Negatively Charged Disordered Regions are Prevalent and Functionally Important Across Proteomes

Lavi S. Bigman [1], Junji Iwahara [2] and Yaakov Levy [1]*

*1 - **Department of Chemical and Structural Biology,** Weizmann Institute of Science, Rehovot, Israel*
*2 - **Department of Biochemistry and Molecular Biology,** Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, TX 77555, United States*

***Correspondence to Yaakov Levy:** koby.levy@weizmann.ac.il (Y. Levy)*
https://doi.org/10.1016/j.jmb.2022.167660
Edited by Edward Lemke

Abstract

Intrinsically disordered regions (IDRs) of proteins are often characterized by a high fraction of charged residues, but differ in their overall net charge and in the organization of the charged residues. The function-encoding information stored via IDR charge composition and organization remains elusive. Here, we aim to decipher the sequence–function relationship in IDRs by presenting a comprehensive bioinformatic analysis of the charge properties of IDRs in the human, mouse, and yeast proteomes. About 50% of the proteins comprise at least a single IDR, which is either positively or negatively charged. Highly negatively charged IDRs are longer and possess greater net charge per residue compared with highly positively charged IDRs. A striking difference between positively and negatively charged IDRs is the characteristics of the repeated units, specifically, of consecutive Lys or Arg residues (K/R repeats) and Asp or Glu (D/E repeats) residues. D/E repeats are found to be about five times longer than K/R repeats, with the longest found containing 49 residues. Long stretches of consecutive D and E are found to be more prevalent in nucleic acid-related proteins. They are less common in prokaryotes, and in eukaryotes their abundance increases with genome size. The functional role of D/E repeats and the profound differences between them and K/R repeats are discussed.

Intrinsically Disordered Regions (IDRs) of proteins, which are linked to various biological functions,[1–5] are often characterized by a high content of charged amino acids. The high charge content of such IDRs, compared with that of foldable sequences, favors interactions with the solvent and may disfavor their folding into a unique three-dimensional structure.[6–9] The structural and dynamic properties of the IDR depend on its charge composition, which is often measured as the net charge per residue (NCPR) or as the fraction of positively or negatively charged residues (f_+ or f_-, respectively). In addition to the charge content, the organization of the charges along the sequence (measured, for example, by the charge pattern parameter, κ[10] can impact the conformational ensemble of IDRs,[11] and modulate the interaction of IDR-bearing proteins with DNA[12,13] and other proteins.[14–21] A key question remains: how is the function of IDRs encoded in their sequence and, specifically, how does the organization of negatively and positively charged residues encode function?

Here, to understand better the principles underlying the effect of charge organization on function, we quantify the properties of IDRs from three different proteomes (human, mouse, and yeast) in terms of the number of IDRs in the proteome, their length, charge content, and charge organization. To explore the linkage between charge characteristics and function, the

proteins were classified according to their function (based on Gene Ontology (GO) molecular function annotations) to three classes: nucleic acid related proteins (NARs), enzymes (whose function is not related to nucleic acids), and other (the rest of the proteins).

The human proteome contains 22,813 IDRs (defined with fIDPnn[22] as regions with at least 15 consecutive residues and with score > 0.3) of which 6,273 are NARs, 3,450 are enzymes, and 9,072 are in the control group (the remaining proteins have no annotated function). As shown in Figure 1(A), ~60% of human enzymes (triangles), 40% of control (empty circles) and 40% of NAR proteins (filled circles) have no IDRs. By contrast, the presence of 2–4 IDRs is most common in NARs, which serve as disordered tails or flexible linkers bridging two folded domains. Similar results were obtained for the yeast and mouse proteomes (Figure S1).

To quantify further the differences between the proteins in the NAR and enzyme classes, we analyzed cases with a single (Figure 1(B)) or two IDRs (Figure 1(C)). Figure 1(B) shows that NARs, enzymes and proteins in the control group possess a larger fraction of negatively charged (NCPR < 0, red bars) than positively charged IDRs (NCPR > 0, blue bars). Similarly, when only highly charged IDRs (NCPR > |0.3|) are considered, proteins from all functional groups have more negatively than positively charged IDRs (inset). For proteins with two terminal IDRs, 18–40% of the proteins have two negatively charged tails (red bars),18–37% have two positively charged tails (blue bars), and 44–62% of the proteins have one positively charged and one negatively charged tail (purple bar) in the three functional classes (Figure 1(C)). The prevalence of proteins with two oppositely charged tails may suggest that these proteins share a common function that can take advantage of dual charge character. One possibility is to form

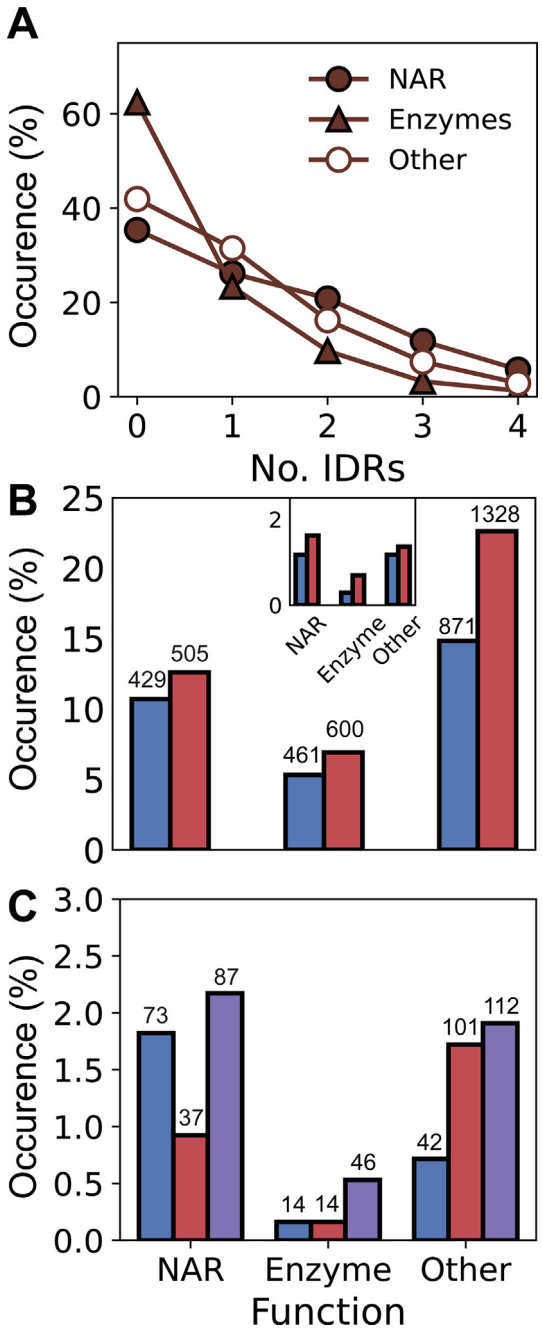


Figure 1. Occurrence and charge properties of IDRs in three functional protein classes. (A) Proteins in the human proteome (15,875 proteins with annotated function) were divided into three functional classes: nucleic acid related proteins (NAR; 3,838 proteins), enzymes (whose function is not dependent on nucleic acids; 4,308 proteins), and a control group containing the rest of the annotated proteins in the human proteome (7,729 proteins). Proteins were classified based on Gene Ontology (GO) molecular function. For each functional class, the fraction of proteins that have 0, 1, 2, 3, or 4 IDRs is shown (see main text for detailed description). IDRs are defined as stretches of at least 15 residues whose disorder fIDPnn score[22] is greater than 0.3. Similar analyses for the yeast and mouse genomes are shown in Figure S1. (B) For proteins with a single IDR, the occurrence of proteins with a negatively (red bars) or positively charged IDR (blue bar) is shown. The value on top of each bar shows the number of proteins in that sub-class. Inset: proteins with a single IDR that have net charge per residue of NCPR < −0.3 (red bars) or NCPR > 0.3 (blue bars). (C) For proteins with two terminal IDRs (at both the C- and N-termini), the occurrence of three possible cases are plotted: the two IDRs are both negatively or both positively charged (red and blue bars, respectively), or one IDR is positively charged and the other is negatively charged (purple bar). The number of proteins in each sub-class is shown on top of each bar. In panels B and C, the percentage occurrence is calculated according to the total number of proteins in each functional class.

biomolecular condensates, but other functions may be involved as well.

The length of an IDR can also affect its function. Figure 2(A) shows a contour plot of IDR length versus NCPR values for all human proteins. The quite symmetric shape of the contour plot suggests that when considering all human IDRs, the average length and NCPR are similar for positively and negatively charged IDRs. Most IDRs are of ~20 residues and with NCPR ~ 0. Some IDRs are much longer (>1000 residues) and some have NCPR values close to unity (*i.e.*, near-polyelectrolytic IDRs), with the negatively charged IDRs being more highly charged than the positively charged IDRs.

To obtain better resolution, we focused on highly charged IDRs (*i.e.*, with NCPR > |0.3|) in the three protein functional classes. Figure 2(B) shows a violin plot of the length of highly negatively (red) or highly positively charged IDRs (blue) for each protein class. It is apparent from the width of the distributions that there are more negatively charged IDRs than positively charged IDRs across protein classes. Moreover, negatively charged IDRs are longer (Figure 2(B)) and have slightly greater NCPR absolute values (Figure 2(C)) than positively charged IDRs. In addition, negatively charged IDRs on NARs are longer than negatively charged enzymatic and control IDRs (Figure 2 (C)). Similar results were obtained for the yeast and mouse proteomes (Figure S2).

The observation that there are clear differences in length and NCPR values between highly negatively and highly positively charged IDRs (Figure 2) led us to ask whether the charge organization within IDR sequences is different for positively and negatively charged IDRs. To that end, we calculated the length of residue repeats involving either positively charged residues (Lys and Arg, denoted as K/R repeats) or negatively charged residues (Asp or Glu, D/E repeats) in each IDR. Figure 3(A) shows contour plots of the length of the longest charged repeat in each IDR versus the fraction of charged residues. The length of the longest K/R (D/E) repeat is plotted versus f_+(f_-) and presented as a blue (red) contour map. Whereas IDRs have D/E repeats with lengths approaching 30 amino acids (red f_- plots), the longest K/R repeat in IDRs only reaches 10 amino acids (blue f_+ plots). Moreover, on average, the IDRs of the NAR proteins have longer D/E repeats than the IDRs of the enzymes or the control group (see the mean values indicated on the plots). A complementary way to quantify the degree of charge segregation is by the κ parameter for each IDR.[10] The closer the

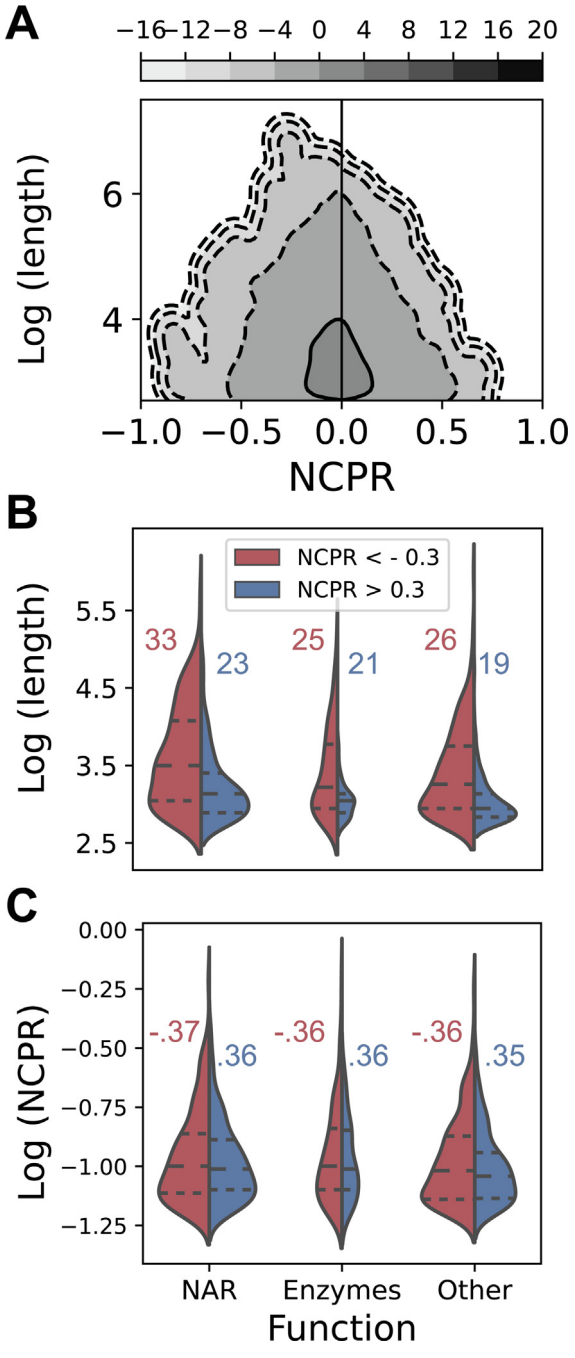


Figure 2. Negatively charged IDRs are longer and more highly charged than positively charged IDRs. (A) Contour plot showing the length (logarithmic scale) versus net charge per residue (NCPR) for all human proteins. (B and C) Violin plots showing the distribution of length (B) and the NCPR (C) in each functional class, for highly negatively (NCPR < −0.3; red) and highly positively charged IDRs (NCPR > 0.3; blue). The median length (B) and NCPR (C) values for each group are shown in the corresponding color. The p-value of the negative and positive IDRs for the NAR, enzyme and control groups are 2×10^{-14}, 4×10^{-9}, and 1×10^{-18}, respectively, with respect to length, and are 0.1, 0.4, and 0.05, respectively, for NCPR. NCPR values are plotted as absolute numbers. Data for yeast and mouse are shown in Figures S2 and S3, respectively.

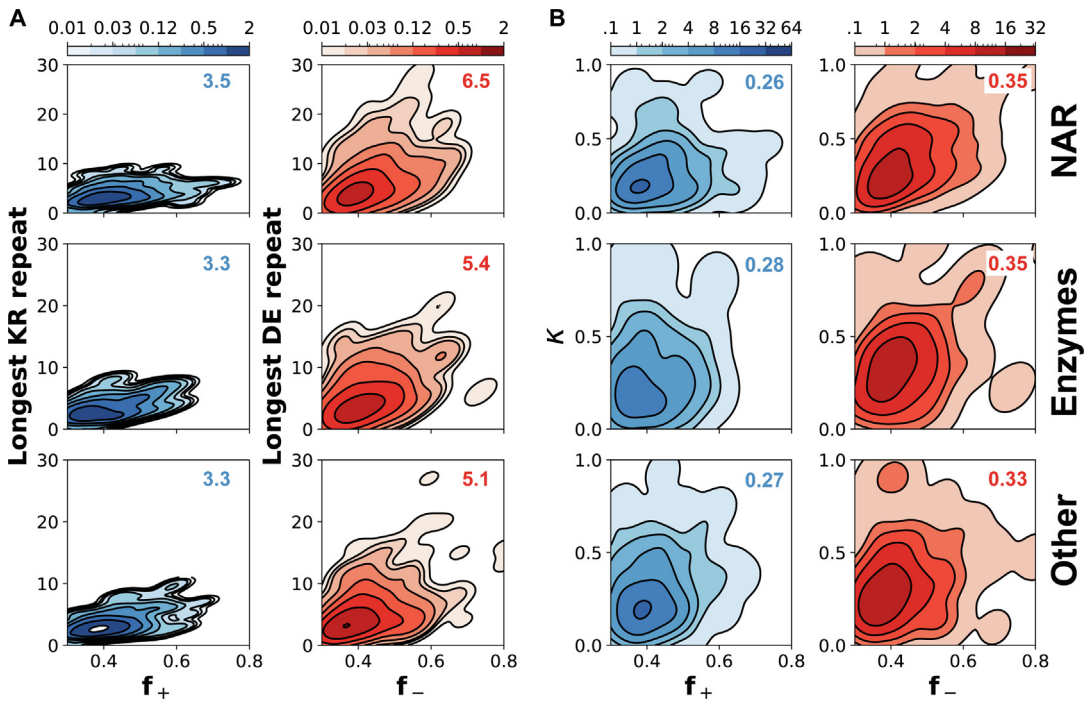
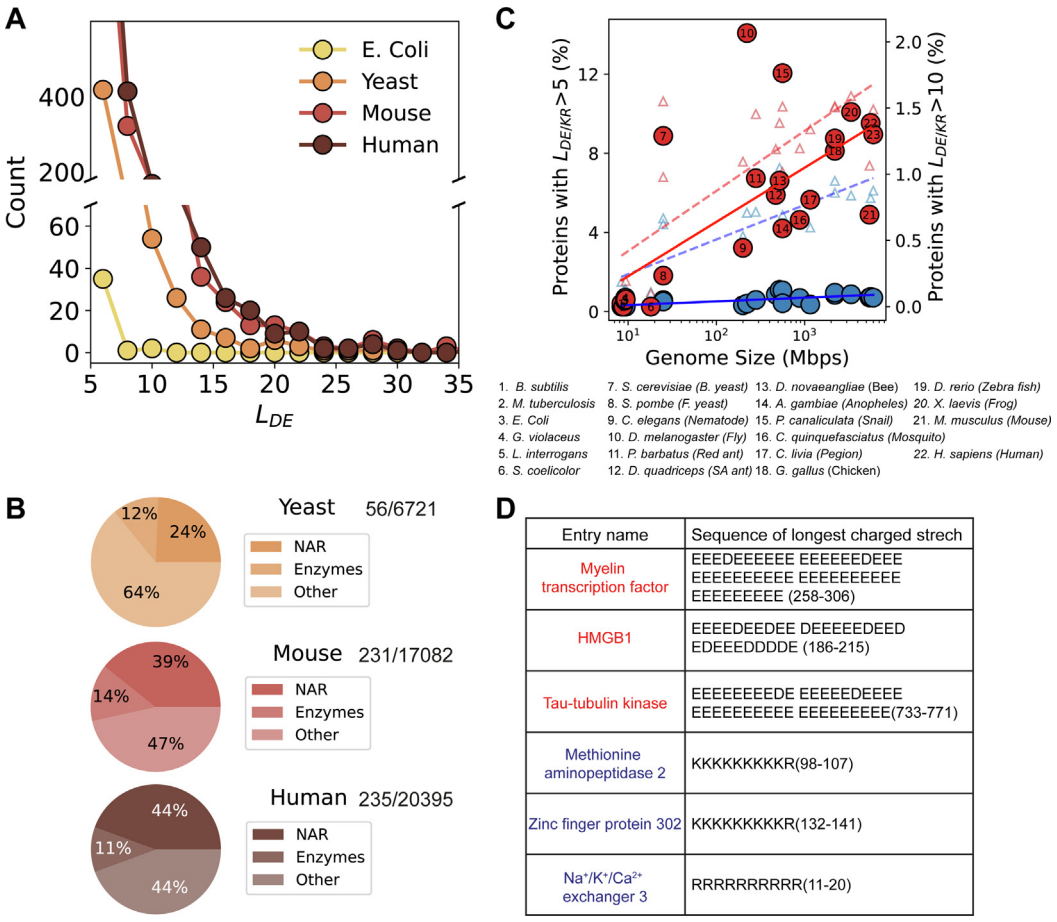


Figure 3. Charge segregation in charged IDRs. (A and B) The length of consecutive positively charge residues (K/R repeats, blue) or negatively charged residues (D/E repeats, red) (A) and the charge organization along the sequence of the longest stretch as measured by κ (B) are shown as a function of the fraction of positively (f_+) or negatively charged (f_-) residues in a given IDR sequence. Mean values for the longest K/R or D/E repeats (in amino acids) and of κ are indicated on each panel. The κ values were calculated using CIDER.[47]

value of κ is to 1, the more segregated the charges are. The results of the κ analysis further support our observation that charges are more segregated in negatively charged IDRs compared with positively charged IDRs (Figure 3(B)). Similar results were obtained for the yeast (Figure S3) and mouse (Figure S4) proteomes. We note that identifying the IDRs using the IUPRED predictor[23] (instead of the fIDPnn predictor[22]) resulted with different number of IDRs per organism and in each functional group, yet the IDRs had similar characteristics, indicating that our conclusions are not sensitive to the disorder predictors.

The observation that highly negatively charged IDRs in NARs are characterized by more significant charge segregation than IDRs in other functional classes may imply that the function of IDRs with long D/E repeats is related to nucleic acids. To further characterize the D/E repeats, we analyzed all D/E repeats in proteomes from multiple organisms by counting the length of the longest D/E repeat in each protein (L_{DE}). For comparison, the length of K/R repeats (L_{KR}) was analyzed too. Figure 4(A) shows the distribution of L_{DE} in the human, mouse, yeast, and *Escherichia coli* (*E. coli*) proteomes. *E. coli* has no proteins with $L_{DE} > 10$, but the three eukaryotic organisms have many proteins with $L_{DE} > 10$. The pie charts in Figure 4(B) show the functional classification of proteins with $L_{DE} > 10$. In all investigated eukaryotic proteomes (yeast, mouse, and human), ~1% of the proteins have $L_{DE} > 10$. Furthermore, while 24% of yeast proteins with $L_{DE} > 10$ are NARs, in the mouse and human proteomes 39 and 44%, respectively, of the proteins with $L_{DE} > 10$ are NARs. Looking at a higher resolution of protein function, we found that within the NAR proteins in the human and mouse proteomes with $L_{DE} > 10$, the occurrence of DNA-binding proteins is at least double than that of RNA-binding proteins (in human: 25% and 13% for DBPs and RBPs respectively, and in mouse 26% and 9% for DBPs and RBPs, respectively). This analysis further supports the idea that, in many cases, the function of long D/E repeats is related to nucleic acid binding.

Based on the latter observation, an additional plausible inference may be that long D/E repeats are more abundant in organisms with more complex genomes. To directly test this idea in a broad context, we estimated the fraction of proteins with K/R or D/E repeats longer than either 5 or 10 residues as a function of genome size for 22 different organisms. Figure 4(C) (left Y axis; triangles) shows that the fraction of proteins with $L_{KR} > 5$ or $L_{DE} > 5$ increases with genome size, but the D/E repeats (red) are about twice as abundant as K/R repeats (slopes of 1.3 ($R^2 = 0.6$) and 0.7 ($R^2 = 0.6$), respectively). The difference between the D/E and K/R repeats is striking when



Entry name	Sequence of longest charged strech
Myelin transcription factor	EEEDEEEEEE EEEEEEDEEE EEEEEEEEEE EEEEEEEEEE EEEEEEEEE (258-306)
HMGB1	EEEEDEEDEE DEEEEEDEED EDEEEDDDDE (186-215)
Tau-tubulin kinase	EEEEEEEEDE EEEEEDEEEE EEEEEEEEEE EEEEEEEEE(733-771)
Methionine aminopeptidase 2	KKKKKKKKKR(98-107)
Zinc finger protein 302	KKKKKKKKKR(132-141)
$Na^+/K^+/Ca^{2+}$ exchanger 3	RRRRRRRRRR(11-20)

Figure 4. Function of D/E repeats across organisms. (A). The number of proteins having D/E repeats of various lengths (L_{DE}) is shown for *E.coli* (yellow), yeast (orange), mouse (red), and human (brown) proteomes. (B) Pie charts showing the distribution of proteins with D/E repeats longer than 10 amino acids (*i.e.*, $L_{DE} > 10$) across three functional classes: NARs, enzymes, and other using the same color code as in (A). To the right of each pie chart, the fraction indicates the number of proteins with $L_{DE} > 10$ out of the total number of proteins in the respective proteome. (C) Percent of proteins in which the length of the longest D/E or K/R repeat is $L_{DE/KR} > 5$ (empty triangles; left axis) or $L_{DE/KR} > 10$ (filled circles; right axis) is shown as a function of genome size for 22 organisms. Red and blue symbols show data for stretches of negatively (L_{DE}) and positively (L_{KR}) charged amino acids, respectively. The numbers on the red circles correspond to the organism that each data point represents, as listed beneath panel C. Symbols 1–6 overlap, and all of them represent bacteria. The solid and dashed lines show the best linear fit to the data of $L_{DE/KR} > 5$ and >10, respectively for D/E (red) and K/R (blue) repeats. (D) Examples of proteins with long D/E (red) or K/R (blue) repeats. The sequences of the stretches are shown with the position of these amino acids in the protein sequence given in brackets. These examples include both NAR proteins (myelin transcription factor, HMGB1, zinc finger protein 302) and proteins with other functions (tau-tubulin kinase, methionine aminopeptidase 2 and $Na^+/K^+/Ca^{2+}$ exchanger 3).

examining longer repeats. Whereas the fraction of proteins with $L_{DE} > 10$ (right Y axis; circles) increases with genome size (slope = 0.1, R^2 = 0.6), the fraction of proteins with $L_{KR} > 10$ barely changes with genome size (slope = 0.01, R^2 = 0.5).

To emphasize the difference between stretches of positively and negatively charged residues, we show that while the D/E repeats in the human proteome can reach up to 49 residues (in the extreme case of the myelin transcription factor), K/R repeats are only up to 10 residues long (see Figure 4(D)), illustrating the profound difference between these two types of charged IDRs. Possible explanations for why negatively charged D/E repeats are longer and more common than positively charged K/R repeats include that K/R repeats are more prone to proteolysis[24] and that they may slow down translation kinetics in the ribosome because the exit tunnel is negatively charged.[25,26]

Based on the observations that long D/E repeats are more likely to be found in NAR than in enzymes and that the fraction of proteins with long D/E repeats increases with genome size, we conclude that the function of long D/E repeats is related, in

many cases and across organisms, to nucleic acids. Several previous studies support the idea that NARs are enriched with IDRs,[27] and specifically that the function of highly negatively charged IDRs is related to nucleic acids. For example, it was shown that a DE-rich region of the transcription factor Sox11 autoinhibits DNA binding both in vivo and in vitro.[28] Similar observations were made for the regulatory factor RFX1, where a 17-amino acid stretch containing 14 negatively charged amino acids inhibits DNA binding,[29] and for the upstream binding factor (UBF), which contains two D/E repeats (L_{DE} = 21 and 18 amino acids).[30] It was also suggested that structural D/E repeats may play a role in gene regulation through nucleic acid mimicry.[31] An extreme example comes from a recent study that combined NMR measurements and molecular-dynamics simulations, which demonstrated that the 30 amino-acid long D/E repeat in the C-terminus of the DNA-binding protein HMGB1 dynamically autoinhibits its interaction with DNA through fuzzy electrostatic interactions.[32] Similarly, another study showed that a 10 amino-acid long sequence containing 9 D/E repeats (and a single non-acid middle residue) in the RNA binding protein Nop15 increases protein stability and RNA binding specificity.[33] Autoinhibition was also shown for DNA-binding proteins with IDRs that do not have D/E repeats, such as in the case of p53.[34] Interestingly, also in the case of p53, autoinhibition is electrostatically driven by a phosphorylation-dependent switch.[35]

We should point out that autoinhibition via D/E repeats also occurs for other types of proteins. For example, phospholipase Cβ3 (PLCβ3) undergoes autoinhibition via D/E repeats (L_{DE} = 11), which interact with the positively charged surface of the PLCβ3 catalytic domain.[36] Neural Wiskott-Aldrich syndrome protein (N-WASP), a regulator of actin cytoskeleton formation whose functional domain (WHI) is positively charged, is also autoinhibited via its acidic region containing D/E repeats (L_{DE} = 11).[37] A probable reason why D/E repeats are found more frequently among NAR proteins is that they possess positively charged domains for DNA/RNA-binding, which D/E repeats can inhibit via electrostatic interactions.

An additional unique function of a highly negatively charged IDR was recently shown for DAXX, a D/E rich protein that contains a 50 amino acid long segment with 35 negatively charged residues. It was shown that this protein prevents aggregation, solubilizes pre-existing aggregates, and unfolds misfolded species of model substrates and neurodegeneration-associated proteins.[38] Finally, we focused here on repeats of negatively charged amino acids and their function, but it is noteworthy that repeats of other amino-acid are found in eukaryotic organisms, and that they have important biological functions.[39,40]

Interestingly, glutamate (E) is more frequently found than aspartic acid in D/E repeats. In the human proteome, the frequency ratio n(E)/n(D) is 3.1 among D/E repeats with L_{DE} > 10, whereas the overall n(E)/n(D) ratio among all proteins is 1.5. In the mouse proteome, the corresponding ratios are 3.1 and 1.4. This strong bias toward E rather than D suggests that charge is not the only factor important for D/E repeats as, compared with poly-E, the shorter side chains of poly-D cause stronger charge repulsion and favor a more expanded backbone structure.[41] Thus, the strong bias toward E in D/E repeat sequences might be related to their structural compactness or to the suitability of their charge density for the functional role of D/E repeats.

In conclusion, in this study, we analyzed the sequence features of negatively charged IDRs associated with different functions across several proteomes. We found that highly negatively charged IDRs are longer and more highly charged than positively charged IDRs. In addition, highly negatively charged IDRs whose function is related to nucleic acids (NARs) are longer, more charged, and their charged residues are more segregated than IDRs in enzymes or in the control group. This finding suggests that, whereas positively charged IDRs are functionally important because they can directly influence protein interactions with nucleic acids via attractive electrostatics (and consequently affect binding thermodynamics[42] and kinetics[12,43,44]), negatively charged IDRs are functionally important because they can regulate nucleic acid binding. Our study clearly shows that negatively charged IDR are also used for other biological functions. One example is the negatively charged N-terminal tails that coat microtubules, whose precise patterning of negatively charged residues shapes the energy landscape for microtubule-mediated protein translocation for a variety of proteins.[45,46]

Moreover, we found that negatively charged IDR are characterized by long D/E repeats whereas the K/R repeats are much shorter. We found that the function of almost 50% of the proteins with L_{DE} > 10 in the mouse and human proteomes is related to nucleic acids, and that the fraction of proteins with long D/E repeats per organism increases with genome size. Therefore, we conclude that highly negatively charged IDRs and, more specifically, IDRs with long D/E repeats, are likely to play an important role in regulating transcription and translation, as these processes are ultimately controlled by the interactions of proteins with nucleic acids.

CRediT authorship contribution statement

Lavi S. Bigman: Conceptualization, Data curation, Visualization, Methodology, Software, Formal analysis. **Junji Iwahara:** Conceptualization. **Yaakov Levy:** Conceptualization, Methodology.

DATA AVAILABILITY

Data will be made available on request.

Acknowledgments

This work was supported by Grant 2020624 (to Y. L.) from the United States – Israel Binational Science Foundation and by Grant MCB-2026805 (to J.I.) from the National Science Foundation.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2022.167660.

Received 28 February 2022;
Accepted 24 May 2022;
Available online 31 May 2022

Keywords:
disordered regions;
repeat sequences;
D/E repeat;
polyampholytes;
electrostatics

Abbreviations:
IDRs, Intrinsically Disordered Regions; NCPR, Net Charge Per Residue; NAR, Nucleic Acid Related; GO, Gene Ontology

References

1. Van Der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., et al., (2014). Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **114**, 6589–6631.
2. Uversky, V.N., (2013). The most important thing is the tail: Multitudinous functionalities of intrinsically disordered protein termini. *FEBS Lett.* **587**, 1891–1901.
3. Oldfield, C.J., Dunker, A.K., (2014). Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu. Rev. Biochem.* **83**, 553–584.
4. Das, R.K., Ruff, K.M., Pappu, R.V., (2015). Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **32**, 102–112.
5. Babu, M.M., (2016). The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.* **44**, 1185–1200.
6. Uversky, V.N., Gillespie, J.R., Fink, A.L., (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins Struct. Funct. Genet.* **41**, 415–427.
7. Müller-Späth, S., Soranno, A., Hirschfeld, V., Hofmann, H., Rüegger, S., Reymond, L., Nettels, D., Schuler, B., (2010). Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14609–14614.
8. Hofmann, H., Soranno, A., Borgia, A., Gast, K., Nettels, D., Schuler, B., (2012). Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci.* **109**, 16155–16160.
9. Bianchi, G., Longhi, S., Grandori, R., Brocca, S., (2020). Relevance of electrostatic charges in compactness, aggregation, and phase separation of intrinsically disordered proteins. *Int. J. Mol. Sci.* **21**, 1–30.
10. Das, R.K., Pappu, R.V., (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.
11. Mao, A.H., Crick, S.L., Vitalis, A., Chicoine, C.L., Pappu, R. V., (2010). Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 8183–8188.
12. Vuzman, D., Levy, Y., (2010). DNA search efficiency is modulated by charge composition and distribution in the intrinsically disordered tail. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21004–21009.
13. Vuzman, D., Levy, Y., (2012). Intrinsically disordered regions as affinity tuners in protein-DNA interactions. *Mol. Biosyst.* **8**, 47–57.
14. Hazra, M.K., Levy, Y., (2020). Charge pattern affects the structure and dynamics of polyampholyte condensates. *Phys. Chem. Chem. Phys.* **22**, 19368–19375.
15. Hazra, M.K., Levy, Y., (2021). Biophysics of phase separation of disordered proteins is governed by balance between short- And long-range interactions. *J. Phys. Chem. B* **125**, 2202–2211.
16. Schuler, B., Borgia, A., Borgia, M.B., Heidarsson, P.O., Holmstrom, E.D., Nettels, D., Sottini, A., (2020). Binding without folding – the biomolecular function of disordered polyelectrolyte complexes. *Curr. Opin. Struct. Biol.* **60**, 66–76.
17. Laptenko, O., Tong, D.R., Manfredi, J., Prives, C., (2016). The Tail That Wags the Dog: How the Disordered C-Terminal Domain Controls the Transcriptional Activities of the p53 Tumor-Suppressor Protein. *Trends Biochem. Sci.* **41**, 1022–1034.
18. Heidarsson, P.O., Mercadante, D., Sottini, A., Nettels, D., Borgia, M.B., Borgia, A., Kilic, S., Fierz, B., et al., (2022). Release of linker histone from the nucleosome driven by polyelectrolyte competition with a disordered protein. *Nature Chem.* **14**, 224–231.

19. Somjee, R., Mitrea, D.M., Kriwacki, R.W., (2020). Exploring relationships between the density of charged tracts within disordered regions and phase separation. *Pacific Symp. Biocomput.* **25**, 207–218.
20. Dinic, J., Marciel, A.B., Tirrell, M.V., (2021). Polyampholyte physics: Liquid–liquid phase separation and biological condensates. *Curr. Opin. Colloid Interface Sci.* **54**, 101457.
21. Das, S., Amin, A.N., Lin, Y.H., Chan, H.S., (2018). Coarse-grained residue-based models of disordered protein condensates: Utility and limitations of simple charge pattern parameters. *Phys. Chem. Chem. Phys.* **20**, 28558–28574.
22. Hu, G., Katuwawala, A., Wang, K., Wu, Z., Ghadermarzi, S., Gao, J., Kurgan, L., (2021). flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nature Commun.* **12**, 1–8.
23. Mészáros, B., Erdös, G., Dosztányi, Z., (2018). IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337.
24. Hosaka, M., Nagahama, M., Kim, W.S., Watanabe, T., Hatsuzawa, K., Ikemizu, J., Murakami, K., Nakayama, K., (1991). Arg-X-Lys/Arg-Arg motif as a signal for precursor cleavage catalyzed by furin within the constitutive secretory pathway. *J. Biol. Chem.* **266**, 12127–12130.
25. Leininger, S.E., Rodriguez, J., Vu, Q.V., Jiang, Y., Li, M.S., Deutsch, C., O'Brien, E.P., (2021). Ribosome Elongation Kinetics of Consecutively Charged Residues Are Coupled to Electrostatic Force. *Biochemistry* **60**, 3223–3235.
26. Lu, J., Deutsch, C., (2008). Electrostatics in the Ribosomal Tunnel Modulate Chain Elongation Rates. *J. Mol. Biol.* **384**, 73–86.
27. Wang, C., Uversky, V.N., Kurgan, L., (2016). Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. *Proteomics* **16**, 1486–1498.
28. Wiebe, M.S., Nowling, T.K., Rizzino, A., (2003). Identification of novel domains within Sox-2 and Sox-11 involved in autoinhibition of DNA binding and partnership specificity. *J. Biol. Chem.* **278**, 17901–17911.
29. Katan-Khaykovich, Y., Shaul, Y., (2001). Nuclear import and DNA-binding activity of RFX1: Evidence for an autoinhibitory mechanism. *Eur. J. Biochem.* **268**, 3108–3116.
30. Ueshima, S., Nagata, K., Okuwaki, M., (2017). Internal Associations of the Acidic Region of Upstream Binding Factor Control Its Nucleolar Localization. *Mol. Cell. Biol.* **37**
31. Chou, C.C., Wang, A.H.J., (2015). Structural D/E-rich repeats play multiple roles especially in gene regulation through DNA/RNA mimicry. *Mol. Biosyst.* **11**, 2144–2151.
32. Wang, X., Greenblatt, H.M., Bigman, L.S., Yu, B., Pletka, C.C., Levy, Y., Iwahara, J., (2021). Dynamic Autoinhibition of the HMGB1 Protein via Electrostatic Fuzzy Interactions of Intrinsically Disordered Regions. *J. Mol. Biol.* **433**, 167122.
33. Zaharias, S., Zhang, Z., Davis, K., Fargason, T., Cashman, D., Yu, T., Zhang, J., (2021). Intrinsically disordered electronegative clusters improve stability and binding specificity of RNA-binding proteins. *J. Biol. Chem.* **297**, 100945.
34. Krois, A.S., Dyson, H.J., Wright, P.E., (2018). Long-range regulation of p53 DNA binding by its intrinsically disordered N-terminal transactivation domain. *Proc. Natl. Acad. Sci.* **115**, E11302–E11310.
35. Sun, X., Dyson, H.J., Wright, P.E., (2020). A phosphorylation-dependent switch in the disordered p53 transactivation domain regulates DNA binding. *Proc. Natl. Acad. Sci. U. S. A.* **118**
36. Esquina, C.M., Garland-Kuntz, E.E., Goldfarb, D., McDonald, E.K., Hudson, B.N., Lyon, A.M., (2019). Intramolecular electrostatic interactions contribute to phospholipase Cβ3 autoinhibition. *Cell. Signal.* **62**, 109349.
37. Suetsugu, S., Miki, H., Takenawa, T., (2001). Identification of Another Actin-related Protein (Arp) 2/3 Complex Binding Site in Neural Wiskott-Aldrich Syndrome Protein (N-WASP) That Complements Actin Polymerization Induced by the Arp2/3 Complex Activating (VCA) Domain of N-WASP*. *J. Biol. Chem.* **276**, 33175–33180.
38. Huang, L., Agrawal, T., Zhu, G., Yu, S., Tao, L., Lin, J.B., Marmorstein, R., Shorter, J., Yang, X., (2021). DAXX represents a new type of protein-folding enabler. *Nature* **597**, 132–137.
39. Karlin, S., Brocchieri, L., Bergman, A., Mrázek, J., Gentles, A.J., (2002). Amino acid runs in eukaryotic proteomes and disease associations. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 333–338.
40. Lobanov, M.Y., Galzitskaya, O.V., (2012). Occurrence of disordered patterns and homorepeats in eukaryotic and bacterial proteomes. *Mol. Biosyst.* **8**, 327–337.
41. Lu, H., Wang, J., Bai, Y., Lang, J.W., Liu, S., Lin, Y., Cheng, J., (2011 2011,). Ionic polypeptides with unusual helical stability. *Nature Commun.* **21** (2), 1–9.
42. Vuzman, D., Hoffman, Y., Levy, Y., (2012). Modulating protein-DNA interactions by post-translational modifications at disordered regions. *Pac. Symp. Biocomput.*, 188–199.
43. Vuzman, D., Azia, A., Levy, Y., (2010). Searching DNA via a "Monkey Bar" Mechanism: The Significance of Disordered Tails. *J. Mol. Biol.* **396**, 674–684.
44. Levy, Y., Onuchic, J.N., Wolynes, P.G., (2007). Fly-casting in protein-DNA binding: Frustration between protein folding and electrostatics facilitates target recognition. *J. Am. Chem. Soc.* **129**, 738–739.
45. Bigman, L.S., Levy, Y., (2020). Tubulin tails and their modifications regulate protein diffusion on microtubules. *Proc. Natl. Acad. Sci.* **117**, 201914772.
46. Bigman, L.S., Levy, Y., (2020). Protein Diffusion on Charged Biopolymers: DNA versus Microtubule. *Biophys. J.* **118**, 3008–3018.
47. Holehouse, A.S., Das, R.K., Ahad, J.N., Richardson, M.O. G., Pappu, R.V., (2017). CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys. J.* **112**, 16–21.

Supporting Information for the paper:

Negatively charged disordered regions are prevalent and functionally important across proteomes

SI Methods

Protein functional classification

Proteins that their Gene Ontology (GO) molecular function annotation contained the words “DNA binding” or “RNA binding”, were defined as “Nucleic Acid Related” (NAR) proteins. Proteins that their GO molecular functions contained the words “ase activity” were defined as “Enzymes”, and the rest of the proteins were in the Control group. Proteins that did not have a manually annotated GO molecular function were not included in our analysis.

Disorder prediction

Disorder propensity for the human proteome in the main text was predicted using the fIDPnn algorithm (1), which had the highest F1 score in the recent Critical Assessment of Intrinsic Disorder prediction (3). Since fIDPnn requires extensive computational resources, disorder of the mouse and yeast proteomes were preformed using Iupred2A (2), which is orders of magnitude faster (see Figs S1-5). For comparison, we also include in the SI the analysis of the human proteome using Iupred2.

SI Figures

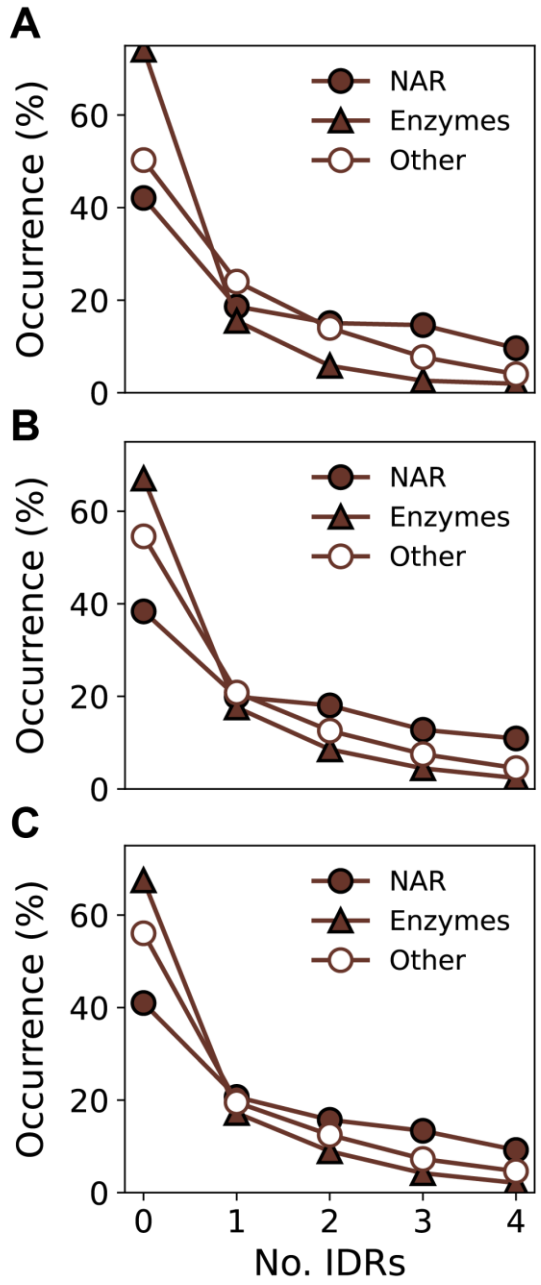


Figure S1. Occurrence and charge properties of IDRs in three functional protein classes. (A) Proteins in the Yeast proteome (4,478 proteins with annotated function) were divided into three functional classes: Nucleic Acid Related proteins (NAR; 1143 proteins), Enzymes (whose function is not dependent on nucleic acids; 1,639 proteins) and a Control group with the rest of the annotated proteins in the human proteome (1,696 proteins). Proteins were classified based on Gene Ontology (GO) molecular function. For each functional class, the fraction of proteins that have either 0, 1, 2, 3 or 4 IDRs are shown IDRs are defined as stretches of at least 15 residues whose disorder Iupred2 score(2) is greater than 0.5. (B) same as in (A) but for mouse. In Mouse there are 13,963 proteins with annotated function, 2,771 NARs, 4,187 enzymes and 7,005 in the control group. (C) Same as A and B, but for Human.

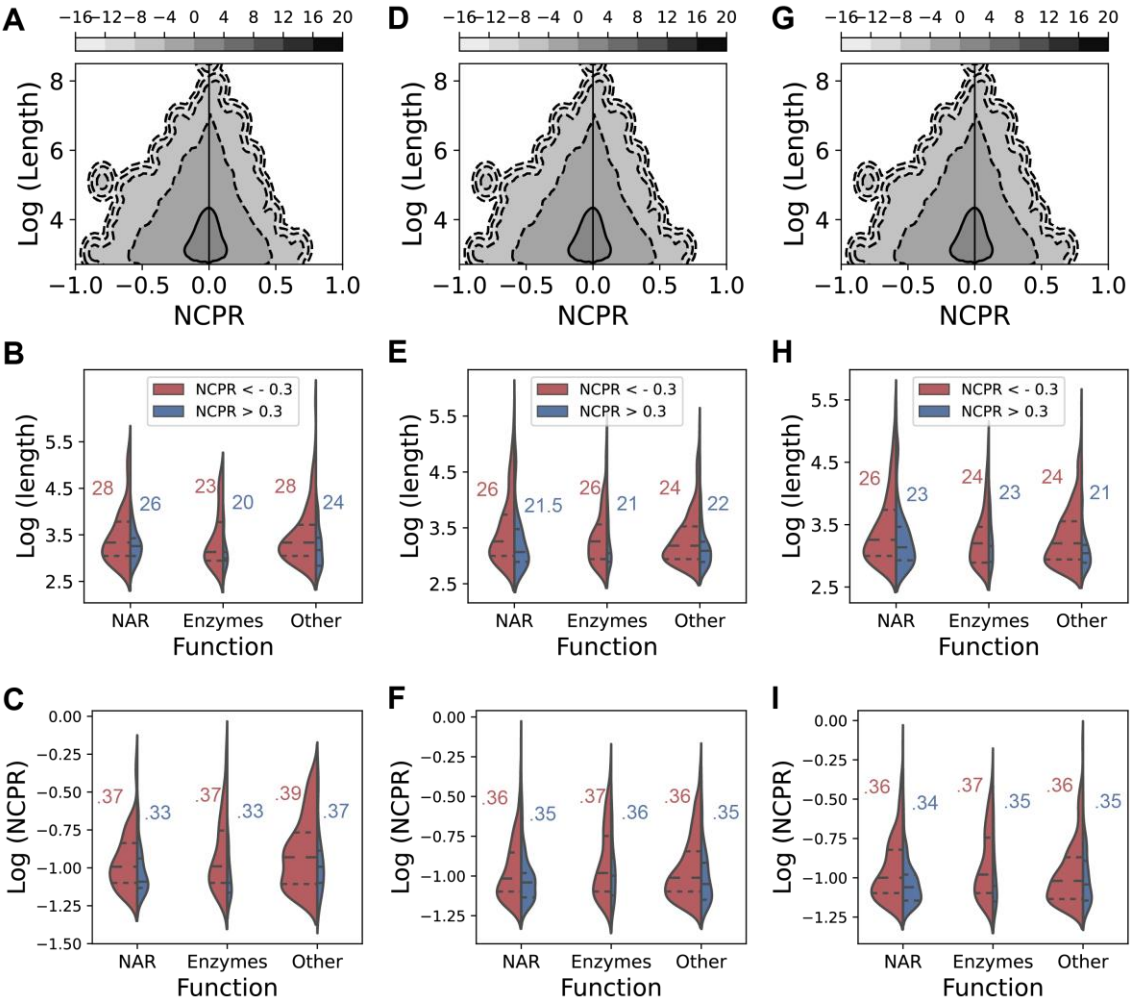


Figure S2. Negatively charged IDRs are longer than positively charged IDRs and more highly charged. (A) Contour plot showing the length (logarithmic scale) versus NCPR for all human proteins. (B) Violin plot showing the distribution of length in each functional class, for highly negatively (NCPR < -0.3; red) and positively charged IDRs (NCPR > 0.3; blue). The median values of length of each group is shown in the corresponding color. The p-value of the negative and positive IDRs for the NAR, Enzyme and control groups are $4x10^{-2}$, $4x10^{-2}$, and, $6x10^{-2}$, respectively. (C) Similar to (B), but showing NCPR. The median values of the NCPR for the six groups of IDRs are highlighted on the plot. The p-value of the negative and positive IDRs for NAR, Enzyme and control groups are $3x10^{-2}$, $1x10^{-2}$, 0.3, respectively. The sign of the NCPR of the negatively charged IDR is neglected in the plot. (D-F) same as A-C, but for Mouse. P-values for the length for NAR, Enzyme and control groups are: $2x10^{-3}$, $2x10^{-2}$, and, $6x10^{-4}$, respectively. P-values for NCPR of the NAR, Enzyme and control groups are: $4x10^{-3}$, $9x10^{-2}$, and, $1x10^{-2}$, respectively. (G-I) same as A-C, but for Human. P-values for the length for NAR, Enzyme and control groups are: $3x10^{-4}$, 0.2, and, $1x10^{-5}$, respectively. P-values for NCPR of the NAR, Enzyme and control groups are: $4x10^{-7}$, $9x10^{-3}$ and, 0.2, respectively.

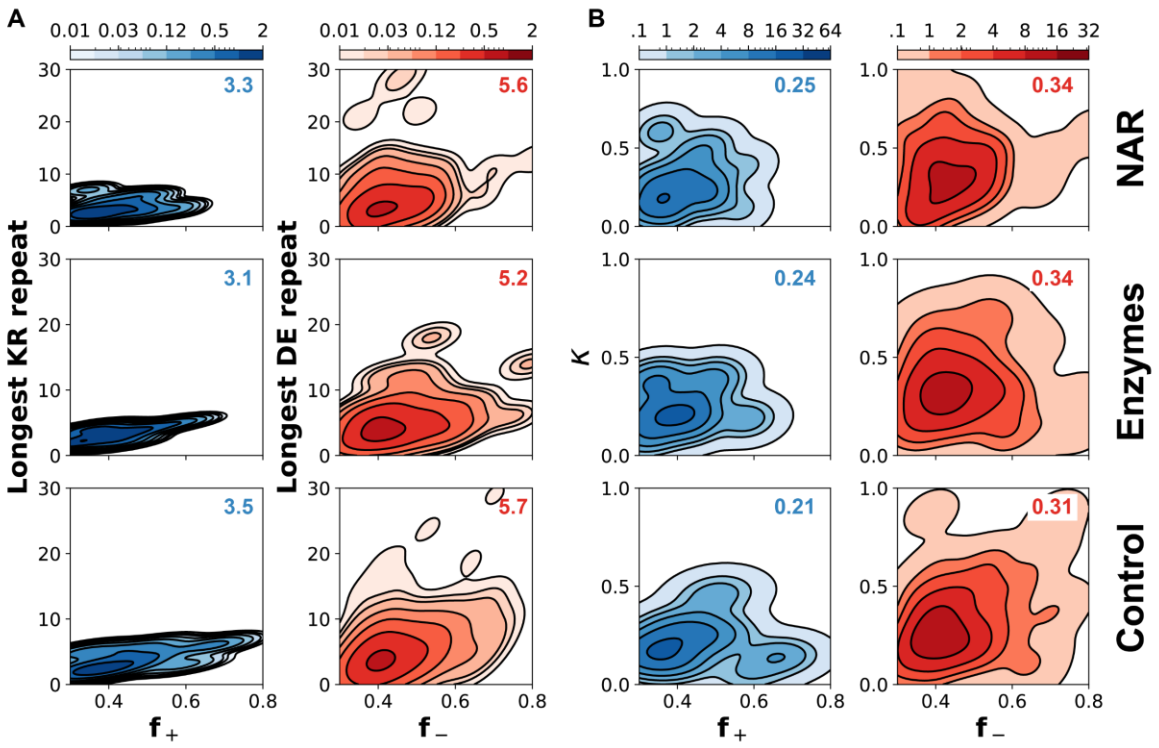


Figure S3. Charge segregation in charged IDRs in Yeast. (A) The length of the longest stretch of consecutive positively (KR repeats, blue) or negatively charged (DE repeats, red) residues, is shown as a function of the fraction of positively (f_+) or negatively charged (f_-) residues in a given IDR sequence. (B) Same as in (A), but showing κ (see main text for definition) as a function of the f_+ and f_-. Mean values the longest KR or DE repeats (in amino acids) and of κ are shown on each panel. Data are shown for NAR proteins (top row), Enzymes (middle row) and the control group (bottom row).

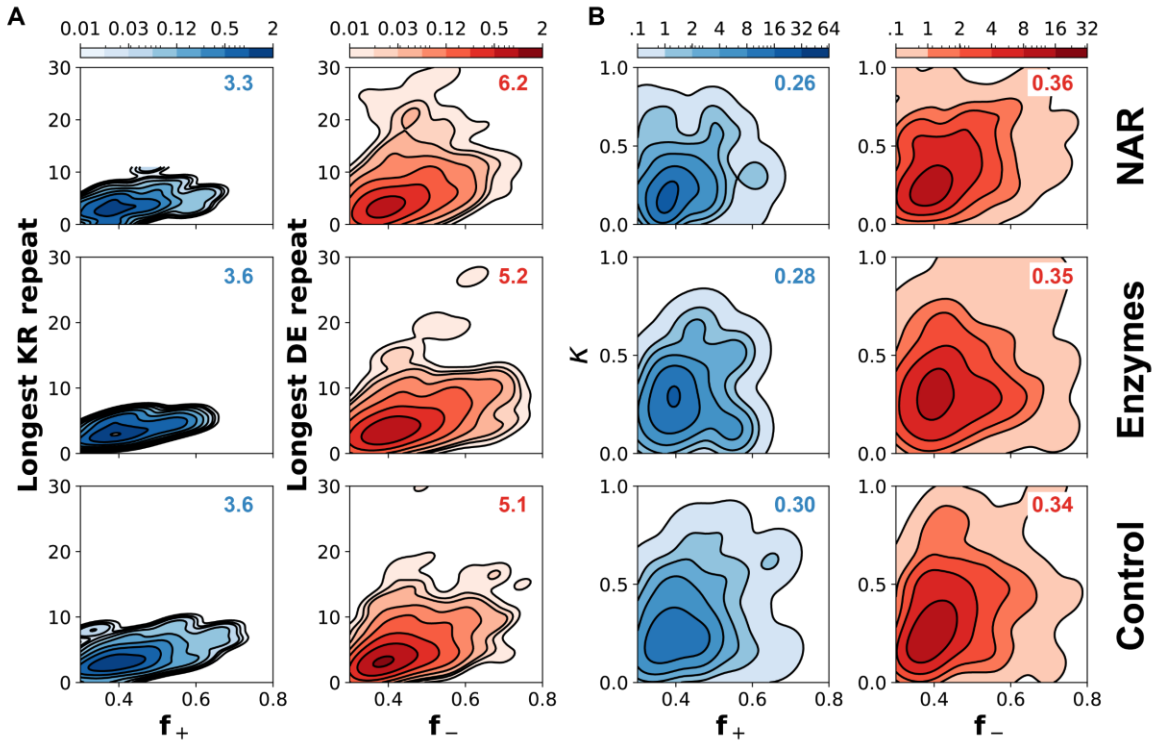


Figure S4. Charge segregation in charged IDRs in Mouse. Same as Fig. S3, but showing data for the mouse proteome.

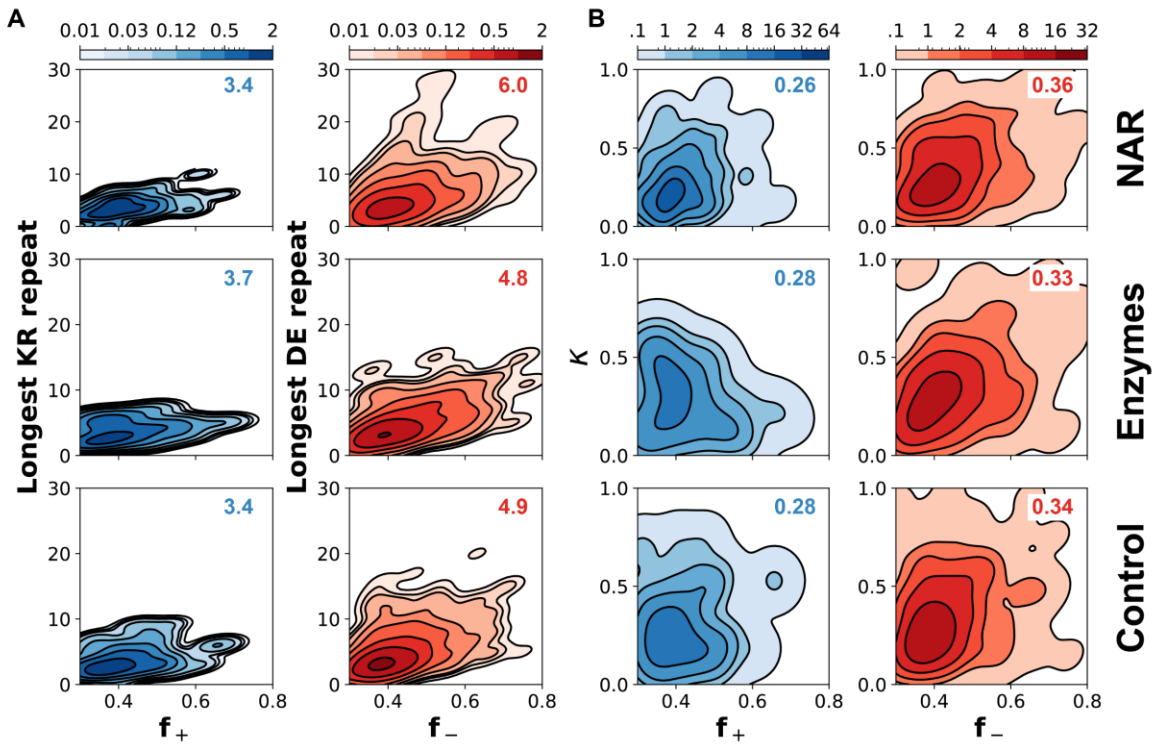


Figure S5. Charge segregation in charged IDRs in Human. Same as Fig. S3-4, but showing data for the human proteome.