Biophysical Journal, Volume *112*

Supplemental Information

CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins

Alex S. Holehouse, Rahul K. Das, James N. Ahad, Mary O.G. Richardson, and Rohit V. Pappu

CIDER: Resources to analyze and classify sequence-ensemble relationships of intrinsically disordered proteins

Supporting Information

Alex S. Holehouse*, Rahul K. Das, James N. Ahad, Mary O. G. Richardson, and Rohit V. Pappu*

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

*Corresponding Authors

Running title: CIDER: Resources to Analyze IDPs

Contents

This supporting information contains an extensive discussion on the analysis of the complete set of IDPs identified in sixteen model organisms. In addition, we provide extended technical information on CIDER and localCIDER, and offer a detailed discussion on the parameter κ. Section 1 defines a number of parameters used in the manuscript and throughout the supporting information. Sections 2 – 5 describe results from the analysis of IDPs from sixteen model organisms, and set the stage for further investigation. Sections 6 – 8 provide addition details regarding CIDER and localCIDER. Finally, section 9 includes a detailed and pedagogical overview of the parameter κ

(kappa), including discussion of its underlying statistical form and ways to compute expected values given a sequence composition.

1. Abbreviations and parameter definitions

Included below are definitions of a number of parameters used throughout this work.

f_+ – Fraction of positively charged residues in a sequence (between 0 and 1)

f_- – Fraction of negatively charged residues in a sequence (between 0 and 1)

FCR – Fraction of charged residues (between 0 and 1)

$$\mathrm{FCR} = (f_+ + f_-)$$

NCPR – Net charge per residue (between -1 and 1)

$$\mathrm{NCPR} = (f_+ - f_-)$$

2. Data used for analysis

The following organisms were included in our full proteome analysis: *H. sapiens*, *R. norvegicus*, *M. musculus, G. gallus*, *A. thaliana, D. rerio, D. melanogaster, C. elegans, P. falciparum, D. discoideum, N. crassa, S. cerevisiae, S. pombe, C. albicans*, *B. subtilis*, and *E. coli*.

All proteomic data were obtained from the UniProt reference proteomes (1), downloaded from the EBI FTP server (http://www.ebi.ac.uk/reference_proteomes). A list of these proteomes is provided at the end of this subsection. DisProt sequences were taken from the DisProt download (DisProt Release 7.03), which after redundancy filtering includes 744 disordered fragments of over 30 residues (2).

Disorder data for each proteome was taken from the MobiDB 2.0 consensus prediction data (3). MobiDB combines disorder predictions from ten disorder predictors. A consensus prediction is generated as a majority vote based on those ten predictors, with a classification of ‘disordered’ or ‘structured’ assigned to each residue in each protein from the proteome. The result of this consensus disorder prediction was then post-

processed to remove short islands (≤ 3 residue) of disorder or order to create a less fragmented set of regions. Specifically, if an identified region - either a disordered region or a structured region - was found to be less than four residues long it was converted into the type of its surrounding regions. We compared results with and without this post-processing and found no difference in terms of the parameters reported in this study, though clearly this post-processing influences the number and size of IDRs identified, an aspect not examined in this work. The presence of short islands of order within disordered regions is primarily an artifact of combining multiple semi-overlapping predictors, as illustrated by figure S1. The threshold of three or fewer residues was selected as a value of half the thermal blob length-scale (6-7 residues,(4)), i.e., substantially shorter than a length scale over which persistent structure would be expected.

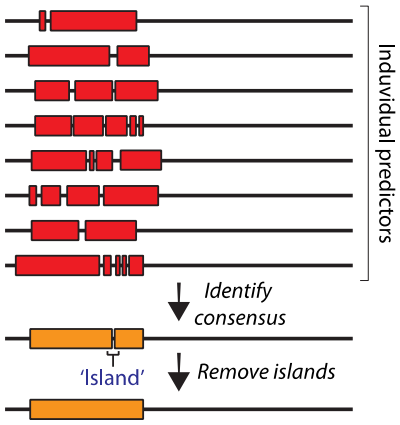

Figure S1: Schematic showing the creation of a consensus disordered region from multiple predictors, followed by the removal of a short 'order' island. Disorder predictors typically generate profiles with multiple short interruptions.

The use of a consensus score, rather than relying on a single disorder predictor, helps to avoid any intrinsic biases in various predictors. It creates a more stringent threshold for defining a region as disordered, but ensures that, to the best of our ability, regions predicted as 'disordered' are utilizing approaches from multiple predictors to avoid false positives. In retrospective analysis we repeated much of the work done here using a single disorder predictor (IUPred (5)) and found highly analogous results (data not shown) suggesting that IUPred provides a robust stand-alone prediction.

MobiDB provides a consensus prediction based on a set of ten disorder predictors (3). In addition to these ten predictors, MobiDB also allows for the inclusion of structural

information from the Protein Data Bank (PDB) to further annotate structural preferences within a region. For our analysis, we used the MobiDB consensus data from disorder predictors alone, rather than also including additional information from the PDB. This decision was made based on two considerations. Firstly, many IDPs are known to undergo coupled folding and binding. The PDB contains a large number of structures representing protein regions that have been shown to fold in the context of a partner, and while relevant for function, this does not appear to be relevant for knowing the region's intrinsic structural propensity as an autonomous unit. As a result, MobiDB's approach of using structural data to categorically rule out a region as disordered is highly appealing, but may unintentionally yield false negatives in some circumstances. As a specific example, the protein PUMA (p53 up-regulated modulator of apoptosis) is predicted to be disordered, and yet the mouse variant (UniprotID Q99ML1) contains a region (residues 130-155) that has been structurally characterized by NMR (PDB ID 2ROC) (6). As a result of this apparently alpha-helical region, MobiDB defines this region as structured (Figure S2A). However, this region adopts a stable helix only upon binding to its partner Mcl1 (Figure S2B), and has been experimentally shown to be disordered in the unbound state (7), although recent studies show a roughly 40% likelihood that PUMA adopts helical conformations in its unbound form (8).

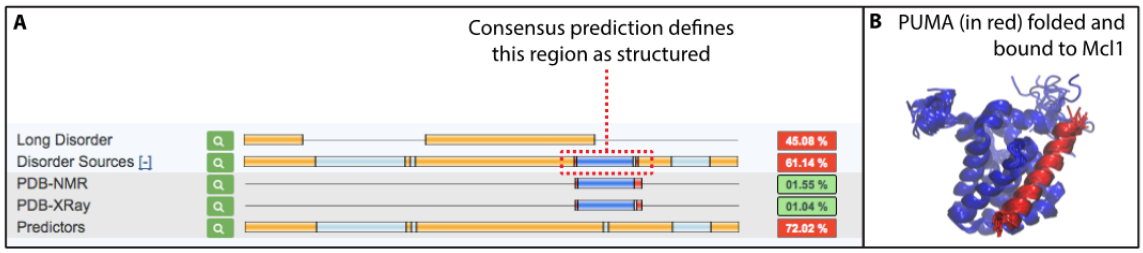

Figure S2: Example of structural data incorrectly informing on a folded region. Panel A shows a screenshot from the MobiDB website, and highlights the fact that the full consensus prediction approach used assigns the region in blue to be folded. Panel B shows the NMR structure of PUMA bound to Mcl1 (PDB ID 2ROC).

Secondly, the number of PDB structures available varies significantly between different organisms. The number of entries per organism is shown in Table S1 below. Note that this does not capture the structural redundancy (*i.e.*, there are frequently many structures of the same protein) but does provide a general overview of the inequality of depositions in the database. These numbers are based on an analysis performed in August 2015.

Organism	Number of PBD structures
Homo sapiens	37183
Rattus norvegicus	2612
Mus musculus	6834
Gallus gallus	1635
Arabidopsis thaliana	952
Danio rerio	202
Drosophila melanogaster	905
Caenorhabditis elegans	306
Plasmodium falciparum	649
Dictyostelium discoideum	153
Neurospora crassa	69
Saccharomyces cerevisiae	4485
Schizosaccharomyces pombe	334
Candida albicans	123
Escherichia coli	13538
Bacillus subtilis	1510
Thermotoga maritima	655

Table S1: Summary of the number of structures in the PDB by organism

Given that the analysis carried out in this study compares multiple proteomes, we felt that it was important to use a uniform approach across all the primary (sequence) data. If structural data were included, we would intrinsically bias sequences from organisms that have been studied in greater structural detail, towards being more likely to identify structured regions. This could have the unintended consequence of allowing a region to be classed as disordered in one organism but structured in another *if* structural data had been obtained in one species but not in the other.

Having obtained the set of disordered regions associated with a proteome, each disordered region greater than thirty residues and with a proline content of less than 15% was used for further analysis. A threshold of thirty residues was chosen to match the general consensus of ‘long’ disorder (5). The threshold of 15% for proline content is in keeping with the original definition of the diagram-of-states (4), and the fact that a growing body of evidence suggests that enrichment in proline drives ensembles to be more expanded than one might naïvely expect based on FCR alone (9-11). The influence of proline residues is explored systematically in other work, and the patterning of charged and proline residues is quantified by the parameter Ω (9).

Proteome-wide statistics for various quantities are shown in Table S2. The "*percentage 'long' disorder*" represents the percentage of the proteome from each organism which is encompassed by a single disordered region stretching thirty residues or longer. Other estimates in the literature do not use this 30 residue threshold (and use a less stringent disorder classification), and as such find a much higher percentage of disorder in the proteomes from these organisms.

Reference proteome	**Organism**	**Number proteins**	**Num. disorder regions (>30 res)**	**Percentage 'long' disorder**
UP000005640_9606	*Homo sapiens*	20 882	23 437	18.6%
UP000002494_10116	*Rattus norvegicus*	21 866	21 529	17.4%
UP000000589_10090	*Mus musculus*	22 129	22 448	17.3%
UP000000539_9031	*Gallus gallus*	15 749	16 949	16.6%
UP000006548_3702	*Arabidopsis thaliana*	27 221	17 192	11.5%
UP000000437_7955	*Danio rerio*	25 642	25 125	16.5%
UP000000803_7227	*Drosophila melanogaster*	13 674	15 489	19.8%
UP000001940_6239	*Caenorhabditis elegans*	20 274	12 716	13.1%
UP000001450_36329	*Plasmodium falciparum*	5 162	7 274	14.6%
UP000002195_44689	*Dictyostelium discoideum*	12 732	13 703	20.3%
UP000001805_367110	*Neurospora crassa*	9 756	11 927	23.9%
UP000002311_559292	*Saccharomyces cerevisiae*	6 720	5 381	14.6%
UP000002485_284812	*Schizosaccharomyces pombe*	5 104	3 407	11.7%
UP000000559_237561	*Candida albicans*	8 354	6 450	16.6%
UP000000625_83333	*Escherichia coli*	4 305	274	1.15%
UP000001570_224308	*Bacillus subtilis*	4 197	382	1.75%
UP000008183_243274	*Thermotoga maritima*	1 851	47	0.42%

Table S2: Summary proteomic statistics relevant for this study

These data represent a total of 243,644 proteins, 203,683 disordered regions, and an average “percentage long proteome disorder” of 16.6% in eukaryotes and 1.45% in non-hyperthermophilic prokaryotes (*E. coli* and *B. subtilis*). The significant depletion of long disordered regions in the hyperthermophile *T. maritima* is in line with previous work (12). With so few disordered regions in *T. maritima,* it was excluded from further analysis in this study to avoid the introduction of misleading biases.

3. Key results from the analysis of proteome-scale disordered regions from multiple organisms

We first examined how disordered regions are distributed across the diagram-of-states (Figure S3A). For the human, rat, mouse, and chicken proteomes the distribution across R1-R5 was highly similar, and generally matched the DisProt distribution. For other organisms (notably *D. melanogaster*, *P. falciparum*, and a number of fungi) large deviations from the distribution seen in humans were observed. In all cases, relatively few polyelectrolytes (R4/R5) were identified, and those found were almost exclusively negatively charged. We also found that the fraction of charged residues (FCR) varied between different organisms (Figure S3B), as do the distributions of κ values (Figure S3C). Taken together, these results show that the distribution of charge density and patterning vary across organisms, although similar global trends are also observed. An additional takeaway from this analysis is that by these measures, DisProt encompasses a good representation of IDPs for describing the sequences in the human proteome. One could have imagined that DisProt might have been enriched in charged IDPs, but this analysis firmly shows that DisProt provides a representative snapshot of the human IDPs, at least in terms of amino acid composition.

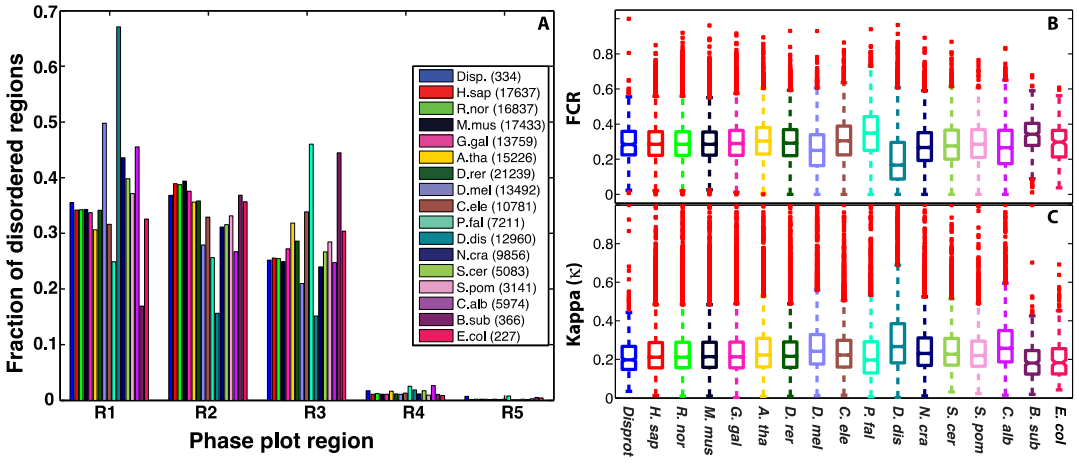

Figure S3: Sequence properties of disordered regions across sixteen proteomes and DisProt. Legend numbers indicate the total number of disordered regions identified. S3A shows fractional populations of the diagram-of-states regions for all IDPs from sixteen different model organisms and the DisProt database. While broadly similar trends are observed, there are substantial differences between different organisms. S3B is a box-plot showing the distribution of FCR values for all IDPs taken from the same set of organisms and DisProt. The central box defines the first quartile, median, and third quartile from the data. Similarly, S3C is a box plot showing the distribution of κ values taken from the same set of organisms and DisProt.

Having established that the median FCR for disordered regions varies across different organisms, we asked if the distribution of κ values observed for naturally occurring sequences varied with FCR. To answer this question, we focused on polyampholytic sequences (absolute net charge per residue |NCPR| < 0.25, FCR > 0). Based on anecdotal evidence, κ appears has the most significant influence on the conformational behavior of sequences which display the dual traits of an intermediate-to-high FCR and a near neutral overall charge – *i.e.,* strong polyampholytes. For comparison, we generated a random prior model by taking each disordered region and performing a fully randomizing shuffle of the sequence. To facilitate the generation of such a background, an efficient method for performing sequence shuffling is implemented in localCIDER. This process generates a composition and size-matched dataset with identical FCR and NCPR distributions, but where the κ of each sequence has been altered. By constructing a random prior we can examine how κ varies as a function of FCR in the absence of *any* selective pressure for sequence patterning. This is an oversimplification given the fact that many other residues show local sequence compositional preference, but is a simple and consistent approach to generate a conceptually important random prior.

Figure S4A shows a comparison of median FCR vs. median κ across the different organisms. The statistically expected behavior obtained from the composition matched random prior is that κ should be inversely correlated with FCR, as shown by the red dashed line. In naturally occurring sequences we found a strong inverse correlation between κ and FCR with a steeper gradient than would be expected from randomly shuffled sequences; the gradient for the random prior (red dashed line) is –0.24, while the gradient for the naturally occurring sequences (black solid line) is –0.54.

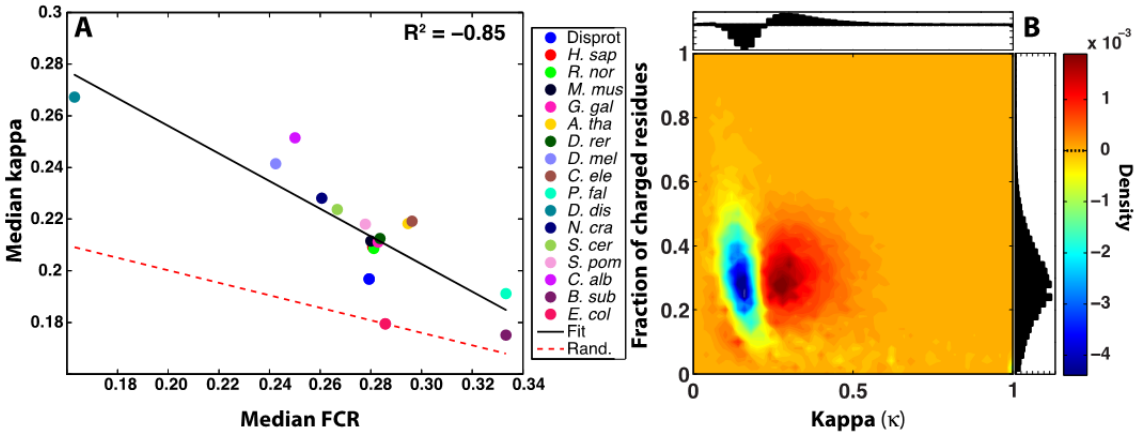

Figure S4: Panel A examines the relationship between median FCR and κ across the different organisms. Panel B is a 2D histogram difference map, and shows regions that are enriched (hotter colors) or depleted (cooler colors) in naturally occurring sequences with respect to a randomly scrambled composition matched background set of sequences. We found that naturally occurring sequences are enriched for sequences with higher κ values, suggesting the evolutionary selection for charge segregation.

In further analysis, we examined the 2D probability distribution of FCR and κ for all species relative to the same random background (Figure S4B). We found a significant over-representation of intermediate-κ and intermediate-FCR disordered regions in naturally occurring sequences (red region). Throughout naturally occurring sequences we found an absence of high κ / high FCR sequences, in line with anecdotal experimental results where highly charged sequences with a high κ are often aggregation prone due to strong electrostatic attraction between oppositely charged patches.

4. FCR vs. κ - further analysis

Figure S4B shows a two-dimensional density difference plot, generated by creating two 2D histograms (Figure S5A, S5B) and subtracting the random distribution from the naturally occurring distribution. As described previously, for this analysis (κ vs. FCR) we focused on polyampholytic sequences. The raw 2D distributions used to generate Figure S3B are included below. In these 2D histograms a bin size of 0.02 is used for κ and for the FCR.

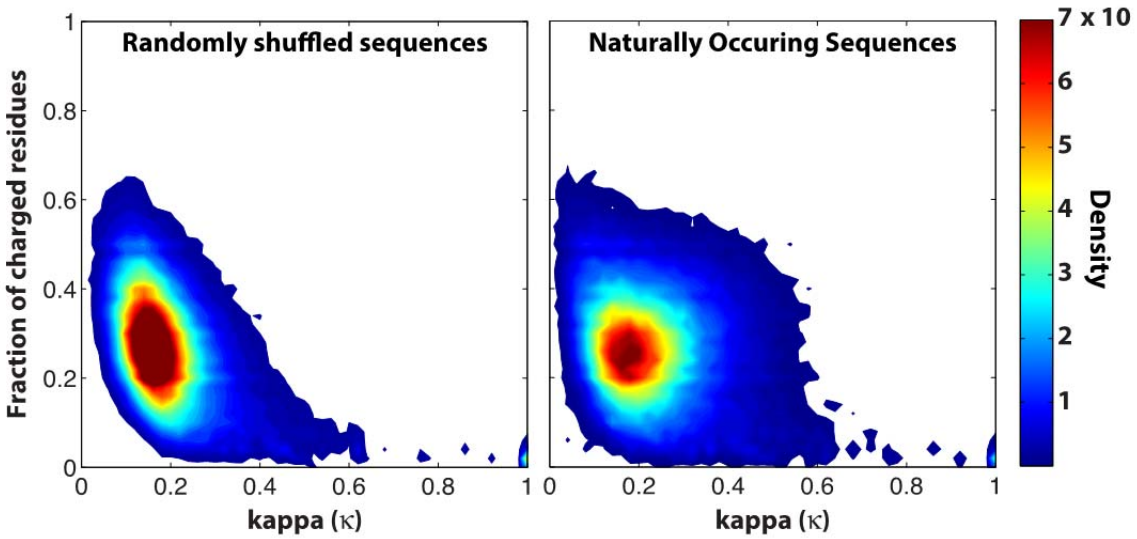

Figure S5: 2D histogram showing the density of sequences associated with a specific κ value and FCR for all polyampholytic disordered regions. Fig S4B is the differences between these two 2D histograms.

Figure S6 shows the distribution of κ values, comparing all naturally occurring sequences with the random-prior sequences. As expected based on the 2D distributions shown in Figure S4 and S5, we find that naturally occurring sequences show a broader distribution of κ values. Notably, substantially more sequences have a higher κ value than would be expected from a random distribution. Again, this result is in line with naturally occurring sequences having more 'charge blocks' – local regions of high net charge density – than one would expect by random chance.

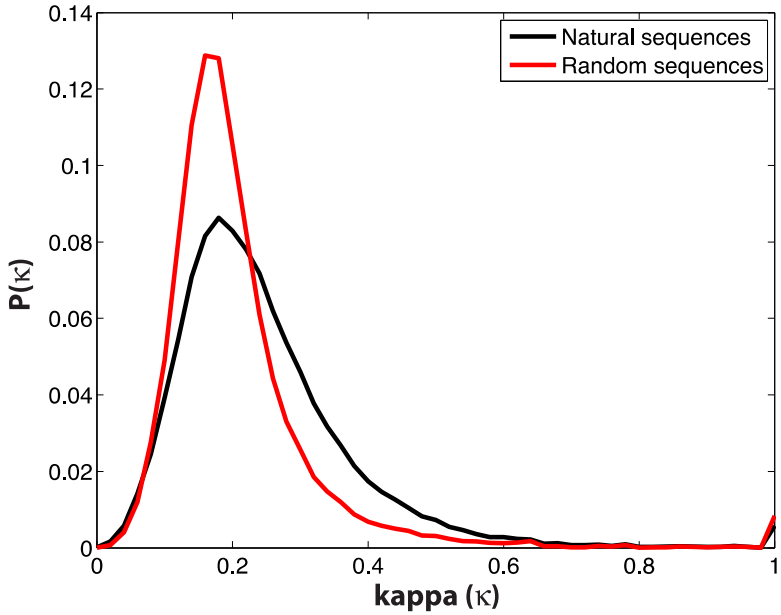

Figure S6: Histogram generated distributions of κ□values from naturally occurring sequences (black) and from the same sequences after unbiased sequence scrambling (red)

The data in Figure S5 can also be represented by determining the median FCR values for a specific κ□range and considering how the median FCR varies with κ. In this analysis, as shown in Figure S7, the complete set of naturally occurring sequences are sub-divided into bins based on κ. For the sequences in each bin, the median FCR value is calculated, and the median FCR vs. κ range is plotted. This analysis does not provide information regarding the number of sequences associated with a given FCR, but allows different organisms to be compared with one another in terms of how FCR and κ co-vary. Only bins with 50 or more sequences are plotted. This analysis reproduces the trends observed in Figure S4A – the median κ□and median FCR are inversely correlated with one another. Again, this implies that there are few sequences where κ and FCR are simultaneously high. We found that sequences with a low κ□value are strongly biased towards a high charge fraction, whereas sequences with a high κ□value are generally depleted in charged residues. Beyond these observations, extracting meaningful proteome-wide conclusions from these data is difficult. While charge distribution described by κ is an important component in determining an IDP's ensemble, there are many other contributing factors that vary on a case-by-case basis. As a result,

these data provide a general, big picture summary of the expected trends, but over-interpretation should be avoided.

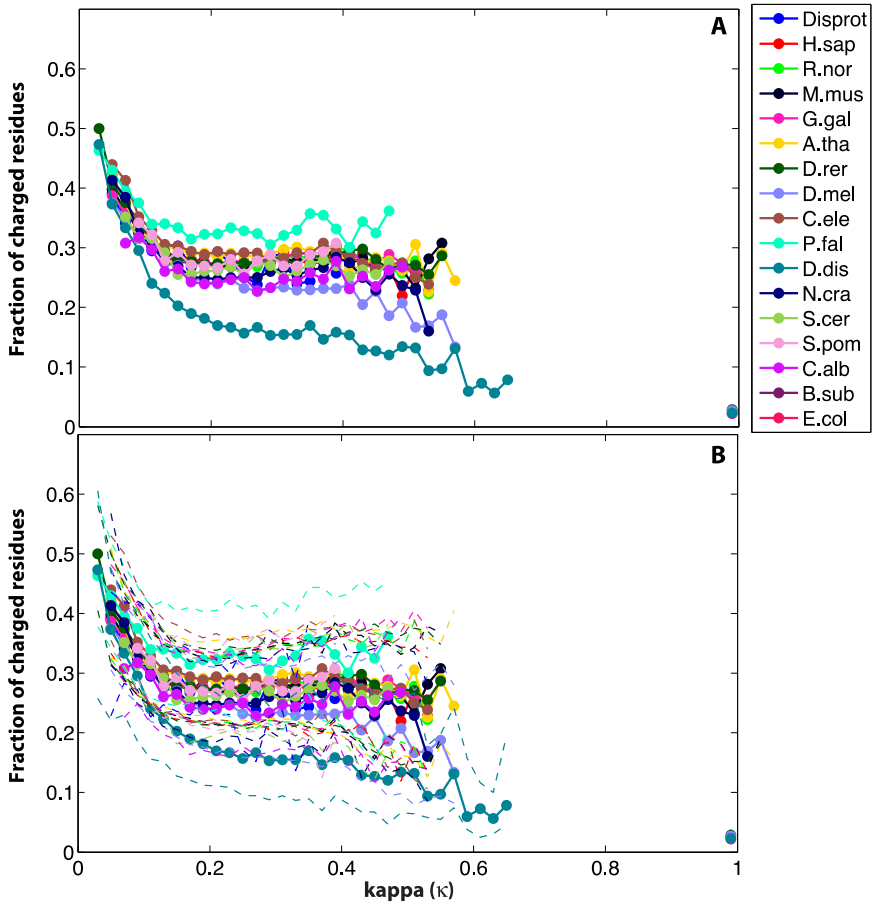

Figure S7: Binned FCR representation of the relationship between κ and FCR. Panel A shows the data without the interquartile range bounds, while Panel B shows the same data with interquartile range bounds.

An important yet nuanced observation from the data presented thus far is that for each organism there exists a range of FCR values where a wide variety of κ values are observed; for many organisms this FCR range lies in the interval of ~0.2 to ~0.4 (figure S7). This overlaps with the R2 region on the diagram-of-states (figure 2, main text), one of the regions (along with R3) where κ has the greatest influence on conformational properties. Finally, R2 is generally the region on the diagram-of-states with the greatest number of disordered regions (Figure S3A). Taken together, these results suggest that a significant fraction of naturally occurring IDPs taken from a wide range of different organisms display sequence properties where charge patterning would be expected to play a major role in determining their conformational ensemble. This is a necessary but not sufficient result to assert that charge patterning is an important feature for proteins

from many different organisms, but sets the stage for a deeper investigation into specific examples through higher resolution analysis.

Given the preceding discussion combined with the inverse correlation between FRC and κ, we offer a plausible interpretation of our results. For sequences with a high FCR there appears to be a strong selective pressure towards well mixed sequences (low κ). At intermediate FCR ($0.2 \leq FCR \leq 0.4$) sequences experience a range of selective pressures both for lower than expected and higher than expected κ values giving rise to a wider distribution than would be expected in the absence of any selective pressure. Finally, at low FCR ($0 \leq FCR \leq 0.2$) charge patterning becomes less influential, and as a result these sequences experience weaker selective pressure. For a true assessment of the conservation associated with sequence patterning, an analysis should consider paralogous IDRs across many different species. While not examined here, clearly localCIDER is well placed to aid in this kind of sequence analysis. It is also worth emphasizing that many other factors (conserved recognition motifs, amino acid composition, residual secondary structure) will all play a role in determining the evolutionary landscape, although charge patterning as measured by κ are one pair of relevant sequence features.

5. FCR vs. NCPR

In the previous section we examined proteome-wide distributions of κ and FCR. Analogously, we can examine how NCPR varies with FCR. Figure S8 shows the 2D histogram - using the same approach as in figure S2 – of FCR vs. NCPR. To maintain bin number parity, the NCPR bin size is 0.04 (ranging from -1.0 to 1.0) while the FCR bin size remains at 0.02 (ranging from 0.0 to 1.0). For these analyses we did not filter out any polyampholyte sequences, but instead used all available sequences.

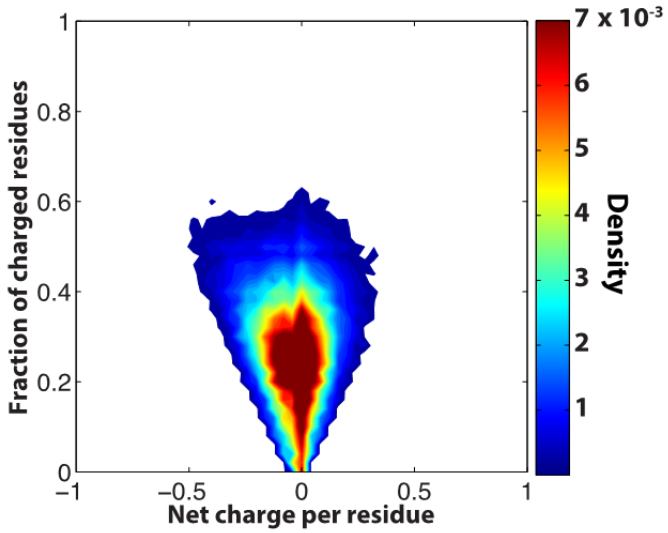

Figure S8: 2D histogram of FCR vs. NCPR. We find the majority of disordered regions are polyampholytes, with an FCR of between ~0.18 and 0.35.

This analysis shows that, generally speaking, there is a depletion of polyelectrolytes across the discorded regions in naturally occurring proteomes, as observed in figure S3A. Figure S9 shows how NCPR is distributed across the different organisms.

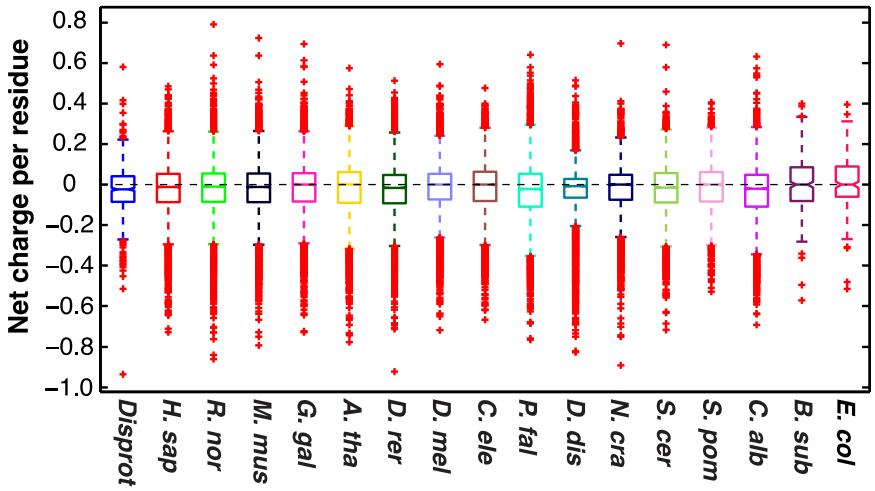

Figure S9: Net charge per residue (NCPR) distribution across organisms. Similar trends are observed over the wide variety of organisms examined.

We can further examine this distribution using the sequence binning approach employed in the FCR vs. κ analysis in figure S7. Figure S10 shows that the same trends with respect to charge are observed across all organisms – as the FCR of regions increases, the NCPR tends towards being increasingly negative (acidic).

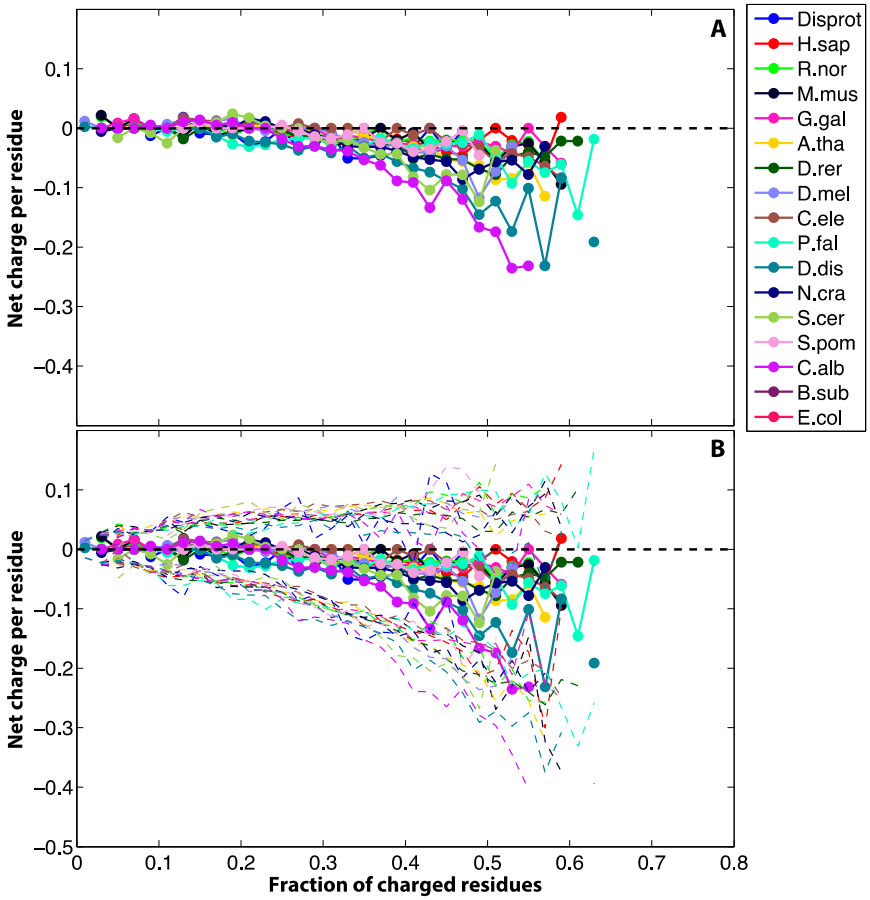

Figure S10: The consistent trends observed suggest that more highly charged disordered regions will have a modest net negative charge. Some organisms (e.g., *C. albicans* or *P. falciparum*) contain disordered regions that are more strongly acidic at a high FCR. Panel A shows the median NCPR when sequences are binned according to their FCR values. Panel B shows the same information as panel A, and includes the interquartile range as dashed lines. The thick black dashed line represents the neutral NCPR value.

An interesting observation that emerges from these data is that sequences with a high fraction of charged residues are most likely to carry a net negative charge. Moreover, there is a near total absence of highly charged sequence with a net positive charge observed across all organisms. An important caveat to consider in all this analysis is that we make no attempt to de-convolve disordered regions into sub-domains, such that all properties examined are the average over each contiguous disordered region. Given the distinct conformational preferences associated with IDPs of different sequence composition, we have no reason to assume that disordered regions could not be divided into sub-domains, where long disordered regions (e.g. > 200 residues) may contain functionally and conformationally discrete subdomains. The identification of such

domains represents a future goal that will be achievable using localCIDER, but is beyond the scope of this work.

6. CIDER webserver details

The CIDER webserver was written in the Python programming language (https://www.python.org/) using the Django web applications framework (https://www.djangoproject.com/). The user interface was built using the Bootstrap front-end framework (http://getbootstrap.com/). CIDER is deployed using an Apache webserver (http://httpd.apache.org/), running on OpenSuse Linux (https://www.opensuse.org/). No user information is stored and no information - other than usage statistics - are saved.

7. localCIDER software package details

localCIDER runs on OSX, Linux and Windows, and requires minimal resource overhead. Plotting is carried out by matplotlib (http://matplotlib.org/) and numerical analysis by numpy (http://www.numpy.org/). localCIDER and its associated documentation are hosted freely on GitHub (https://github.com/), which is also used for version control and feature requests. For more information see http://pappulab.github.io/localCIDER/. A list of the full range of sequence analysis functions can be found at http://pappulab.github.io/localCIDER/

8. Community involvement, open source software, and extended acknowledgments

As an open source project, community involvement is a key part of developing a tool to suit the general needs of our audience. Individuals from institutions around the world have contributed ideas, bug reports, and code fixes. By encouraging this general community involvement, CIDER and localCIDER serve their roles as general-purpose analysis tools. We are grateful for the extensive feedback, thoughts and ideas provided by our colleagues, both directly and in conversation at various meetings over the years. Scientific software should be held to the same (or higher) standard than commercial software, and by allowing our projects to be open source and built with user interaction in mind we can ensure that usability is not compromised by the convenience of rapid development.

Notably, we wish to thank Paul Nobrega, Davide Mercadante, Katra Kolšek, Alex Chin, Thomas Pranzatelli, Gül Zerze, Max Staller, Andrea Soranno, Carlos Hernández and Luke Wheeler for code contributions, bug reports, feedback, suggestions and helpful discussion regarding the design and implementation of CIDER and localCIDER.
We also thank GitHub for providing free educational access to their micro plan. For more information, please see https://education.github.com.

9. Understanding κ

For completeness, we provide a detailed discussion on the statistical and practical properties of the patterning parameter κ (kappa). This section is divided into several subsections. In **section 9.1** we revisit how κ is defined, providing an intuitive overview combined with a mathematical definition. In **section 9.2** we provide a method to formally calculate the number of charge permutants, and describe the probability mass function (PMF) associated with κ for a given sequence composition. **Section 9.3** describes how the expected κ value varies across the diagram-of-states with complete enumeration of expected values for all possible compositions over a range of different sequence lengths. Finally, in **section 9.4** we offer some general rules of thumb when thinking about how a sequence' κ value may influence conformation. **Section 9.4** also offers some notes of caution regarding how one should or should not treat the parameter.

9.1 - Defining κ

The ideas presented in this subsection were first described in previous work (4). We include them here for completeness. The parameter κ is a measure of the mixing of oppositely charged residues along the primary sequence of a protein, where this mixing is effectively quantifying how similar the local charge distribution is when compare to the global charge distribution. Specifically, the local charge distribution is assessed based on five and six residue sub-fragments (blobs). For a sequence where charged residues are globally well mixed with respect to one another, local sequence properties and global properties will mirror one another. For a sequence where charged residues are highly segregated, local properties will be consistently divergent from the global properties. κ is a parameter that formally describes this similarity/difference, and is normalized against a maximally segregated sequence to ensure $0 < \kappa \leq 1$. A graphical summary of how κ maps to protein sequences is shown in Figure S11.

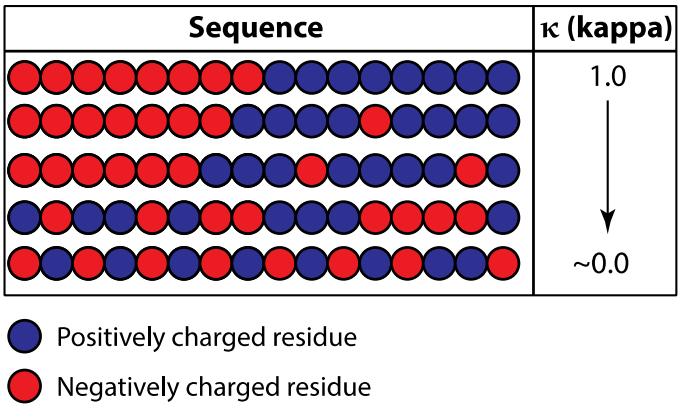

Figure S11: Graphical description of how **κ** varies with sequence patterning. A high κ value is associated with a highly segregated sequence, while a low κ value is associated with a well-mixed sequence. It is worth noting that were are using highly charged polyampholytes here because they graphically illustrate the relationship between κ and patterning well, but as a relevant parameter, κ also applies to much less charged naturally occurring sequences.

To compare local vs. global properties, it is necessary to define a comparison metric. Such a metric should be normalized by sequence length to allow the comparison of regions of different lengths (e.g., a local six residue blob compared with the full *n* residue sequence). For κ, the metric used is **charge asymmetry** (σ), which is defined as follows

$$\sigma = \frac{(f_+ - f_-)^2}{(f_+ + f_-)} \qquad (1)$$

Here, f_+ and f_- represent the fraction of positively and negatively charged residues. To carry out a complete comparison of the global sequence properties with the local sequence properties, we perform a comparison of all possible blobs with the full sequence, normalized by the number of blobs. Specifically, each blob is *g* residues long, meaning a sequence of *n* residues is subdivided into ($n - g + 1 = N_{\text{blobs}}$) blobs. For a complete comparison of global vs. local properties we introduce a new parameter, (δ) which defines a permutant-specific comparison between global and local charge asymmetry, and is defined by:

$$\delta = \frac{\sum_{i=1}^{N_{\text{blobs}}} (\sigma_i - \sigma)^2}{N_{\text{blobs}}} \qquad (2)$$

Here, σ defines the charge asymmetry for the full sequence while σ_i defines the charge asymmetry associated with the *i*-th blob. A graphical schematic of how the summation terms in δ are calculated is shown in Figure S12 (where *g* = 6);

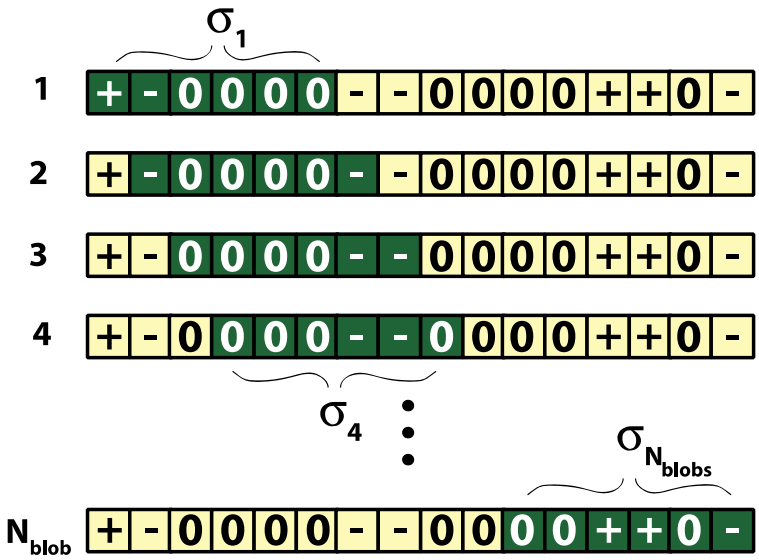

Figure S12: Graphical schematic showing how the summation term in the δ calculation represents a sliding window for determining the σ for each overlapping blob of *g* residues (where in this case *g* = 6).

Having calculated δ for the sequence of interest, we introduce a normalization factor to ensure that we have a parameter (κ) that ranges from 0 to 1. This normalization factor ($\delta_{\max}$) represents the δ associated with the maximally segregated sequence, such that we define κ as shown in equation (3).

$$\kappa = \left(\frac{\delta}{\delta_{\max}} \right) \qquad (3)$$

Finally, for the full definition of κ we need to define the blob size (*g*) – *i.e.,* what is the length-scale that we consider 'local'. The value selected for *g* was chosen to reflect the number of residues that give rise to a chain length at which the interplay between chain-chain, chain-solvent, and solvent-solvent interactions are on the order of kT (13). For protein sequences with low proline contents (*i.e.,* less than 15%) this value is 5 to 6

residues. To account for this variability, we use an average of the κ value derived from g = 5 and g = 6 such that the κ value reported is an average of two κ values - one where g = 5 and one where g = 6. As a result, the κ value reported in the original paper and by localCIDER and CIDER is defined by equation (4)

$$\kappa = \frac{\left(\frac{\delta^{g=5}}{\delta^{g=5}_{\max}}\right) + \left(\frac{\delta^{g=6}}{\delta^{g=6}_{\max}}\right)}{2} \qquad (4)$$

Where $\delta^{g=i}$ reflects the δ value calculated for a sequence with a blob size of i.

9.2 - Number and distribution of κ values

Having defined how κ is calculated, it is useful to provide a general sense of the range of κ values that are likely, given a sequence composition. The most intuitive approach to answer this question would be to define a probability mass function (PMF) associated with κ for a given sequence composition. This would provide a statistical description of the likelihood associated with a given κ value, and help offer statistical context for the κ value associated with a naturally occurring sequence – *i.e.,* is it far from or close to the statistically expected value. In the following subsection we examine how κ values are distributed, and how this distributions changes with FCR and NCPR. An important point to reiterate is that many different sequence permutants will have the same value of κ, a consequence of the fact that κ is a scalar parameter trying to capture sequence-encoded patterning.

One approach for generating the κ PMF would be to perform exhaustive enumeration and determine the complete mapping of every possible charge permutant to κ value followed by the creation of a histogram of those κ values. The number of possible charge permutations of a sequence can be calculated by taking the sequence, converting it into a three-letter alphabet representation (negative, neutral, positive) and using the expression defined in equation (5).

$$\begin{aligned}\text{Number of permutations} &= \binom{(n_0 + n_+ + n_-)}{n_0}\binom{(n_+ + n_-)}{n_+}\binom{n_-}{n_-} \qquad (5)\\ &= A \times B \times C\end{aligned}$$

In this expression, n_0, n_+, and n_- represent the number of neutral, positive, and negative residues, respectively, and the notation here shows the product of three "*a choose k*" terms (termed A, B, C for convenient discussion below). For an example of the conversion of an amino acid sequence into a three-letter alphabet representation, see the twenty residue example in Figure S13.

Figure S13: Example of converting a twenty residue peptide from to a three-letter alphabet. This peptide's sequence-properties are $n_0 = 9$, $n_+ = 5$, and $n_- = 6$, giving it an FCR of 0.55 and an NCPR of –0.05.

Equation (5) can be explained by considering the following framework. There are a total of $(n_0 + n_+ + n_-)$ positions in the sequence. n_0 of those positions can be filled by neutral residues in *A* ways. This leaves $(n_+ + n_-)$ positions, which can be filled with positive residues in *B* ways. Finally, there is only one permutation of ways the negative residues can fill, hence $C = 1$. For a 50-residue sequence with 7 positive and 7 negative residues (FCR 0.28, NCPR=0.0) there are approximately 3.2×10^{15} different permutations. For a 100-residue sequence with 20 positive residues and 20 negative residues, there are approximately 1.9×10^{39} different permutations. Based on these numbers it should be clear that complete enumeration of unique sequences and calculation of the associated PMF is not a feasible strategy. However, given the multinomial nature of the number of permutations, the distribution of κ values can be fit to a lognormal distribution. To test this hypothesis, we took all the disordered regions from the human proteome and generated 500 random permutants per region (*i.e.,* 500 × 23,437 = 11,718,500 sequences). For each random permutant we calculated the κ value. Consequently, for each of 23,437 IDRs we have a distribution of κ values generated through random shuffling. For each region, the distribution of κ values was then fit to a lognormal probability distribution, and the goodness of that fit assessed based on the Euclidean distance between empirical histogram and the lognormal fit. A schematic of this process is illustrated in Figure S14.

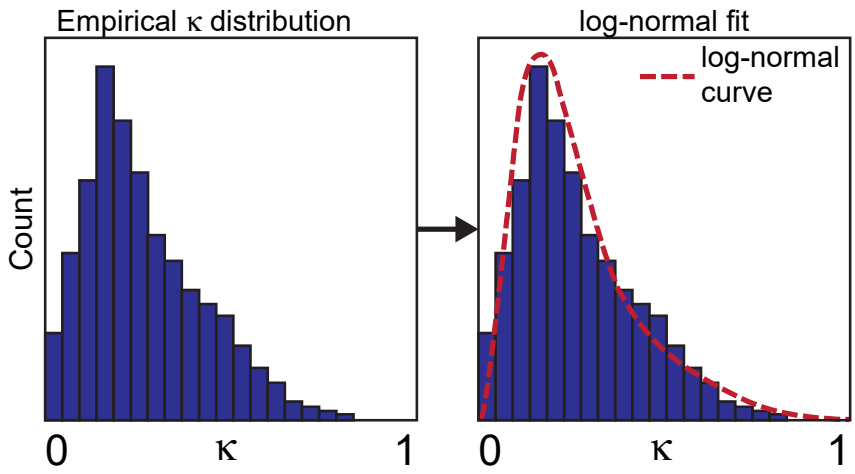

Figure S14: The panel on the left is an empirical histogram of κ values generated by random shuffling of a single IDR. The panel on the right shows a lognormal fit to that histogram (red dashed curve). The goodness of this fit is evaluated by determining the Euclidian distance between the empirical distribution and the lognormal distribution.

With the exception of sequences with a very low fraction of charged residues, we found that the lognormal distribution offers an extremely good fit to all possible regions. Figure S15 shows the goodness of fit plotted in the diagram-of-states plot space (higher numbers indicate a greater deviation from the lognormal curve – *i.e.*, lower is better).

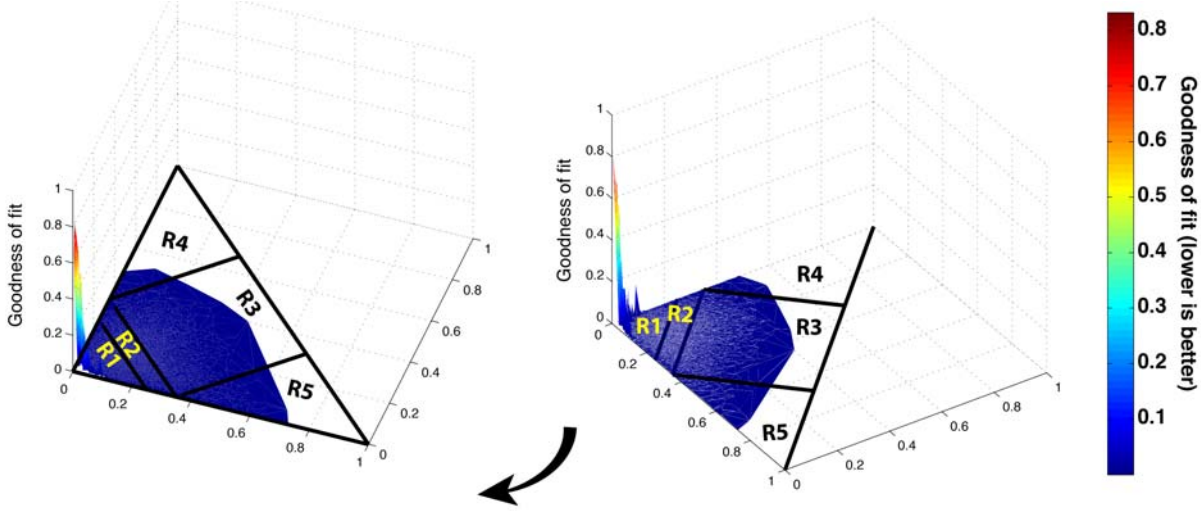

Figure S15: The goodness of fit of the distribution of possible κ values to a lognormal function is shown for all disordered regions in the human proteome shown as a 3D density plot superimposed on the diagram-of-states. The only regions where the fit does poorly is where FCR < 0.1 – *i.e.*, where κ stops being a useful parameter.

To better demonstrate the versatility of the lognormal fit, we randomly selected thirty examples of disordered regions, with six that that were of length 50, 75, 100, 200, and 300 residues. The empirical histogram vs. the lognormal fit is plotted in Figure S15. The goodness of fit across a wide range of lengths quantifies the robustness of the lognormal distribution as a reasonable approximation for the true distribution. Given the fact that the distribution of κ values can be approximated by a lognormal function it is now possible to determine the true random 'likelihood' of realizing the κ value of a naturally occurring sequence, i.e., we can ask "*What is the probability that a sequence with a specific composition will have the observed κ value by random chance?*" This analysis could help identify sequence where the κ value is far from the expected value, implying evolutionary pressure towards a specific κ value.

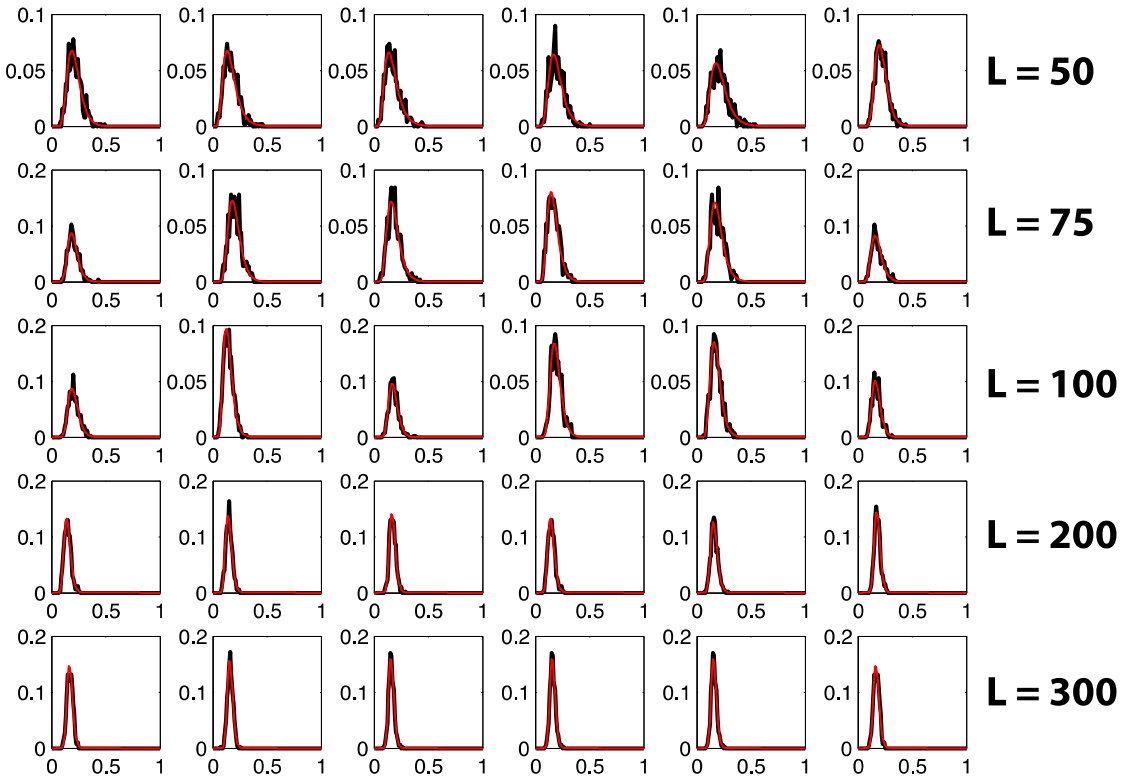

Figure S16: Randomly selected disordered regions and their empirical distribution of κ values (generated by determining the κ value of 500 random permutations of the sequence) compared with a fit lognormal distribution. In each subplot the abscissa (x-axis) is the κ value and the ordinate (y-axis) is the probability of that κ value. Each row contains six randomly selected sequences of length L=X (as defined in the far right hand side of the row). Red curves describe the lognormal fit, while black curves are the empirical histograms.

Using this approach to assess the P(κ) of a real sequence would first involve creating an empirical distribution of possible κ values through repeated random shuffling. Once a distribution of κ values has been generated, a lognormal fit can be performed, and the probability of the κ value of interest determined from that functional form. Based on initial work it appears only 50-100 random permutants are required to build a basis set, from which the lognormal distribution can be generated. This analysis, when performed on the disordered regions in the human proteome, shows that the likelihood of observing IDP sequences with their naturally occurring κ values by random chance is essentially zero. This suggests strong evolutionary pressure away from the statistically expected random prior distribution of charged residues. It is important to remember that the expected value here refers only to the expected value given a uniform background prior. There are many additional constraints which influence how a set of amino acids are distributed in a linear sequence, but as a zeroth order approximation this provides some statistical context of the observed κ value for a given sequence.

9.3. Most likely κ value across diagram-of-state space

Considering the results of section 9.2, for a sequence of some given length and composition we can calculate the statistically expected κ value. If this is done for all sequence compositions of a given length, we can fully explore the κ-to-composition space. Figure S16 shows a 2D heat map of four different sequence lengths (40, 60, 80, 100) where we calculated the κ values for all possible sequence compositions . The color in this heat map reports on expected κ value. A number of features emerge from Figure S17. Firstly, for the vast majority of sequence compositions the expected κ value is between 0.17 and 0.23. This result is relatively insensitive to sequence length. Secondly, in the cases of very strong polyelectrolytes (*i.e.,* FCR >0.5) the expected κ value increases to 0.3 - 0.4. Finally, although not shown here or in Fig S15, when NCPR > 0.9 (*i.e.,* the top left and bottom right corners of the diagram-of-states) the lognormal fitting procedure breaks down in much the same way as it does when FCR < 0.05. This inability to obtain a good fit is a result of one specific class of the residues (positive, negative, or neutral) entirely dominating the sequence composition and causing a rapid drop in the total number of possible sequence permutants.

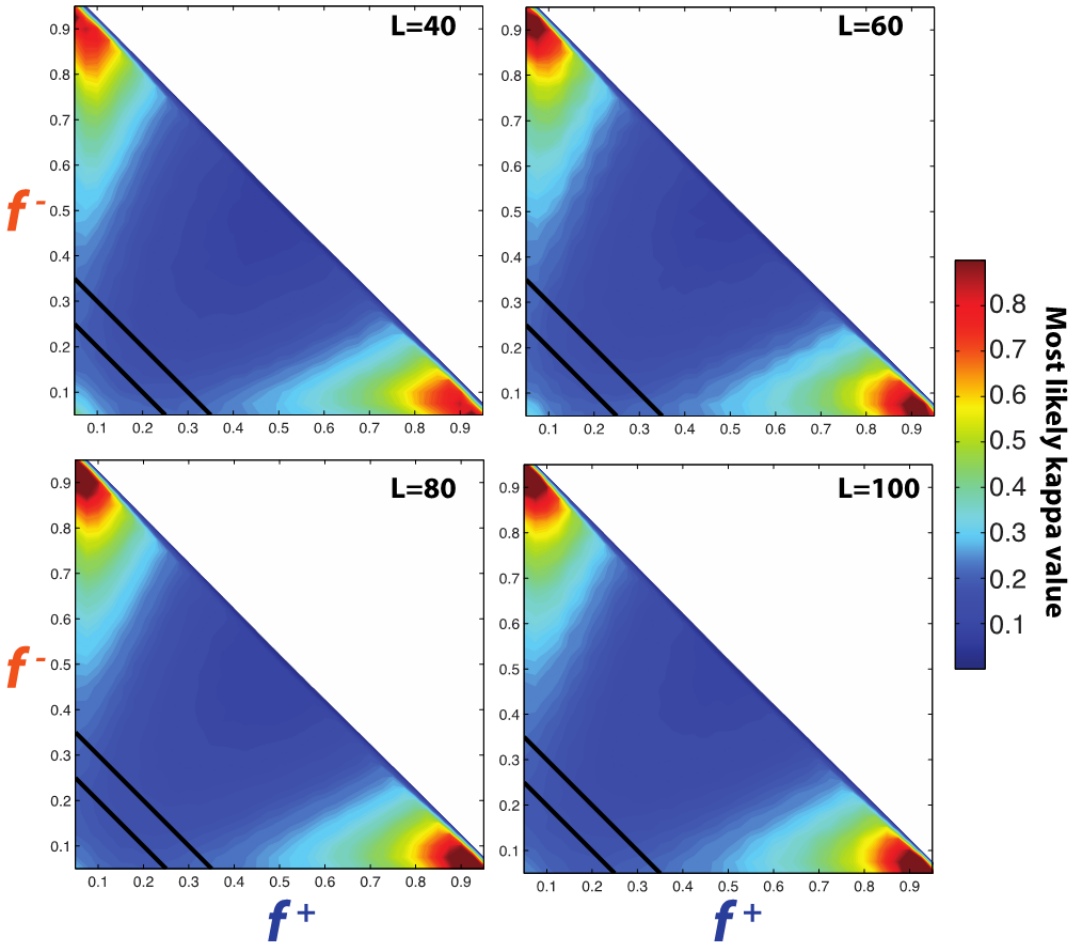

Figure S17: Statistically expected κ value given charge composition based on an unbiased uniform distribution of sequences. Expected values obtained by fitting a lognormal distribution to each sequence.

9.4. Practical comments regarding κ

In this final section, we distill the quantitative and formal descriptions derived in sections 9.1 through 9.3 into some easy to digest expectations regarding κ. For ease of reference we present this final section as a Q & A style discussion. A number of these questions and their answers are repeated on the CIDER webserver help page.

Is κ always a useful / relevant parameter?

No. When sequences have few charged residues (FCR < 0.15) other types of interactions act as the dominant determinants of the underlying ensemble. As an example, proline-rich IDPs are typically more expanded than a charge-composition matched IDP that was depleted for proline residues, a result quantitatively captured by Marsh & Forman-Kay (10). Similarly, local clusters of hydrophobic residues may lead to compaction for an IDP with a low FCR, where that compaction is driven by the hydrophobic effect rather than by charge interaction.

For highly charged disordered sequences we have a number of anecdotal examples where similar global conformational preferences are observed for sequences spanning a range of κ values. These results imply that for some sequences, global conformational properties show robustness to charge distribution, and composition is more important than patterning. In other work, we have observed that even small perturbations to local charge patterning can lead to significant changes in conformation and / or function (14). Fundamentally, κ is readily calculable for every sequence, however, this in no way means that it necessarily is the only determinant of conformational or functional attributes for the sequence in question.

Does κ predict / imply disorder?

No. The κ value of a sequence is in no way related to its propensity for disorder. If it is known that the sequence in question is disordered, then the κ value provides insights regarding the conformational class that is most likely.

When should a sequence be classified as having a high or low κ value?

Anecdotally, for most sequences with an FCR > 0.2 a κ of greater than 0.25 would be considered ‘high’ and lower than 0.12 would be considered ‘low’. This general intuition has come from the analysis of many different sequences over the past few years. The

statistical analysis in section 10.3 suggests that these rough guidelines are grounded in some reality, given the expected κ values.

How different should κ values be to suggest a difference in conformational behavior as dictated by charge patterning?

There is no definitive answer to this question since the relative influence of charge patterning depends on a number of factors. Differences in κ of 0.01 to 0.03 would (most likely) not imply that global charge patterning is likely to be significantly different. However, two sequences with very similar κ values could have quite different relative distributions of charged residues such that the protein's interaction with partners may be very different. As mentioned previously, many sequences may have the same κ value, and it remains to be seen how changing the position of charge clusters while holding κ fixed influences conformation and function.

Is κ the only parameter necessary for considering charge patterning?

Almost certainly not - κ is convenient as it offers a single global description of charge patterning, condensing a high dimensional attribute into a one-dimensional parameter. However, in many cases one might expect this to be insufficient to explain the role of charge patterning. As an example, the number density of sequences associated with a given value of κ is proportional to 10^n where, depending on sequence length, *n* is > 30. As a result, there are an extremely large number of sequence solutions that are consistent with a specific κ value, and other factors may play a key role in influencing conformation and function. These factors may be related to charge (*e.g.*, relative positioning of charged patches), may be determined by patterning of other residue types, or may depend on the intrinsic propensity of a local region for a given secondary structure (*e.g.,* helicity, turns, PPII).

Final comments

The calculation of κ provides simple expectations for the conformational properties of IDPs. It offers a convenient descriptor of global charge distribution that has been shown to offer predictive power in terms of both ensemble behavior and protein function *in vivo* and *in vitro*. When used in conjunction with higher-dimensional parameters (*e.g.,* FCR and linear NCPR) it offers a way to generate expectations regarding conformational behavior of charged IDPs. We anticipate that as our understanding of sequence-to-

ensemble relationships grows, additional parameters may be identified that will offer useful insights.

Supporting Information References

1. UniProt, C. 2015. UniProt: a hub for protein information. Nucleic Acids Res 43:D204-212.
2. Sickmeier, M., J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic, and A. K. Dunker. 2007. DisProt: the Database of Disordered Proteins. Nucleic Acids Res 35:D786-793.
3. Potenza, E., T. Di Domenico, I. Walsh, and S. C. Tosatto. 2015. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. Nucleic Acids Res 43:D315-320.
4. Das, R. K., and R. V. Pappu. 2013. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc Natl Acad Sci U S A 110:13392-13397.
5. Dosztanyi, Z., V. Csizmok, P. Tompa, and I. Simon. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21:3433-3434.
6. Day, C. L., C. Smits, F. C. Fan, E. F. Lee, W. D. Fairlie, and M. G. Hinds. 2008. Structure of the BH3 domains from the p53-inducible BH3-only proteins Noxa and Puma in complex with Mcl-1. J Mol Biol 380:958-971.
7. Rogers, J. M., A. Steward, and J. Clarke. 2013. Folding and binding of an intrinsically disordered protein: fast, but not 'diffusion-limited'. J Am Chem Soc 135:1415-1422.
8. Harmon, T. S., M. D. Crabtree, S. L. Shammas, A. E. Posey, J. Clarke, and R. V. Pappu. 2016. GADIS: Algorithm for designing sequences to achieve target secondary structure profiles of intrinsically disordered proteins. Protein Engineering Design and Selection 29:339-346.
9. Martin, E. W., A. S. Holehouse, C. R. Grace, A. Hughes, R. V. Pappu, and T. Mittag. 2016. Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation. Journal of the American Chemical Society. DOI: 10.1021/jacs.6b10272
10. Marsh, J. A., and J. D. Forman-Kay. 2010. Sequence determinants of compaction in intrinsically disordered proteins. Biophys J 98:2383-2390.
11. Tomasso, M. E., M. J. Tarver, D. Devarajan, and S. T. Whitten. 2016. Hydrodynamic Radii of Intrinsically Disordered Proteins Determined from Experimental Polyproline II Propensities. PLoS Comput Biol 12:e1004686.
12. Burra, P. V., L. Kalmar, and P. Tompa. 2010. Reduction in structural disorder and functional complexity in the thermal adaptation of prokaryotes. PLoS One 5:e12069.
13. Dobrynin, A. V., and M. Rubinstein. 1995. Flory theory of a polyampholyte chain. Journal de Physique II 5:677-695.
14. Das, R. K., Y. Huang, A. H. Phillips, R. W. Kriwacki, and R. V. Pappu. 2016. Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. Proc Natl Acad Sci U S A 113:5616-5621.