVVRRLKEKEEDAERDREARGRDRPSRRDHYERGDRDDRSGQRWSAS-
|F2Q4F9|/951-1008 --SLSAGLPKDERDRDRERERDRERERERERDR-DREREREREQRERERERDRDRERDRER-
|Q2U8X6|/2227-2285 --SGHDRDRERERDRDRDRDRDRDRERERRTGEEEGTRSSSRREEQRDRTRDYERERSR-
|K6VGZ7|/455-513 --RDRSRDRSKSRDRSKSRHRGRHKDRSRDRDRSRDRDRSRDRDRSRDRDRSRDRERRRKR-
(B)
10
20
30
40
|H0Z5I1|/136-184 EKEKEKEKEKEKEKEKEKEKEKEKEKEK-EKEKEKEKEKEKEKEKEKEKE-
|A8QFD0|/165-214 EKEEEKEKEKEKEKEKEKEKEKEKEKEK-EKEKEKEEEKEKEKEEEKEKE-
|A6SM06|/1-49 -KEKEKEKEKEKEKEKEKEKEKEKEKEK-EKEKEKEKEKEKEKEKEKEKEK
|G3HTT6|/9-56 --EKEKEKEKEKEKEKEKEKEKEKEKEK-EKEKEKEKEKEKEKEKEKEKEK
|H0YPI7|/106-154 EKEKEKEKEKEKEKEKEKEKEKEKEKEK-EKEKEKEKEKEKEKEKEKEKE-
|E5AAR1|/9-50 -------REKEKEKEKEKEKEKEKEKEK-EKEKEKEKEKEKEKEKEKEK--
|A6SQB8|/9-56 --EKEKEKEKEKEKEKEKEKEKEKEKEK-EKEKEKEKEKEKEKEKEKEKEK
|Q55FF7|/135-183 EKEKEKEKVKDREKEKEKEKEKEKEKVK-DREKVKDREKEKEKEKERDK--
|A8P8Z0|/97-144 EKEKEKEKEKEKEKEKEEEKEKEKEKEK-EKEKEKEKEEEE--EKEKEKE-
|C5XWS8|/6-49 ----ERERERERERERERERERERERER-ERERERERERERERERERE---

Figure S21: Representative strong polyampholytic regions drawn from sequence databases. Panel A shows examples of naturally occurring sequences that contain long stretches of Asp/Glu-Arg repeats. We retrieved these sequences by performing a BLAST search (http://blast.ncbi.nlm.nih.gov/Blast.cgi) of the UniProtKB database (http://www.uniprot.org/help/uniprotkb) using the PQBP-1 sequence as a template. Panel B shows examples of naturally occurring sequences that contain long stretches of Glu-Lys repeats. We retrieved these sequences by performing a Profile HMM (http://www.biology.wustl.edu/gcg/hmmanalysis.html) search of the UniProtKB database (http://www.uniprot.org/help/uniprotkb) using the sv1 sequence variant of (Glu-Lys)$_{25}$ as an input. For the Profile HMM search we used the HMMER web server (http://hmmer.janelia.org/). In both cases, the sequences were aligned using Clustal (http://www.clustal.org/). We annotate all sequences by their UniProt identifiers. Positively charged residues are colored in blue and negatively charged residues are colored in red.

Table S1: Predictions of disorder tendencies for the sequence variants of (Glu-Lys)$_{25}$ using the meta predictor metaPrDos (7). The disorder scores take values between 0 and 1, with the predicted score approaching unity as the disorder tendency increases.

Sequence Variants	Disorder Tendencies
sv1	0.8629
sv2	0.8591
sv3	0.8477
sv4	0.8470
sv5	0.8520
sv6	0.8576
sv7	0.8577
sv8	0.8660
sv9	0.8544
sv10	0.8517
sv11	0.8505
sv12	0.8596
sv13	0.8458
sv14	0.8557
sv15	0.8429
sv16	0.8923
sv17	0.8453
sv18	0.8567
sv19	0.8648
sv20	0.8600
sv21	0.8457
sv22	0.8359
sv23	0.8455
sv24	0.8520
sv25	0.8414
sv26	0.8439
sv27	0.8454
sv28	0.8429
sv29	0.8435
sv30	0.8423
sv31	0.8359

Table S2: Sequence characteristics of the simulated naturally occurring IDP sequences. The values of δ (and hence δ_{max}) and κ were calculated only for strong polyampholytes.

Sequence	***N***	f_+	f_-	**NCPR**	**FCR**	**σ**	δ_{max}	**κ**	**Disorder tendencies predicted using metaPrDos (7)**
PQBP-1	55	0.36	0.36	0 .0	0.73	0.0	0.72	0.02	0.73
DP00166	48	0.27	0.23	0.04	0.50	0.003	0.47	0.11	0.72
DP00357	47	0.19	0.23	0.04	0.43	0.004	0.38	0.12	0.70
DP00503	52	0.08	0.27	0.19	0.35	0.10	0.24	0.17	0.54
Q5SH22	57	0.04	0.28	0.25	0.32	0.20	0.19	0.20	0.49
DP00246	50	0.12	0.18	0.06	0.30	0.01	0.24	0.29	0.71
DP00011	46	0.11	0.13	0.02	0.24	0.002	-	-	0.52
DP00438	48	0.15	0.06	0.08	0.21	0.03	-	-	0.74
DP00436	48	0.04	0.13	0.08	0.17	0.04	-	-	0.49
DP00348	59	0.12	0.02	0.10	0.14	0.07	-	-	0.46

Table S3: Sequences for PreMolten Globules. This designation and sequence information are taken from the inventory collated by Uversky (8).

Proteins	Sequence
Osteocalcin	YLDSGLGAPVPYPDPLEPKREVCELNPNCDELADHIGFQEAYQRFYGPV
Heat stable protein kinase inhibitor	MTDVETTYADFIASGRTGRRNAIHDILVSSASGNSNELALKLAGLDINKTEGEEDAQRSSTEQSGEAQGEAAKSE
Caldesmon 636-771 fragment	RLEQYTSAVVGNKAAKPAKPAASDLPVPAEGVRNIKSMWEKGNVFSSPGGTGTPNKETAGLKVGVSSRINEWLTKTPEGNKSPAPKPSDLRPGDVSGKRNLWEKQSVEKPAASSSKVTATGKKSETNGLRQFEKEP
pf1 gene 5 protein	MNMFATQGGVVELWVTKTDTYTSTKTGEIYASVQSIAPIPEGARGNAKGFEISEYNIEPTLLDAIVFEGQPVLCKFASVVRPTQDRFGRITNTQVLVDLLAVGGKPMAPTAQAPARPQAQAQAPRPAQQPQGQDKQDKSPDAKA
DARRP-32	MDPKDRKKIQFSVPAPPSQLDPRQVEMIRRRRPTPAMLFRLSEHSSPEEEASPHQRASGEGHHLKSKRSNPCAYTPPSLKAVQRIAESHLQSISNLGENQASEEEDELGELRELGYPREEEEEEEEEDEEEEEDSQAEVLKGSRGSAGQKTTYGQGLEGPWERPPPLDGPQRDGSSEDQVEDPALNEPGEEPQRPAHPEPGT
Manganese stabilizing protein	EGGKRLTYDEIQSKTYLEVKGTGTANQCPTVEGGVDSFAFKPGKYTAKKFCLEPTKFAVKAEGISKNSGPDFQNTKLMTRLTYTLDEIEGPFEVSSDGTVKFEEKDGIDYAAVTVQLPGGERVPFLFTIKQLVASGKPESFSGDFLVPSYRGSSFLDPKGRGGSTGYDNAVALPAGGRGDEEELQKENNKNVASSKGTITLSVTSSKPETGEVIGVFQSLQPSDTDLGAKVPKDVKIEGVWYAQLEQQ
Calreticulin, human C fragment	YDNFGVLGLDLWQVKSGTIFDNFLITNDEAYAEEFGNETWGVTKAAEKQMKDKQDEEQRLKEEEEDKKRKEEEEAEDKEDDEDKDEDEEDEEDKEEDEEEDVPGQAKDEL
Calsequestrin, rabbit	MNAADRMGARVALLLLLVLGSPQSGVHGEEGLDFPEYDGVDRVINVNAKNYKNVFKKYEVLALLYHEPPEDDKASQRQFEMEELILELAAQVLEDKGVGFGLVDSEKDAAVAKKLGLTEEDSIYVFKEDEVIEYDGEFSADTLVEFLLDVLEDPVELIEGERELQAFENIEDEIKLIGYFKNKDSEHYKAFKEAAEEFHPYIPFFATFDSKVAKKLTLKLNEIDFYEAFMEEPVTIPDKPNSEEEIVNFVEEHRRSTLRKLKPESMYETWEDDMDGIHIVAFAEEADPDGYEFLEILKSVAQDNTDNPDLSIIWIDPDDFPLLVPYWEKTFDIDLSAPQIGVVNVTDADSVWMEMDDEEDLPSAEELEDWLEDVLEGEINTEDDDDEDDDDDDDD
SdrD protein, B1-B5 fragment	VYKIGNYVWEDTNKNGVQELGEKGVGNVTVTVFDNNTNTKVGEAVTKEDGSYLIPNLPNGDYRVEFSNLPKGYEVTPSKQGNNEELDSNGLSSVITVNGKDNLSADLGIYKPKYNLGDYVWEDTNKNGIQDQDEKGISGVTVTLKDENGNVLKTVTTDADGKYKFTDLDNGNYKVEFTTPEGYTPTTVTSGSDIEKDSNGLTTTGVINGADNMTLDSGFYKTPKYNLGNYVWED

	TNKDGKQDSTEKGISGVTVTLKNENGEVLQTTKTDKDGKYQFTGLENGTYKVEFETPSGYTPTQVGSGTDEGIDSNGTSTTGVIKDKDNDTIDSGFYKPTYNLGDYVWEDTNKNGVQDKDEKGISGVTVTLKDENDKVLKTVTTDENGKYQFTDLNNGTYKVEFETPSGYTPTSVTSGNDTEKDSNGLTTTGVIKDADNMTLDSGFYKTPKYSLGDYVWYDSNKDGKQDSTEKGIKDVKVTLLNEKGEVIGTTKTDENGKYCFDNLDSGKYKVIFEKPAGLTQTGTNTTEDDKDADGGEVDVTITDHDDFTLDNGYYEEET
Topoisomerase I	MSGDHLHNDSQIEADFRLNDSHKHKDKHKDREHRHKEHKKEKDREKSKHSNSEHKDSEKKHKEKEKTKHKDGSSEKHKDKHKDRDKEKRKEEKVRASGDAKIKKEKENGFSSPPQIKDEPEDDGYFVPPKEDIKPLKRPRDEDDADYKPKKIKTEDTKKEKKRKLEEEEDGKLK
Calreticulin bovine	MLLPVPLLLGLLGLAAADPTVYFKEQFLDGDGWTERWIESKHKPDFGKFVLSSGKFYGDQEKDKGLQTSQDARFYALSARFEPFSNKGQTLVVQFTVKHEQNIDCGGGYVKLFPAGLDQTDMHGDSEYNIMFGPDICGPGTKKVHVIFNYKGKNVLINKDIRCKDDEFTHLYTLIVRPNNTYEVKIDNSQVESGSLEDDWDFLPPKKIKDPDAAKPEDWDDRAKIDDPTDSKPEDWDKPEHIPDPDAKKPEDWDEEMDGEWEPPVIQNPEYKGEWKPRQIDNPEYKGIWIHPEIDNPEYSPDSNIYAYENFAVLGLDLWQVKSGTIFDNFLITNDEAYAEEFGNETWGVTKAAEKQMKDKQDEEQRLHEEEEEKKGKEEEEADKDDDEDKDEDEEDEDEKEEEEEEDAAAGQAKDEL

Table S4: Sequences for Coil-like Proteins. This designation and sequence information are taken from the inventory collated by Uversky (8).

Proteins	Sequence
Vmw65 C-terminal domain	GSAGHTRRLSTAPPTDVSLGDELHLDGEDVAMAHADALDDFDLDMLGDGDSPGPGFTPHDSAPYGALDMADFEFEQMFTDALGIDEYGG
PDE g	MNLEPPKAEFRSATRVAGGPVTPRKGPPKFKQRQTRQFKSKPPKKGVQGFGDDIPGMEGLGTDITVICPWEAFNHLELHELAQYGII
wheat EM protein	MASGQQERSQLDRKAREGETVVPGGTGGKSLEAQENLAEGRSRGGQTRREQMGEEGYSQMGRKGGLSTNDESGGDRAAREGIDIDESKFKTKS
Apo-cytochrome c (acid denatured)	MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
Prothymosin a	MSDAAVDTSSEITTKDLKEKKEVVEEAENGRDAPANGNAQNEENGEQEADNEVDEEEEEGGEEEEEEEEEGDGEEEDGDEDEEAEAPTGKRVAEDDEDDDVETKKQKKTDEDD
g synuclein	MDVFKKGFSIAKEGVVGAVEKTKQGVTEAAEKTKEGVMYVGAKTKENVVQSVTSVAEKTKEQANAVSEAVVSSVNTVATKTVEEAENIAVTSGVVRKEDLRPSAPQQEGEASKEKEEVAEEAQSGGD
b synuclein	MDVFMKGLSMAKEGVVAAAEKTKQGVTEAAEKTKEGVLYVGSKTREGVVQGVASVAEKTKEQASHLGGAVFSGAGNIAAATGLVKREEFPTDLKPEEVAQEAAEEPLIEPLMEPEGESYEDPPQEEYQEYEPEA
fibronectin binding domain B	KKGKGKIARKKGKSKVSRKEPYIHSLKRDSANKSNFLQKNVILEEESLKTELLKEQSETRKEKIQKQQDEYKGMTQGSLNSLSGESGELEEPIESNEIDLTIDSDLRPKSSLQGIAGSNSISYTDEIEEEDYDQYYLDEYDEEDEEEIRL
a synuclein	MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATVAEKTKEQVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNEAYEMPSEEGYQDYEPEA
Stathmin	MASSDIQVKELEKRASGQAFELILSPRSKESVPEFPLSPPKKKDLSLEEIQKKLEAAEERRKSHEAEVLKQLAEKREHEKEVLQKAIEENNNFSKMAEEKLTHKMEANKENREAQMAAKLERLREKMYFWTHGPGAHPAQISAEQSCLHSVPALCPALGLQSALITWSDLSHHH

Section 2 – Details of simulation methods, assessments of simulation quality, and analysis of simulation results.

ABSINTH implicit solvation model and forcefield parameters: Details of the ABSINTH implicit solvation model and underlying forcefield paradigm have been published previously (3, 9, 10). This paradigm uses experimentally measured free energies of solvation for model compounds as inputs. These are coupled with refined parameters for modeling van der Waals interactions (9, 11) that are based on data for heats of fusion. Parameters for charges and the neutral group paradigm were taken from the OPLS-AA/L (12) forcefield, although in theory, the ABSINTH paradigm can be used, with appropriate caveats (9), with charge sets from most standard molecular mechanics forcefields. In the ABSINTH paradigm the effects of solvent-mediated interactions are captured using an implicit representation of the solvent whereas the effects of mobile ions and small solutes are simulated using explicit representations of the cosolutes. In all of our simulations we used explicit representations of Na^+ and Cl^- ions. We used the default parameters of Åqvist (13) instead of the new parameters developed by Mao and Pappu that are based on parameterization against crystal lattice properties (14). This choice is mainly chronological and given the low salt concentrations, we expect only minor differences in the simulation results based on the choice of parameters for solution ions. In accord with the empirical choices recommended in the original ABSINTH work, we set the reference free energies of solvation for the sidechains of Asp, Glu, Arg, and Lys to be more favorable by 30 kcal/ mol vis-à-vis the default values that are based on estimates from experiments. This choice, as explained in the published literature, ensures against artifacts due to the formation of spurious salt-bridges and the use of a fixed charge model (the charges are fixed by the pK_a values of amino acids at neutral pH) rather than a constant pH simulation paradigm (15). Simulations of sequences with proline residues utilize the modified and tested parameters for bond angles, dihedral angles, and Lennard-Jones interactions as described by Radhakrishnan et al. (10).

Details regarding the MMC move sets: The degrees of freedom for the Metropolis Monte Carlo (MMC) simulations include the backbone torsion angles ϕ, ψ, and ω, the sidechain torsion angles χ and the rigid body coordinates of polypeptides and solution ions. The move sets include translation of ions combined with small- and large-scale conformational changes of the polypeptide degrees of freedom. The latter are achieved through a combination of local, pivot, and concerted moves, and their frequencies were prescribed in direct analogy to the decision tree used in previous work (16). A summary of the move sets and their frequencies is presented in Table S3. Conformations of sequences with proline residues utilize the improved move sets published by Radhakrishnan et al. (10) and these include the deformation of pyrrolidine rings, the modeling of ring puckering, and the coupling between ring puckering and backbone degrees of freedom.

Table S3: Details of the move sets used to sample conformational space in MMC simulations. Here, f_Δ denotes the fraction of moves assigned to finite perturbations, whereas the remaining moves attempt full randomization of the respective degrees of freedom. The fourth column denotes maximum step-sizes corresponding to finite perturbations. Rigid body moves have both translational and rotational step-sizes. A small fraction of moves was invested in swaps between nearest-neighbor thermal replicas.

Move Type	**Frequencies**	f_Δ	**step- sizes (max)**
Rigid-body	5%	50%	2 Å / 10^o
Side-chain χ	19%	60%	4^a / 30^o
Concerted rotation	7.6%	-	-
ω	6.84%	90%	5^o
Backbone Pivots ϕ,ψ	49.25%	70%	10^o
Proline ring puckering	12.31%	80%	4^o (dihedral) / 2^o (angular)

[a]Refers to the maximum number of χ-angles perturbed simultaneously for a sidechain move.

Simulation set up: We performed three types of simulations. One set of simulations denoted as EV, FRC, and LJ in the figure legends of the main text corresponds to atomistic reference models for all sequences. In the EV limit, the only interactions included in the simulations are those due to the pairwise repulsions that are part of a standard 12-6 Lennard-Jones potential. Details of simulation results for generic polypeptides in the EV limit have been presented elsewhere (17, 18). The FRC refers to the Flory random coil. Ensembles for this limit were obtained by drawing backbone and sidechain dihedral angles from a database of prior simulations of dipeptides N-acetyl-Xaa-N′-methylamide for all residues Xaa. Conformations are constructed using detailed models for dipeptides and ignoring all interactions beyond the dipeptide. This ensures high fidelity to local conformational preferences and ignores the effects of long-range interactions. Distributions of sequence-specific random globules were generated using simulations based on the Lennard-Jones 12-6 potential and ignoring all other terms in the energy functions following procedures described previously (3, 9, 19).

For simulations with the full ABSINTH model and forcefield, all polypeptides and ions were modeled in atomic detail. The simulations were carried out using spherical boundary conditions. In each simulation, the system comprised the polypeptide chain, neutralizing Na^+, Cl^- ions plus excess ion pairs to mimic 15 *m*M NaCl enclosed within a spherical droplet of radius 75Å. To assess the sensitivity of conformational properties to changes in salt concentration we performed additional simulations for three variants of $(Glu\text{-}Lys)_{25}$ in an excess salt concentration of 125 *m*M NaCl. The choice of 75Å for the droplet radius was justified using the end-to-end distance distributions for all sequences simulated in the EV limit. These simulations revealed that none of the sequences sampled conformations with end-to-end distances larger than 150Å. This argues against artifacts due to confining effects posed by small droplet sizes. The cut-off distances for van der Waals interactions and for electrostatic interactions between charge groups that are net neutral were set to 10Å and 14Å, respectively. No cut-offs were used for electrostatic interactions involving mobile ions and the charged sidechains such as those of Asp, Glu, Arg, and Lys. Finally, the IS limit simulations were preformed by retaining all terms of the ABSINTH potential function except W_{el}.

Thermal Replica Exchange simulations: To enhance the quality of sampling thermal replica exchange (TREx) simulations were utilized. For all the simulations, the temperature schedule comprised fourteen temperatures: [280K, 286K, 293K, 298K, 305K, 315K, 325K, 335K, 350K, 365K, 380K, 395K, 410K, and 430K]. The choice of the temperature schedule was justified from the computed overlap statistics between neighboring temperature replicas (Figure S21). For a pair of windows X and X+1, the overlap fraction was defined as $1-\frac{\int_{E=E_{\min}}^{E=E_{\max}} | P_X(E)-P_{X+1}(E) | \, dE}{2}$. [1] The acceptance ratio between neighboring thermal replicas was always greater than 0.3 (on average) indicating high reliability of the sampling quality (Figure S21). For each thermal replica the simulation was initiated using a conformation drawn at random from a prior simulation in the EV limit. To improve the overall statistics and to assess reproducibility, there were at least three independent TREx simulations for each polypeptide simulated. The standard deviation in the mean across independent TREx simulation was computed to quantify the reproducibility of our results. Each TREx simulation had 4.65×10^7 and 5.15×10^7 MC steps for (Glu-Lys)$_{50}$ permutants and naturally occurring IDP sequences, respectively with the first 1.5×10^6 being discarded as equilibration steps. Swaps between two neighboring replicas were attempted every 50,000 steps.

Simulation analysis: Trajectories were saved every 5,000 steps for simulations of (Glu-Lys)$_{50}$ permutants and every 4,000 steps for simulations of naturally occurring IDP sequences. Polymeric properties were computed every 500 steps and saved every 20,000 steps. Internal distances were computed every 1,000 steps.

Protocol for generating sequence variants for a fixed sequence composition: For a given sequence composition a set of variants with different κ values was generated using the Wang-Landau algorithm (20, 21). The algorithm was used to compute the density of sequence variants $n(\kappa)$ with a given κ. For each move, a swap between two randomly chosen amino acid residues at two different positions was proposed and the new κ (κ_{new}) was computed. The move is accepted if $p < \min\left\{1, \frac{n(\kappa_{\mathrm{new}})}{n(\kappa_{\mathrm{old}})}\right\}$, where p is a pseudo-random number within interval [0, 1]. After each move, the visit histogram $H(\kappa)$ and the density of states $n(\kappa)$ were updated as $H(\kappa) = H(\kappa)+1$ and $\ln n(\kappa) = \ln n(\kappa) + \ln f_i$, respectively. f_i, the modification factor, is a constant and its initial value was set to $\exp(1)$. For a f_i, when the $H(\kappa)$ satisfied the flatness

[1] The formula introduce here to calculate overlaps between energy distributions for pairs of thermal replicas is also used to calculate the overlaps between pairs of p(r) histograms as shown in Figure S4.

criterion, it was reset to zero and the modification factor was set to $f_{i+1} = \sqrt{f_i}$. The simulation was continued until f_{i+1} satisfied the convergence criteria. To avoid any possible bias, originating from the starting permutant, in the generated variants, multiple independent sets of variants were generated using the same approach with the difference being the starting value of κ. This ensured that the combined sets of sequence variants spanned the whole range of κ-values for a given sequence composition.

Annotation of the modified diagram of states: Sequences were obtained from DisProt (5) version 5.9 (released on 02-23-2012) using the following filters: i) *N* (sequence length) > 30, ii) NCPR < 0.3, iii) Fraction of proline residues < 0.3, iv) below the parametric line of Uversky et al. (6) in the charge-hydropathy plot.

Sequence details of naturally occurring IDPs:

PQBP-1: **UniProt ID - O60828:** residues 132-183 of polyglutamine tract binding protein-1.

Sequence:
WPPDRGHDKSDRDRERGYDKVDRERERDRERDRDRGYDKADREEGKERRHHRREE

PQB-P1 is a nuclear protein that is crucial for transcription and RNA processing. These functions are affected due to binding of PQBP-1 with expanded polyglutamine tracts.

(Organism: *Homo sapiens*)

DisProt ID - DP00166: UniProt ID: P19429: residues 163-210 of Troponin I.

Sequence: AKESLDLRAHLKQVKKEDTEKENREVGDWRKNIDALSGMEGRKKKFES

Troponin I monitors calcium levels in the muscle and initiates muscle contractions.

(Organism: *Homo sapiens*)

DisProt ID - DP00357: UniProt ID: P62328: the beta-thymosin/WH2 actin binding domain that is important in regulating actin dynamics and cell motility.

Sequence: WPPMSDKPDMAEIEKFDKSKLKKTETQEKNPLPSKETIEQEKQAGES

(Organism: *Homo sapiens*)

DisProt ID - DP00503: UniProt ID: P40259-1: cytoplasmic domain **(**residues 181-229) of the Ig beta chain of B-cell antigen receptor (BCR) complex.

Sequence:
WPPLDKDDSKAGMEEDHTYEGLDIDQTATYEDIVTLRTGEVKWSVGEHPGQE

(Organism: *Homo sapiens*)

UniProt ID - Q5SH22: Alpha-aminoadipate carrier protein lysW; important for biosynthesis of L-Lysine.

Sequence:
WPPMVGTSPESGAELRLENPELGELVVSEDSGAELEVVGLDPLRLEPAPEEAEDWGE
(Organism: *Thermus thermophilus*).

Disprot ID - DP00246: UniProt ID: P21758-1: cytoplasmic domain (residues 1-50) of Macrophage scavenger receptor types I and II (isoform 1).

Sequence: MAQWDDFPDQQEDTDSCTESVKFDARSVTALLPPHPKNGPTLQERMKSYK

(Organism: *Bos taurus*)

Disprot ID - DP00011: UniProt ID: P50224: residues 216-261 of Catecholamine sulfotransferase 1A3/1A4

Sequence: PEETMDFMVQHTSFKEMKKNPMTNYTTVPQELMDHSISPFMRKGMA

Catecholamine sulfotransferase sulfonates catecholamines as a part of detoxification pathway leading to the formation of readily excretable water soluble metabolites.

(Organism: *Homo sapiens*)

Disprot ID - DP00438: UniProt ID: Q05158: linker region (residues 68-115) of Cysteine and Glycine rich CRP proteins.

Sequence: PKGYGYGQGAGTLNMDRGERLGIKPESSPSPHRPTTNPNTSKFAQKFG

CRP proteins are important for regulatory processes connected to cell growth and differentiation.

(Organism: *Coturnix coturnix japonica*)

Disprot ID - DP00436: UniProt ID: P50477: residues 1-50 of Canavalin.

Sequence: WPPMAFSARFPLWLLLGVVLLASVSASFAHSGHSGGEAEDESEESRAQ

Canavalin is a major storage protein of jack beans.

(Organism: *Canavalia ensiformis*)

Disprot ID - DP00348: UniProt ID: P45481: ACTR binding domain (residues 2059-2117) of CREB-binding protein that is important for transcriptional activation.

Sequence:
PNRSISPSALQDLLRTLKSPSSPQQQQVLNILKSNPQLMAAFIKQRTAKYVANQPGMQ

(Organism: *Mus musculus*)

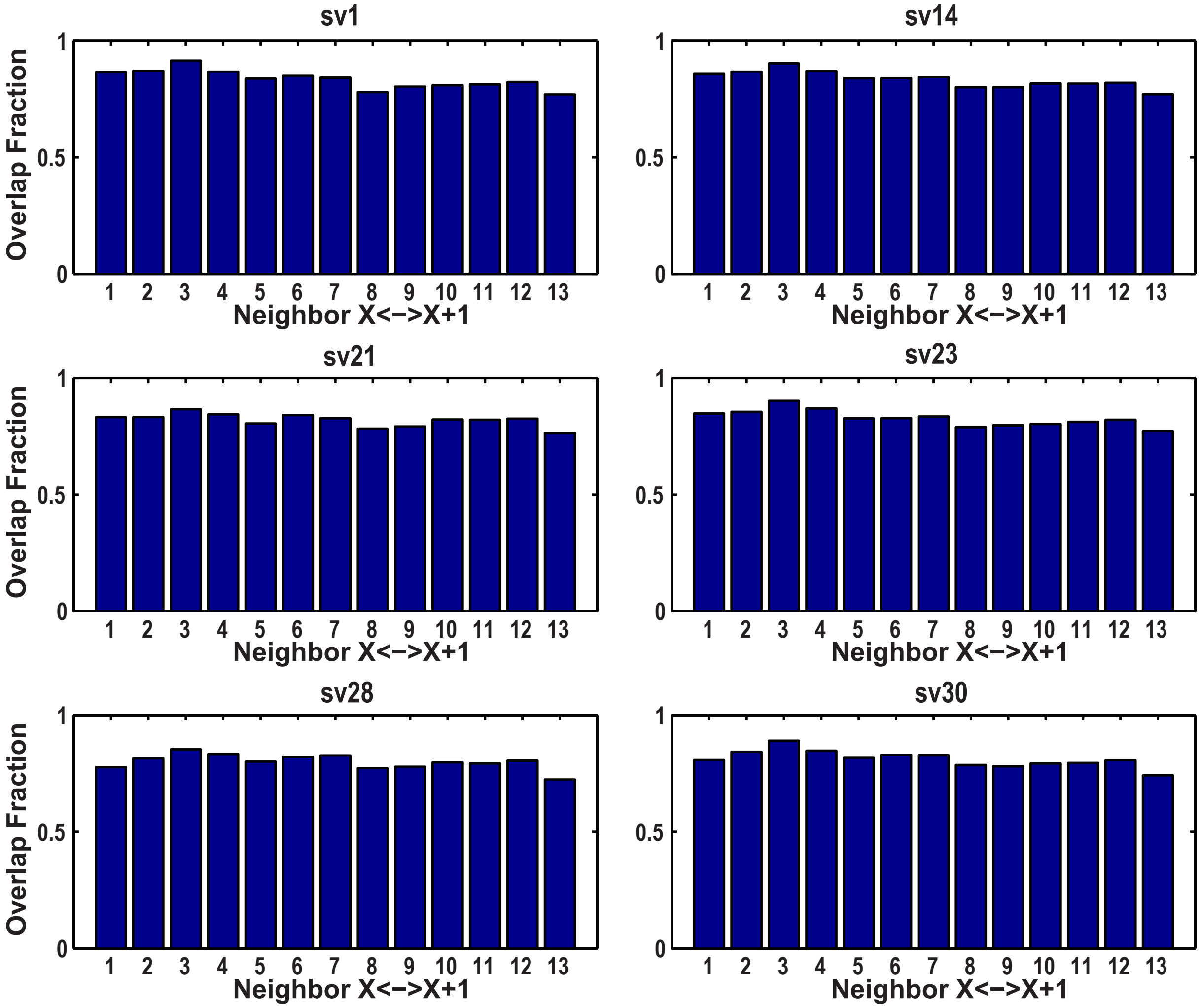

Figure S21: Overlap statistics between neighboring temperature replicas in TREx simulations of different (Glu-Lys)$_{50}$ sequence variants. The statistics are shown for six different sequence variants with low, intermediate and high κ. The bars show the mean overlap statistics over three independent TREx runs.

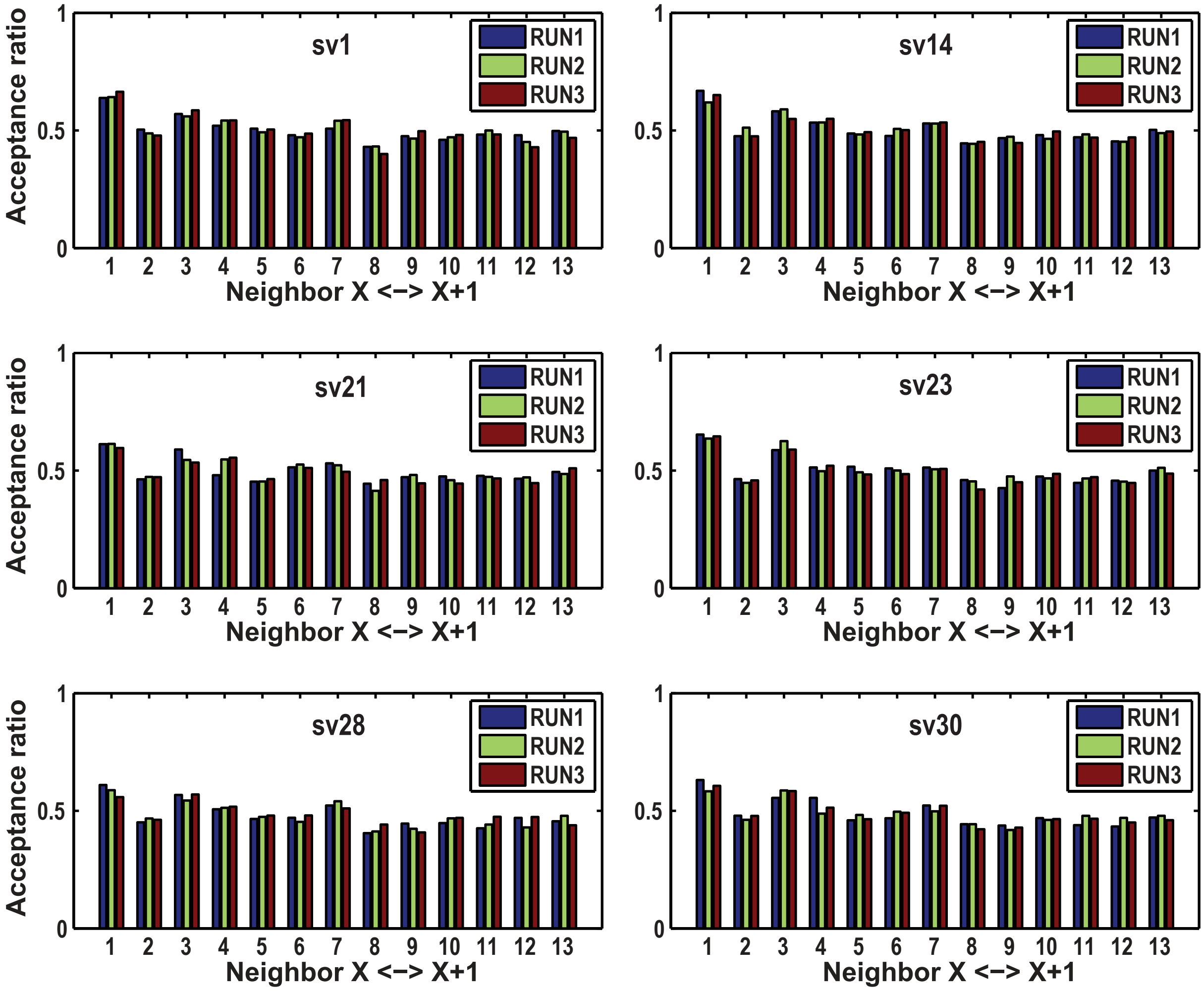

Figure S22: Acceptance ratios of swaps between nearest-neighbor thermal replicas in TREx simulations of six different (Glu-Lys)$_{50}$ sequence variants, with low, intermediate and high κ values. The three bars denote acceptance ratios for three independent TREx runs.

References

1. Nettels D*, et al.* (2009) Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proc. Natl. Acad. Sci. U. S. A.* 106(49):20740-20745.
2. Best RB & Mittal J (2010) Protein Simulations with an Optimized Water Model: Cooperative Helix Formation and Temperature-Induced Unfolded State Collapse. *J. Phys. Chem. B* 114(46):14916-14923.
3. 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. U. S. A.* 107(18):8183-8188.
4. Humphrey W, Dalke A, & Schulten K (1996) VMD: visual molecular dynamics. *J. Mol. Graphics Model.* 14(1).
5. Sickmeier M*, et al.* (2007) DisProt: the database of disordered proteins. *Nucleic Acids Res.* 35:D786-D793.
6. Uversky VN, Gillespie JR, & Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins Struct. Func. Genet.* 41(3):415-427.
7. Ishida T & Kinoshita K (2008) Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics* 24(11):1344-1348.
8. Uversky VN (2002) What does it mean to be natively unfolded? *Eur. J. Biochem.* 269:2-12.
9. Vitalis A & Pappu R (2009) ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* 30(5):673-699.
10. Radhakrishnan A, Vitalis A, Mao AH, Steffen AT, & Pappu RV (2012) Improved atomistic Monte Carlo simulations demonstrate that poly-L-proline adopts heterogeneous ensembles of conformations of semi-rigid segments interrupted by kinks. *J. Phys. Chem. B* 116(23):6862-6871.
11. Wyczalkowski MA, Vitalis A, & Pappu RV (2010) New estimators for calculating solvation entropy and enthalpy and comparative assessments of their accuracy and precision. *J. Phys. Chem. B* 114(24):8166-8180.
12. Kaminski GA, Friesner RA, Tirado-Rives J, & Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B* 105(28):6474-6487.
13. Åqvist J (1990) Ion Water Interaction Potentials Derived from Free-Energy Perturbation Simulations. *J. Phys. Chem.* 94(21):8021-8024.
14. Mao AH & Pappu RV (2012) Crystal lattice properties fully determine short-range interaction parameters for alkali and halide ions. *J. Chem. Phys.* 137(6).
15. Khandogin J & Brooks CL (2005) Constant pH molecular dynamics with proton tautomerism. *Biophys. J.* 89(1):141-157.
16. Das RK, Crick SL, & Pappu RV (2012) N-terminal segments modulate the alpha-helical propensities of the intrinsically disordered basic regions of bZIP proteins. *J. Mol. Biol.* 416(2):287-299.
17. Tran HT, Wang X, & Pappu RV (2005) Reconciling observations of sequence-specific conformational propensities with the generic polymeric behavior of denatured proteins. *Biochemistry* 44(34):11369-11380.
18. Tran HT & Pappu RV (2006) Toward an accurate theoretical framework for describing ensembles for proteins under strongly denaturing conditions. *Biophys. J.* 91(5):1868-1886.

19. Vitalis A, Wang X, & Pappu RV (2007) Quantitative characterization of intrinsic disorder in polyglutamine: Insights from analysis based on polymer theories. *Biophys. J.* 93(6):1923-1937.
20. Wang FG & Landau DP (2001) Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Phys. Rev. E* 64(5).
21. Wang FG & Landau DP (2001) Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* 86(10):2050-2053.