SUPPORTING INFORMATION APPENDIX

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

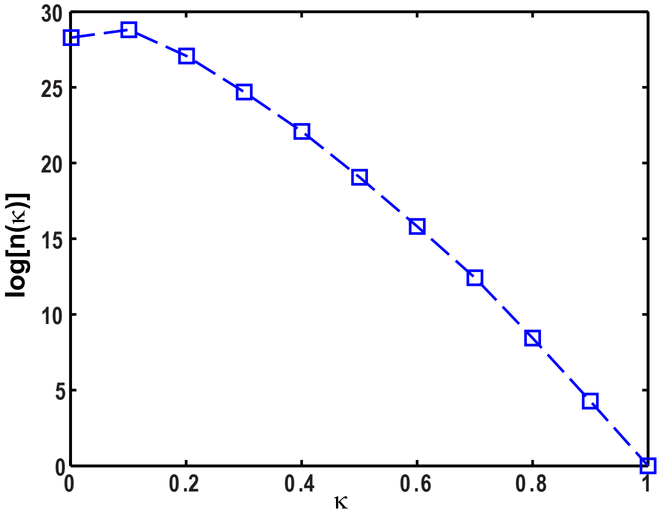

Figure S1: Semi-log plot of the number density of sequence variants of (Glu-Lys)$_{25}$, n(κ), that have similar κ values.

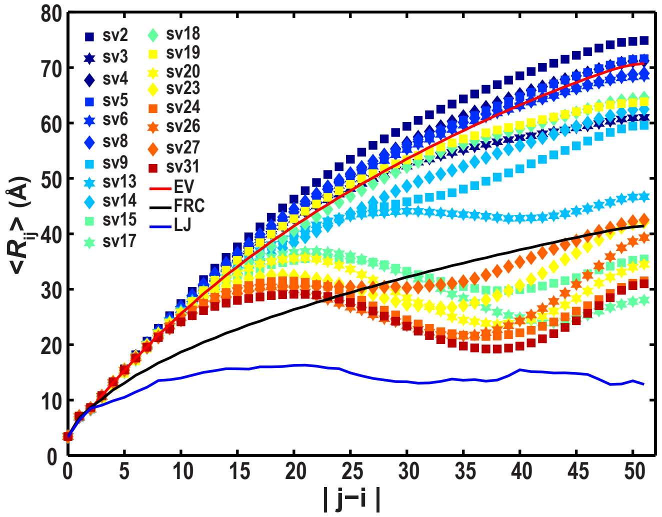

Figure S2: $\langle R_{\mathrm{ij}}\rangle$ profiles for nineteen of the thirty sequence variants from Figure 1 of the main text. Profiles for the remaining sequence variants are shown in panels (A) and (B) of Figure 3 of the main text. The profiles are colored according to their κ values; the cooler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

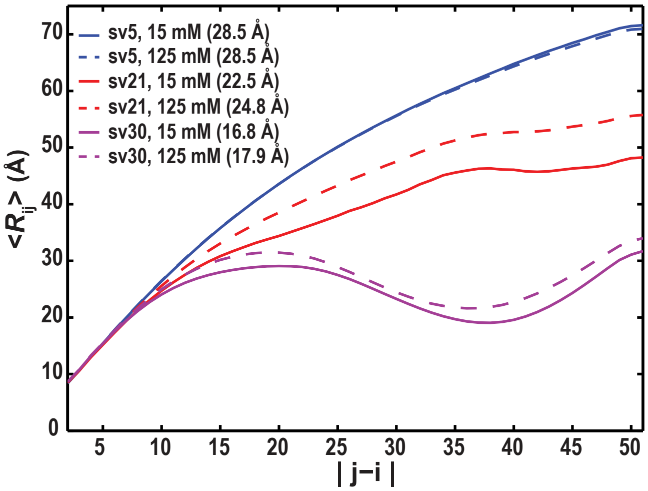

Figure S3: Quantification of the salt dependence of conformational properties for three of the sequence variants shown in Figure 1. The colors identify the sequence variants and the line types identify the profiles in 15 mM NaCl (solid), and 125 mM NaCl (dashed). In each case, the error bars (not shown 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.

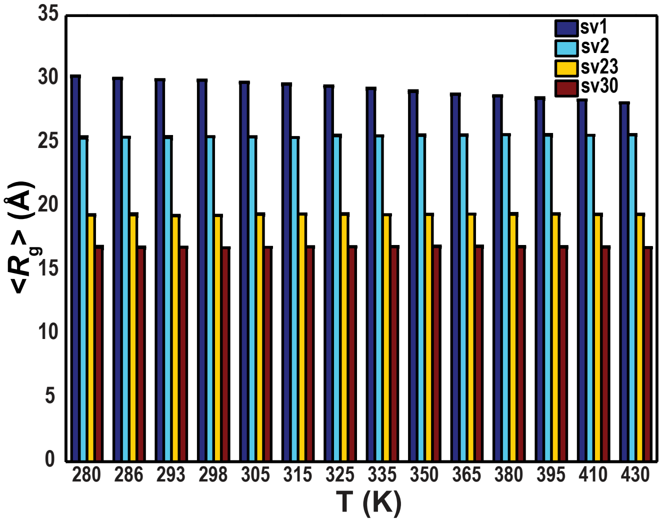

Figure S4: Temperature dependence of $\langle R_g^2\rangle$ for different sequence variants of (Glu-Lys)$_{25}$ (15 mM NaCl). Each stack of bars corresponds to a single temperature and the colors pertain to different sequence variants. Figures S4 and S5 summarize the temperature dependence of conformational 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

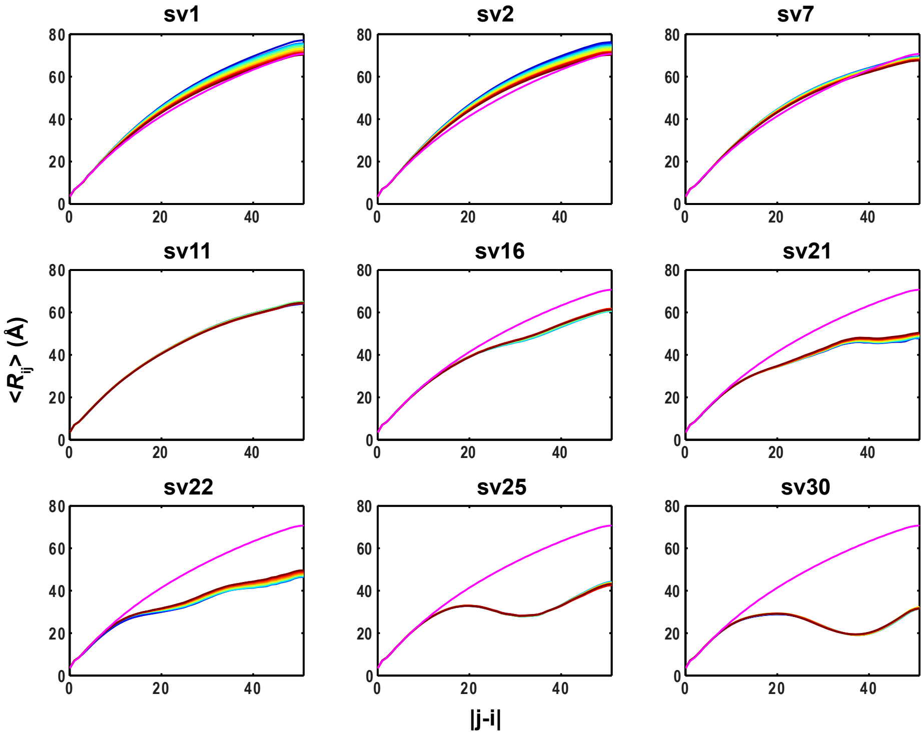

Figure S5: Temperature dependence of the $\langle R_{ij} \rangle$ profiles for selected sequence variants of (Glu-Lys)$_{25}$ to illustrate the weak contraction / expansion of variants with low versus intermediate values of κ. This contraction / expansion results from increased conformational fluctuations 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 $\langle R_{ij} \rangle$ 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.

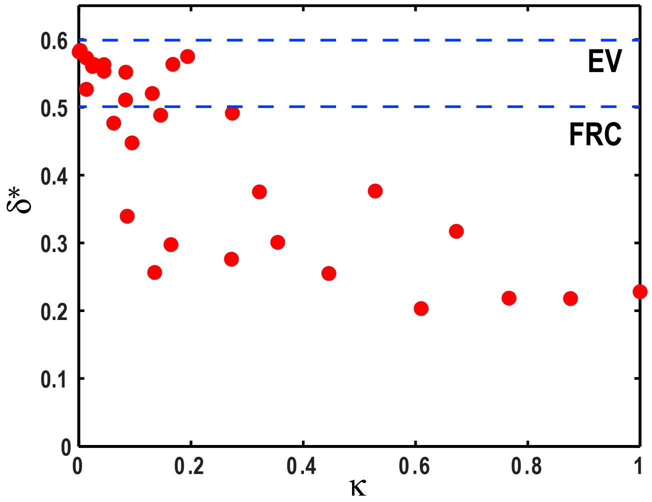

Figure S6: Plot of ensemble-averaged asphericity values δ^* against κ for sequence variants of (Glu-Lys)$_{25}$. For each sequence, δ^* is calculated from the ensemble of eigenvalues for gyration tensors that describe individual conformations as described in previous work (3). The two dashed lines intersect 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.

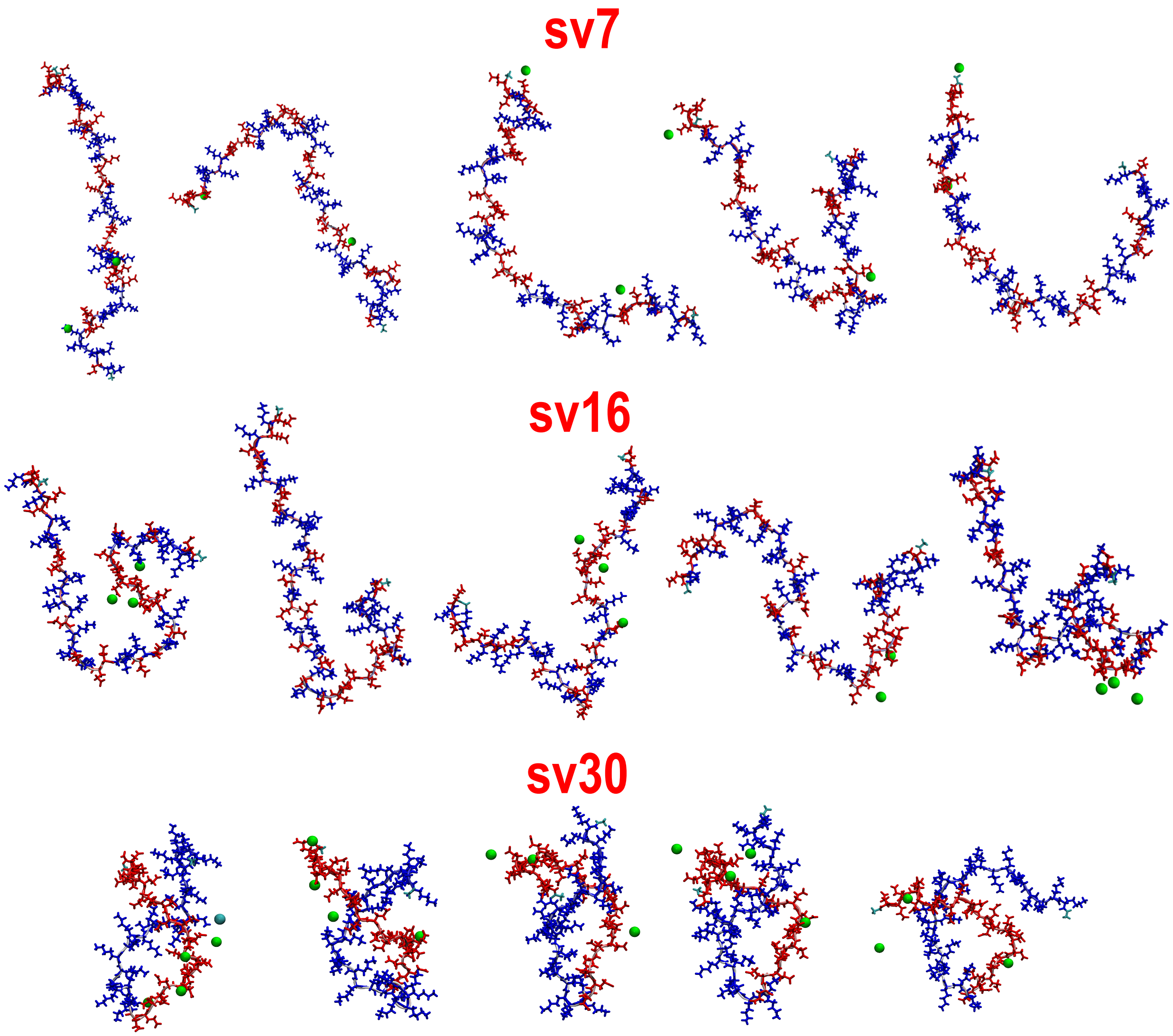

Figure S7: Montage of representative conformations drawn at random from individual Markov chain Metropolis Monte Carlo trajectories (T=298 K and [NaCl] = 15 mM) for different sequence variants of $(\text{Glu-Lys})_{25}$. In this montage, the lysine and glutamate residues are shown in blue and red, respectively. 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).

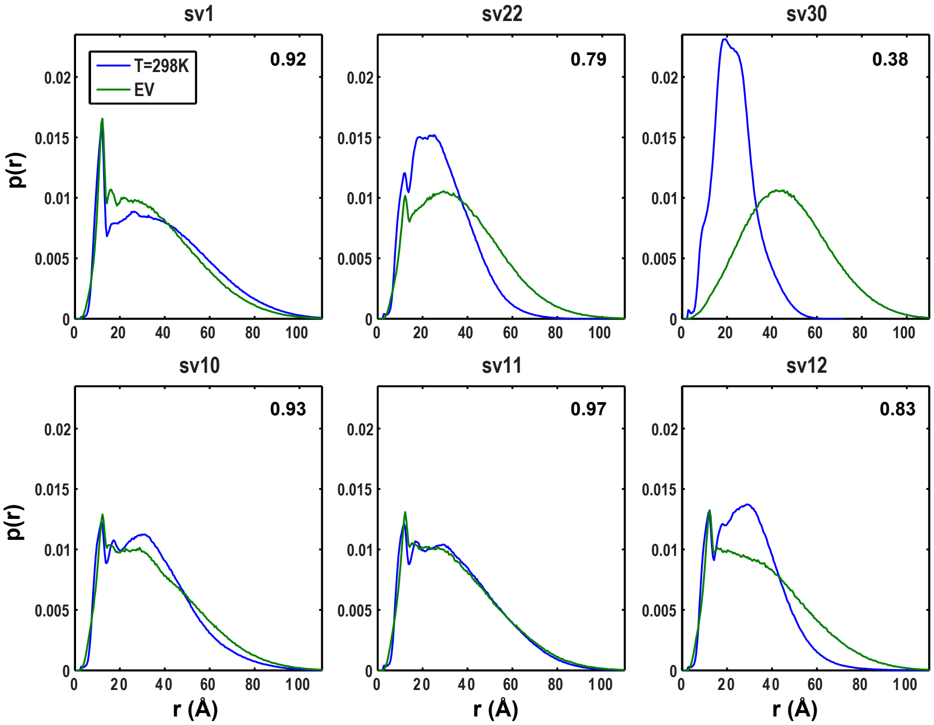

Figure S8: Comparison of histograms p(r) of distances r between the amino and carboxyl sidechain tips of Lys and Glu residues in different sequence variants of the (Glu-Lys)$_{25}$ system. In each panel, the number on the top right corner quantifies the overlap between p(r) histograms calculated in the 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.

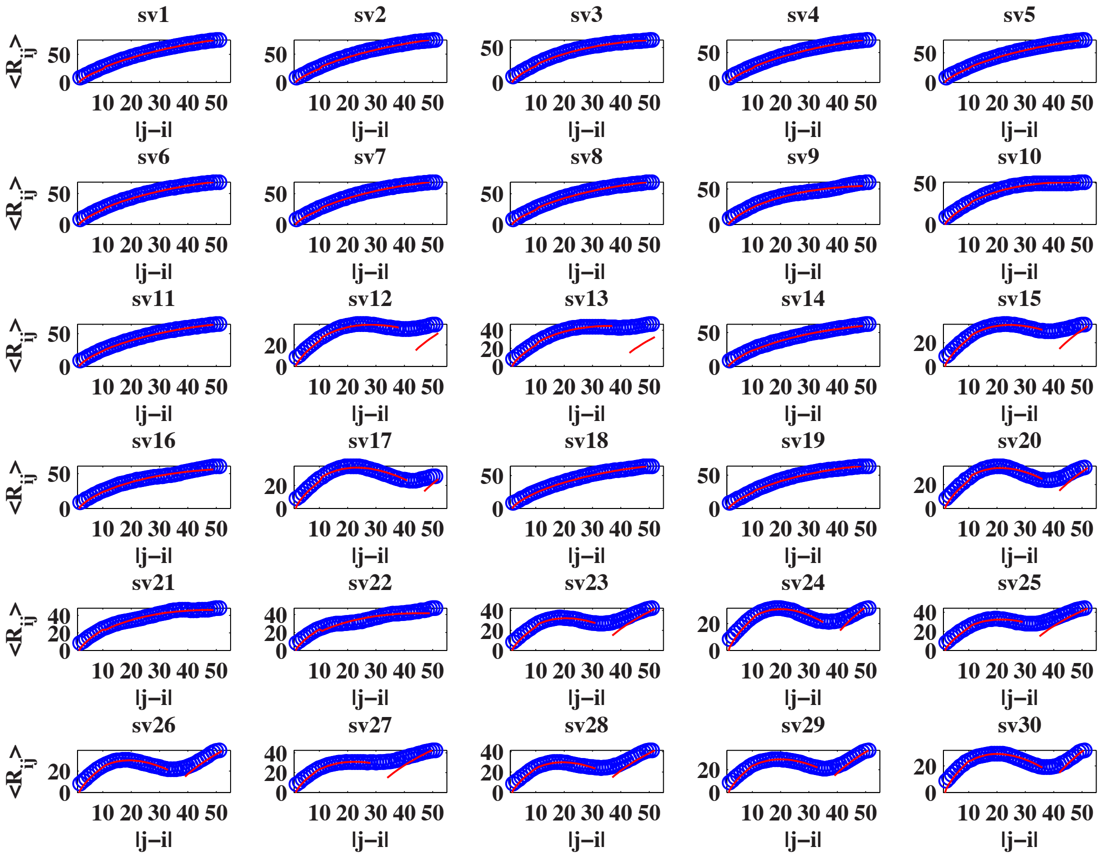

Figure S9: Fits to $\langle R_{\mathrm{ij}} \rangle$ profiles for thirty of the sequence variants of (Glu-Lys)$_{25}$. Details regarding the κ values for each variant are shown in Figure 1 of the main text. The blue circles are the raw data for $\langle R_{\mathrm{ij}} \rangle$ profiles and the red curves 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 $2g \le |\mathrm{j}–\mathrm{i}| \le l_c$ 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.

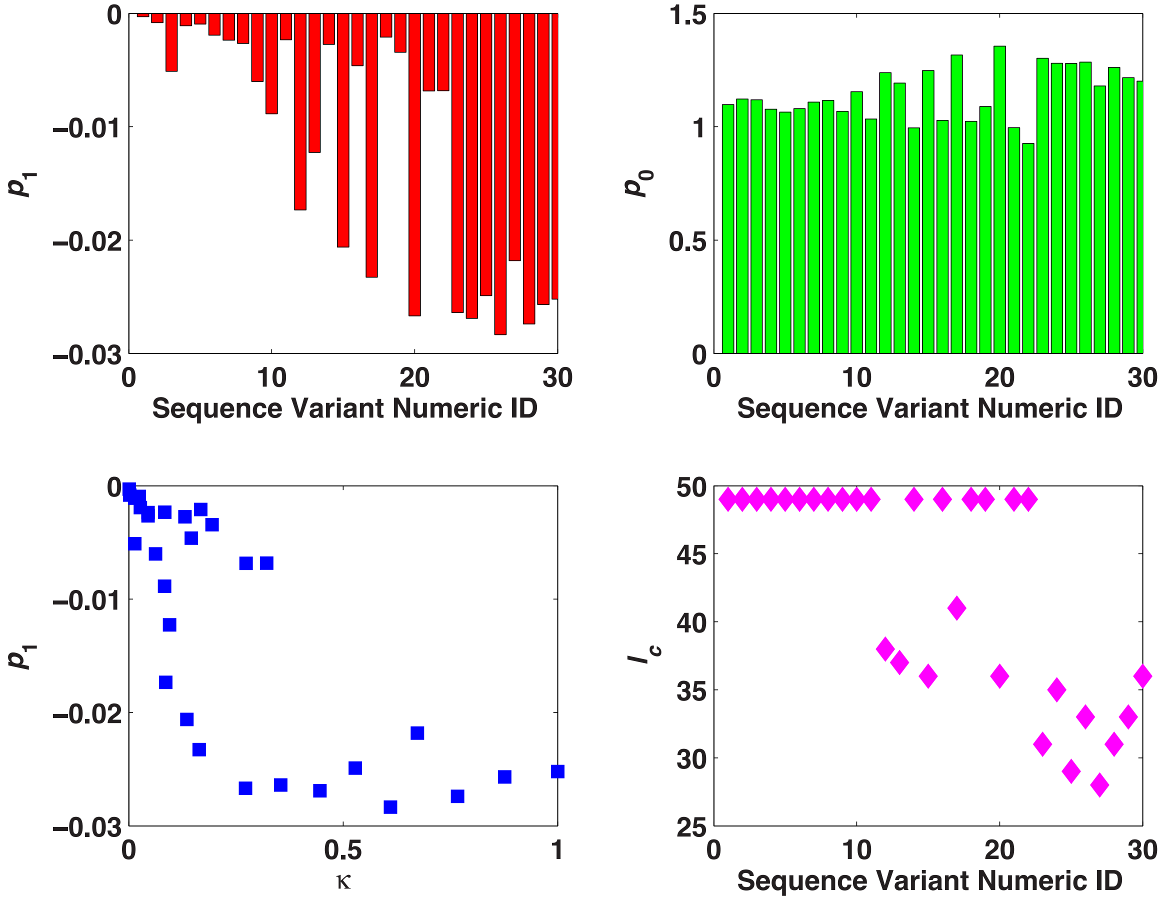

Figure S10: Parameters obtained from numerical fitting of scaling form to results for $\langle R_{ij} \rangle$ profiles for all sequence variants of (Glu-Lys)$_{25}$ shown in Figure 1 of the main text. The top left and top right panels show bar plots of the parameters p_1 and p_0, respectively. In these 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.

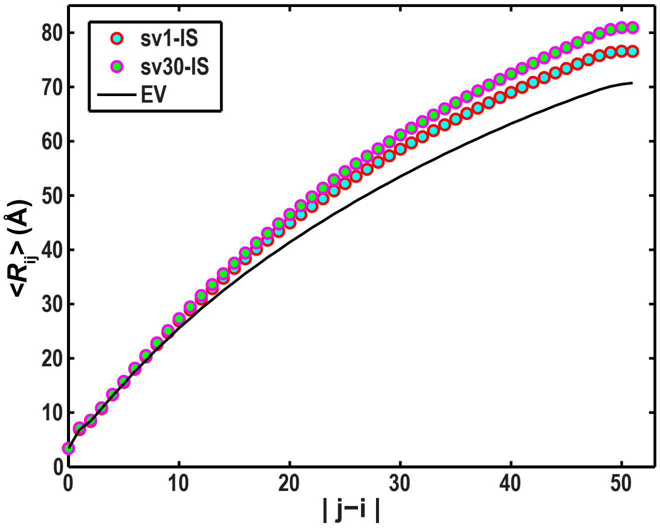

Figure S11: IS limit $\langle R_{\mathrm{ij}} \rangle$ profiles for two of the sequence variants of (Glu-Lys)$_{25.}$ These profiles are compared to that of the EV limit

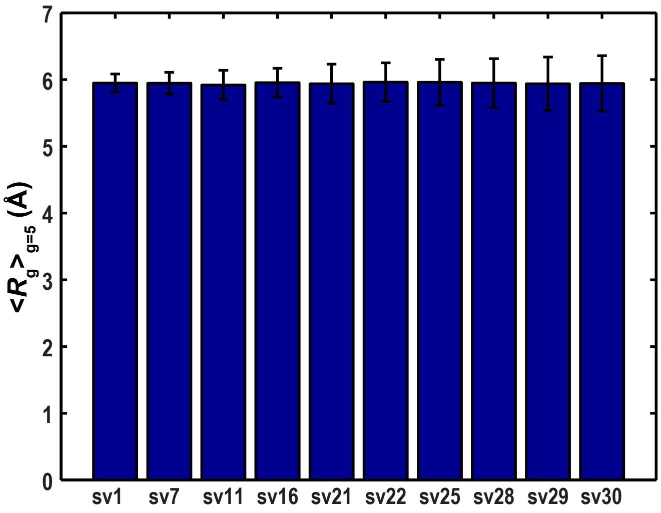

Figure S12: Demonstration of the validity of the blob concept. The plot shows average radii of gyration calculated for segments of length g=5 extracted from different sequence variants of (Glu-Lys)$_{25}$. The average R_g for blobs, denoted as ξ in the main text, is ≈6Å irrespective of the sequence from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

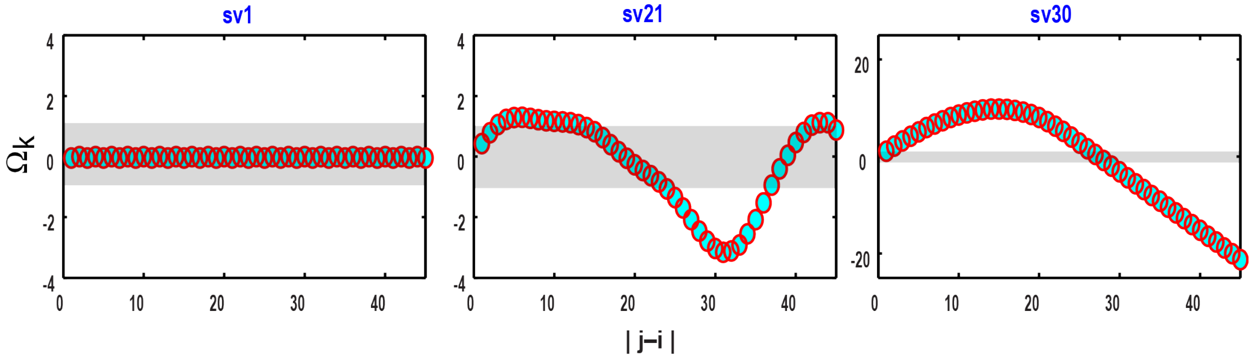

Figure S13: Plots of cumulative sums of the length scale specific Coulomb coupling parameters, $\Omega_k = \sum_{k=1}^{|j-i|} < \Gamma_k >$ for three different sequence variants of (Glu-Lys)$_{25}$. The ordinate corresponds to the cumulative sum of $\mathbf{\Gamma_{ij}}$ and the height of the gray bars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

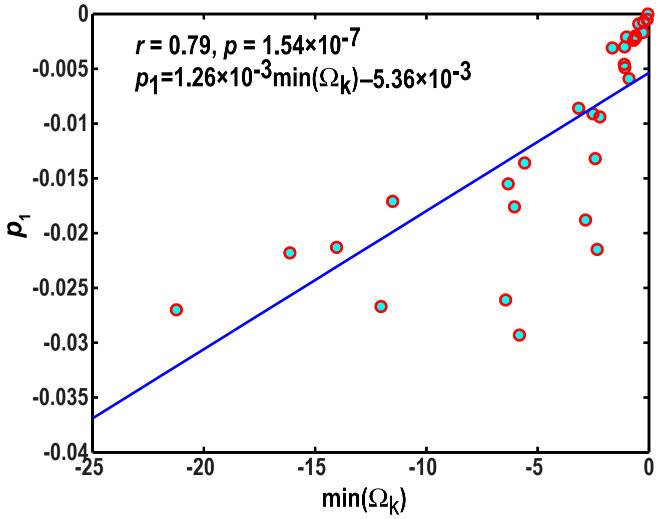

Figure S14: Plot of p_1 against min(Ω_k) for sequence variants of (Glu-Lys)$_{25}$. See caption to Figure S13 for definition of Ω_k. The Pearson *r* value, the *p*-value that quantifies the probability of realizing this correlation at random, and the equation for the line of best fit (shown in blue)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

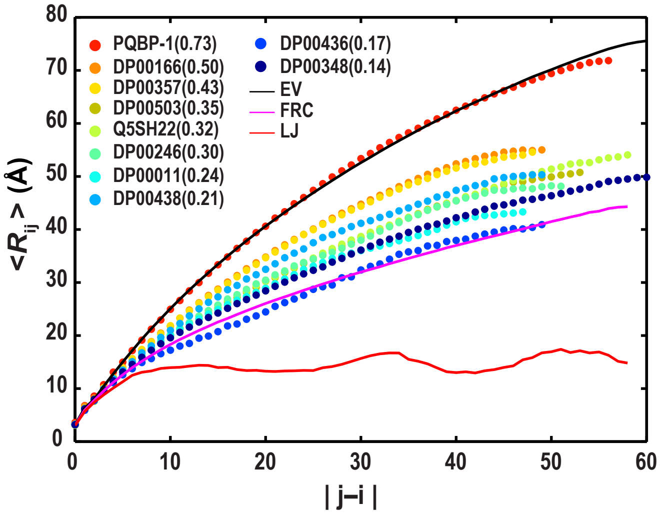

Figure S15: $\langle R_{\text{ij}}\rangle$ profiles for the ten naturally occurring IDPs (see Table S2) simulated in the IS limit. For comparison, the plot includes profiles for the EV limit, Flory random coil, and a maximally compact LJ globule. The legend shows the sequence identifiers and the FCR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

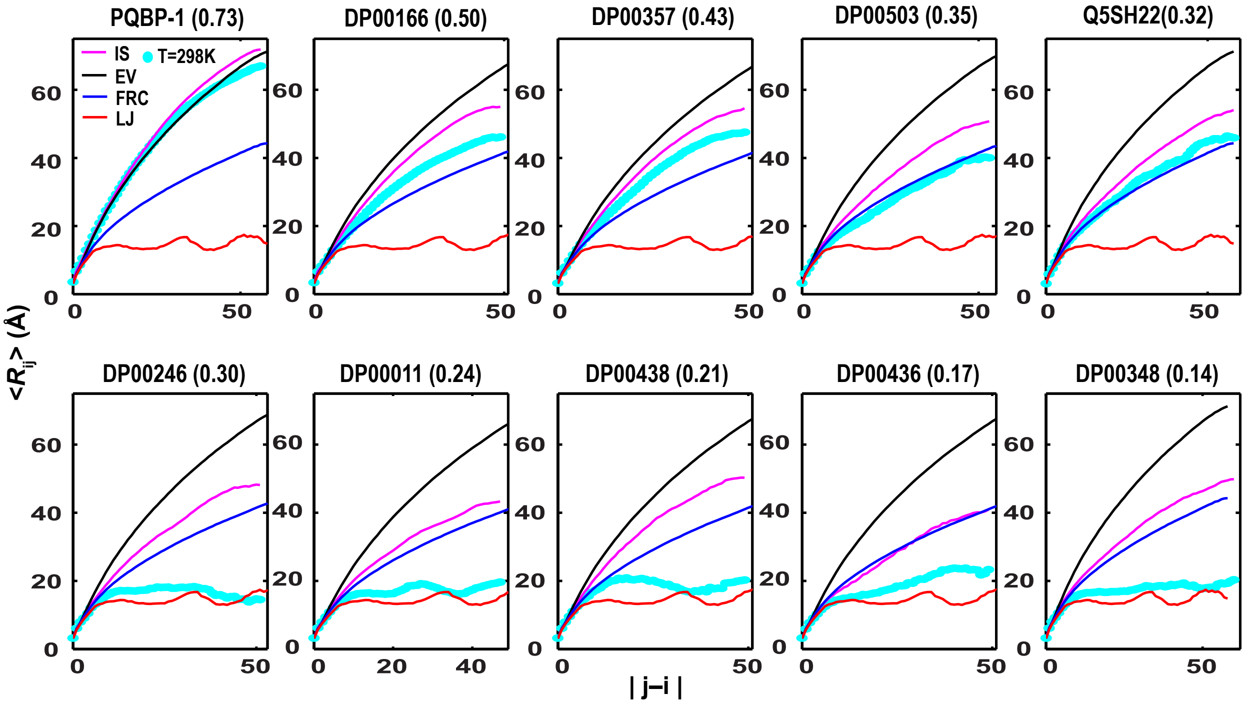

Figure S16: Comparison of IS limit profiles to those obtained using simulations based on the full ABSINTH Hamiltonian for all ten naturally occurring IDPs studied in this work. In each panel, the IS limit profile is shown in magenta, the EV limit profile in black, the profile for Flory random coils in 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.

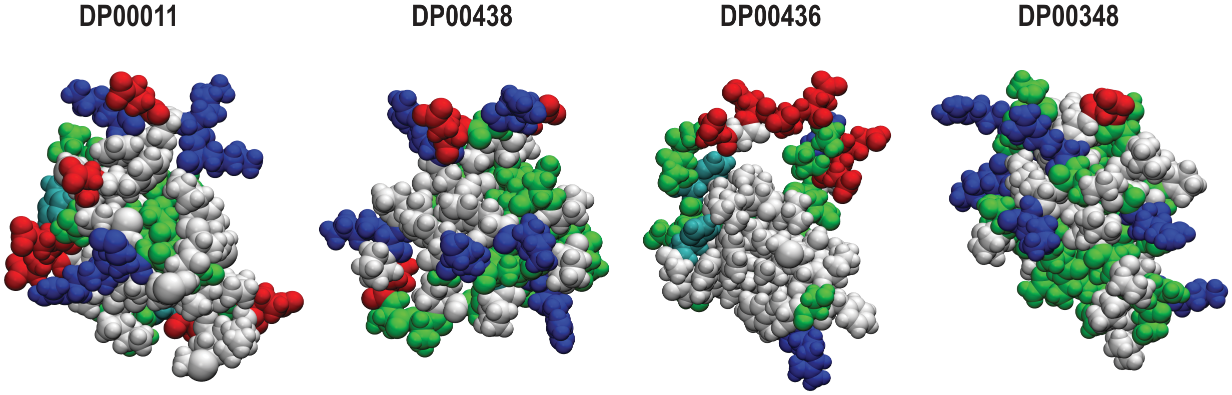

Figure S17: Space filling pictures of representative conformations drawn from the simulated ensembles for weak polyampholytic IDPs. The DisProt identifiers for each sequence are shown in the figure and sequence details are available from Table S2 and section 2 of this *SI Appendix*. The color coding 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).

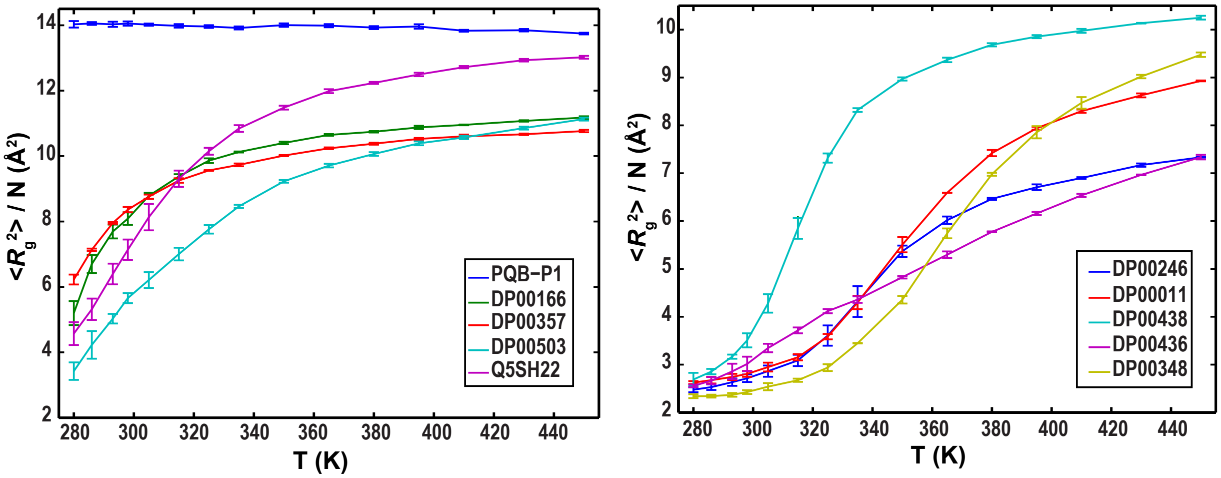

Figure S18: Temperature dependence of $\langle R_g{}^2\rangle/N$ for the ten naturally occurring IDPs. In the interest of clarity we set the scales for the ordinates to be different between the two panels. These results show increased sensitivity of $\langle R_g{}^2\rangle/N$ to changes in temperature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

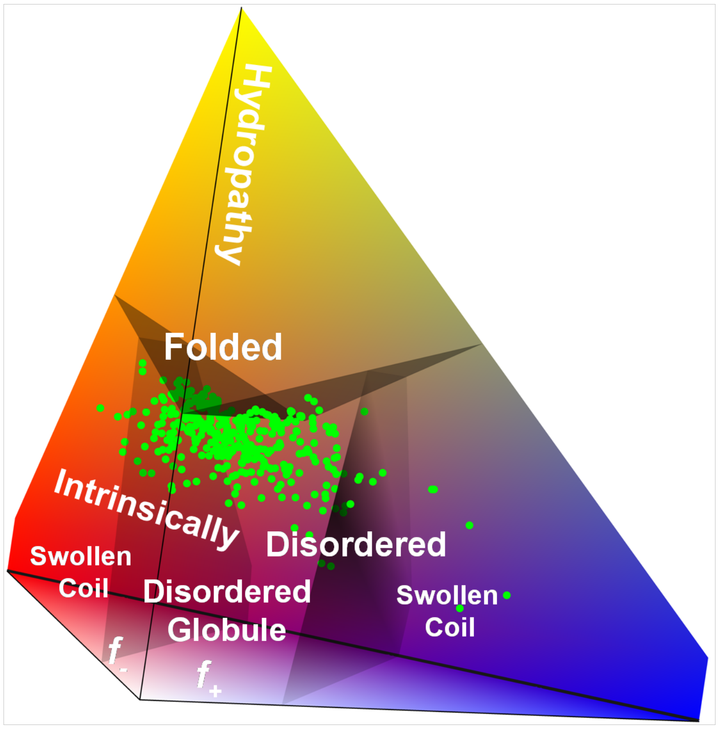

Figure S19: Diagram of states of Mao et al. (3) annotated by a subset of sequences drawn from the DisProt database (5). The three axes denote f_+, f_-, and the hydropathy and all three parameters are calculated from the amino acid composition. The diagram of states of Mao et al. and the designations 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.

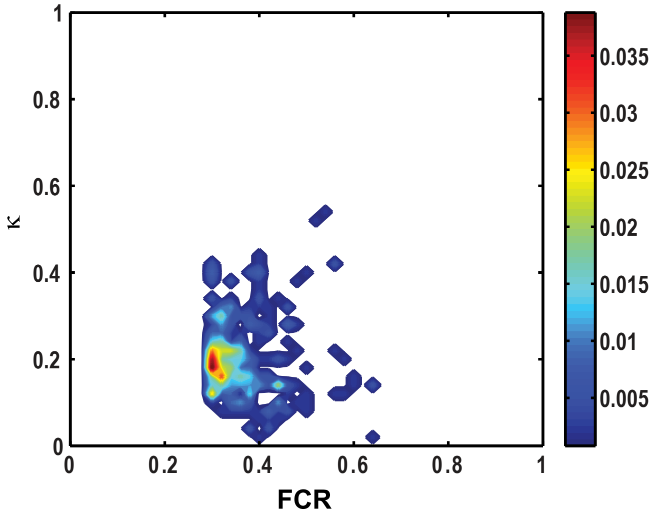

Figure S20: Joint density distribution ρ(FCR,κ) for a subset of the 364 sequences drawn from the DisProt database. As noted in the main text, this analysis focuses on sequences that satisfy the constraints $N > 30$, FCR ≥ 0.3, NCPR < 0.25, and fraction of proline residues < 0.3.

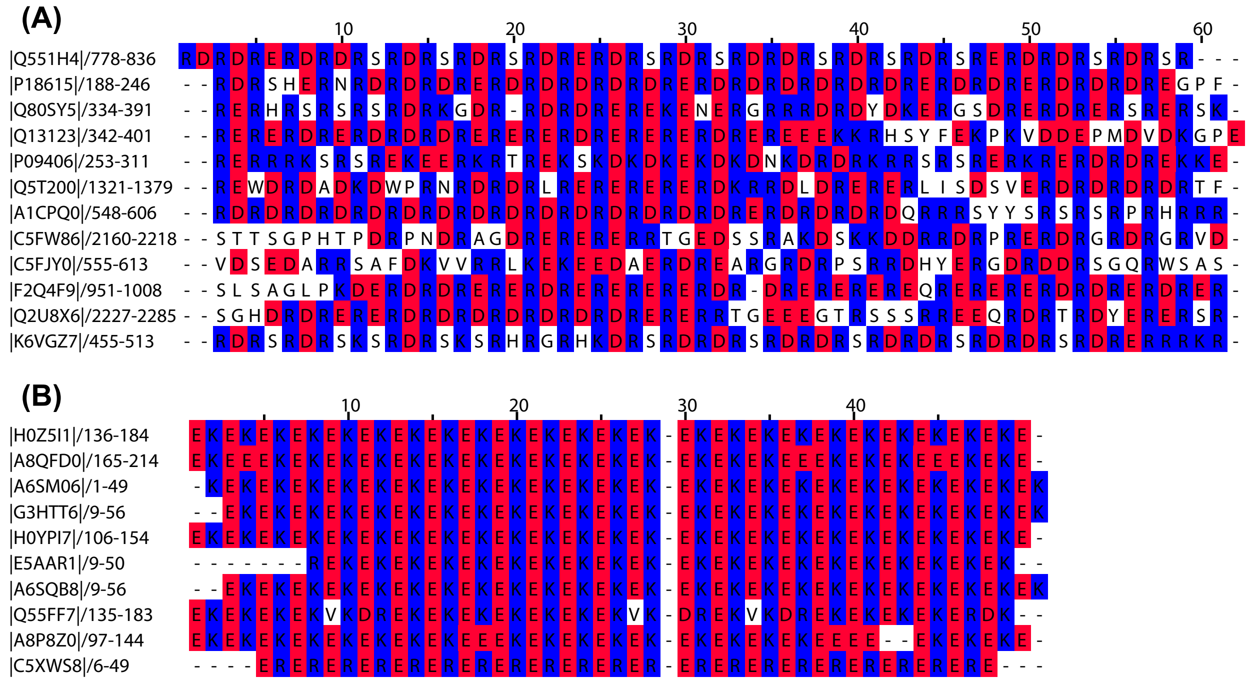

Figure S21: Representative strong polyampholytic regions drawn from sequence databases. Panel A shows examples of naturally occurring sequences that contain long stretches of Asp/Glu-Arg repeats. We retrieved these sequences by performing a BLAST search (http://blast.ncbi.nlm.nih.gov/Blast.cgi) of the 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	MSDAAVDTSSEITTKDLKEKKEVVEEAENGRDAPANGNAQNEENGEQEADNEVDEEEEEGGEEEEEEEEGDGEEEDGDEDEEAEAPTGKRVAEDDEDDDVETKKQKKTDEDD
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*)

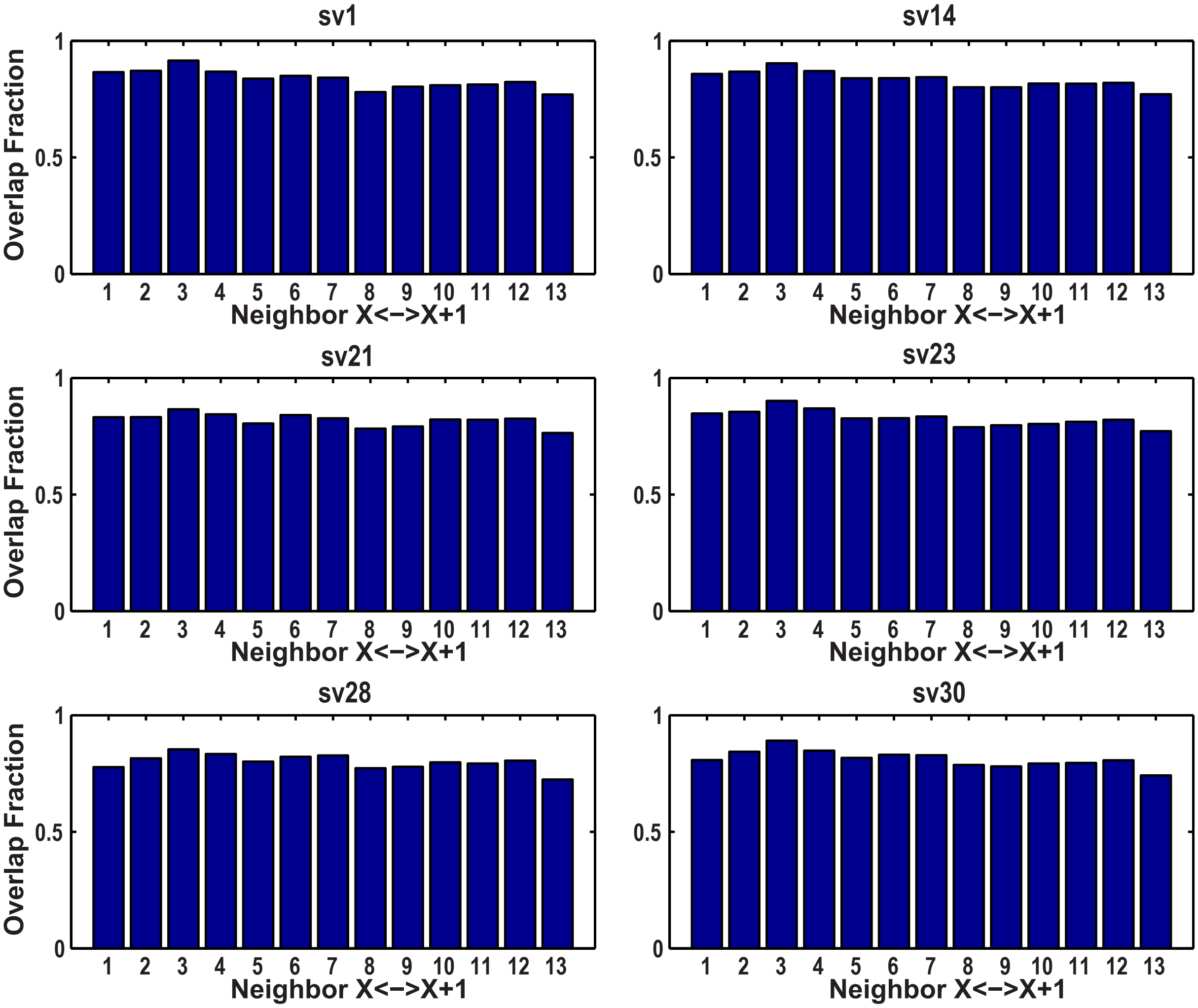

Figure S21: Overlap statistics between neighboring temperature replicas in TREx simulations of different (Glu-Lys)$_{50}$ sequence variants. The statistics are shown for six different sequence variants with low, intermediate and high κ. The bars show the mean overlap statistics over three independent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

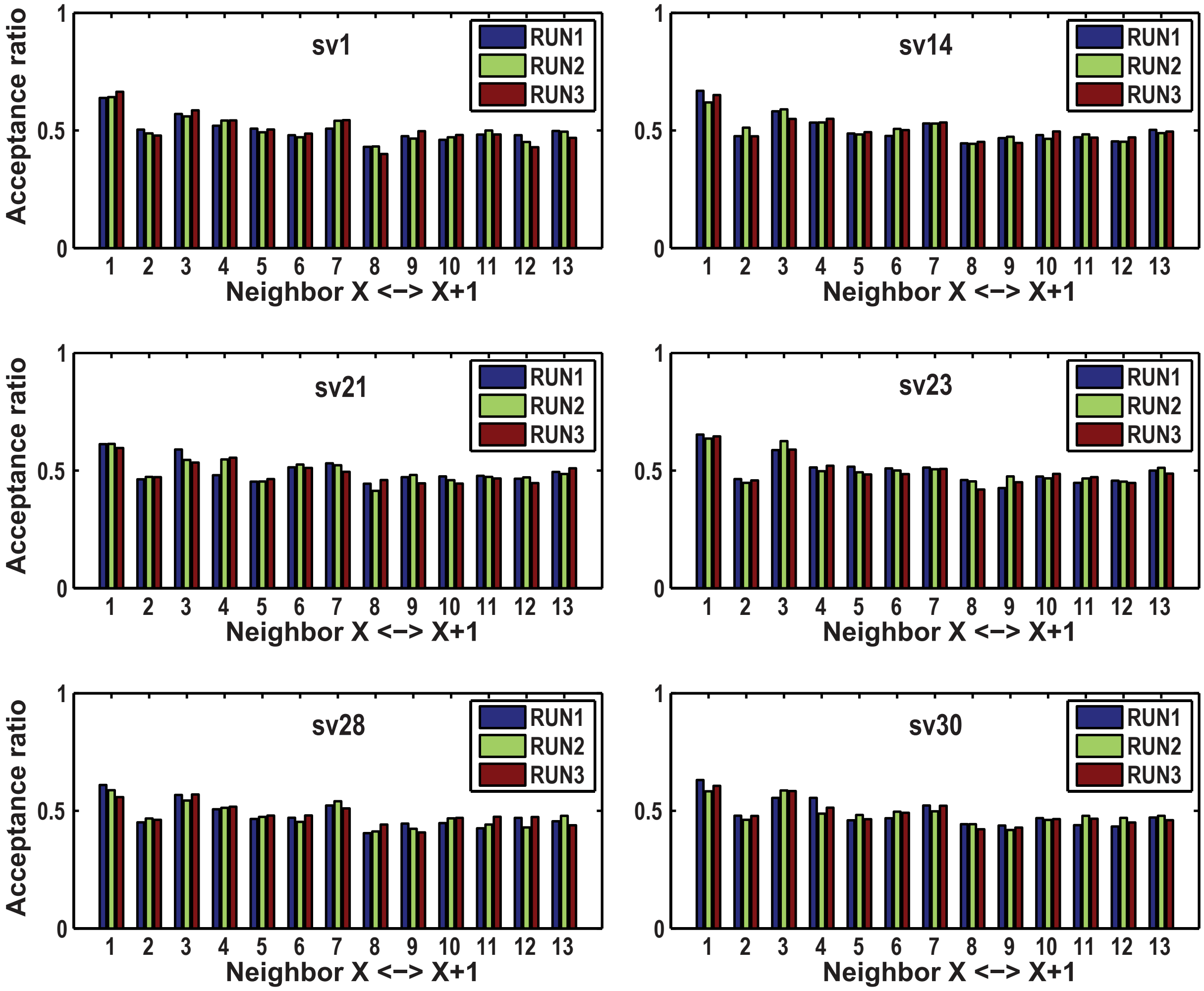

Figure S22: Acceptance ratios of swaps between nearest-neighbor thermal replicas in TREx simulations of six different (Glu-Lys)$_{50}$ sequence variants, with low, intermediate and high κ values. The three bars denote acceptance ratios for three independent TREx runs.

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.