Disease causing mutations are found in hydrophobic regions in both ordered and disordered proteins

Ruchi Lohia[1], Matt Hansen[2], Grace Brannigan[1,3*],

1 Center for Computational and Integrative Biology, Rutgers University, Camden, NJ, USA
2 Department of Genetics and Center of Excellence in Environmental Toxicology, University of Pennsylvania
3 Department of Physics, Rutgers University, Camden, NJ, USA

* grace.brannigan@rutgers.edu(GB)

Abstract

Author summary

Introduction

The physiological significance of intrinsically disordered proteins (IDPs), which can explore a wide range of conformational ensembles in their functional form, is now well-established [1–4]. More than 33% of eukaryotic proteins contain disordered regions longer than 30 residues [3], many of which are involved in critical biological functions, including transcriptional regulation and cell signaling [5]. Long intrinsically disordered regions are particularly abundant among cancer and neurodegenerative-associated proteins [6, 7].

IDP amino-acid sequences tend to be low complexity and include numerous charged residues, often in long repeats [1]. In contrast to ordered proteins, in which a complex sequence encodes a well-defined tertiary structure, an IDP sequence determines a heterogeneous conformational ensemble.

Although IDP sequences are low-complexity and do not encode a well-defined structure, single residue substitutions can still have functional effects that are significant for the organism. More than 20% of disease-associated missense single nucelotide polymorphisms (SNPs) are found in IDPs [11]; although detectable, the relatively subtle functional effects may lead to relatively weak selection pressure, whether positive or negative, allowing the mutation to persist at high frequencies within a population. Numerous structural and simulation studies [12–18] have demonstrated clear effects of single charged-residue insertion, deletion, or substitutions on conformational ensemble and aggregation of IDPs monomers. Single charged residue mutations or post translational modifications that change charges will affect the sequence electrostatics predicted to determine ensemble properties simply from statistical physics models, and in short-chains, can also induce qualitative changes by changing the appropriate regime [9, 12, 19, 20]. Locally, such mutations can modulate residual secondary structure preferences via forming or breaking local salt-bridges or by introducing helix breaking residues [13, 17, 21].

For IDPs with a relatively low fraction of charged residues, typical of the Janus region of the state diagram proposed by Das and Pappu [9, 10], more subtle differences

among neutral amino-acids play an increasingly important role in determining the ensemble. More than 15% of disease-associated IDP polymorphisms are substitutions between two charge-neutral residues [11]. The extent to which such substitutions in IDPs can affect non-local aspects of the conformational ensemble is uncertain; these substitutions directly affect short-range interactions, and structure-based coupling between distant residues in IDPs is expected to be weak. Nonetheless, correlations between secondary structure of distant residues has been frequently observed in IDPs [13, 22]; for example, several cancer mutations in transactivation domain of tumor suppressor p53 can lead to helicity changes in residues sequentially far away from the mutation sites [13].

1 Results and discussion

SNP distributions in ordered and disordered regions

Among the total 58,645 SNPs analyzed, 47.6% lies in disordered regions and remaining 52.3% in ordered regions. When the SNPs were further broken down into Disease-associated (DS) and not-disease-associated (nDS), we find that 64.2% of DS are found in ordered region and 35.8% are found in disordered region (Fig 1). As observed previously [11], we find that disease associated mutations are enriched by 0.49 fold in ordered regions and 0.37 fold depleted depleted in disordered regions (Fig 1).

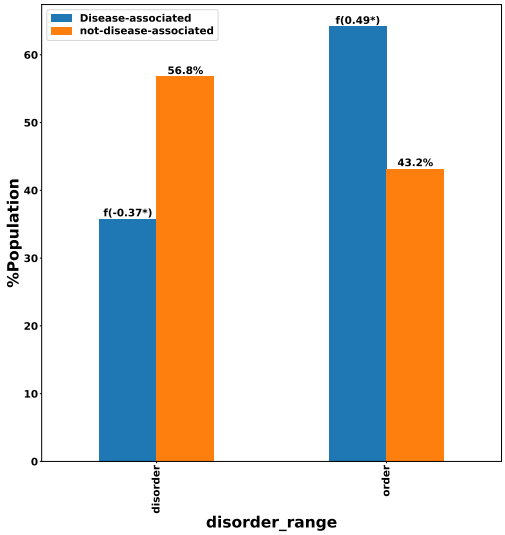

Fig 1. Disease mutations have higher frequency in ordered regions. The population of DS SNPs (blue bar) and nDS SNPs (orange bar) in ordered region and disordered region. The expected population for DS SNPs in ordered and disordered set from nDS SNPs is labeled at the top of orange bars. Fold enrichment in DS SNPs when compared with not-disease-associated SNPs is annotated in the plot as well with f. If p-value from the binomial test is $< .005$, the enrichment is marked with star.

Type of SNP distribution

About 50% of the disease causing SNPs in both ordered and disordered proteins are charge-neutral (Fig 3).

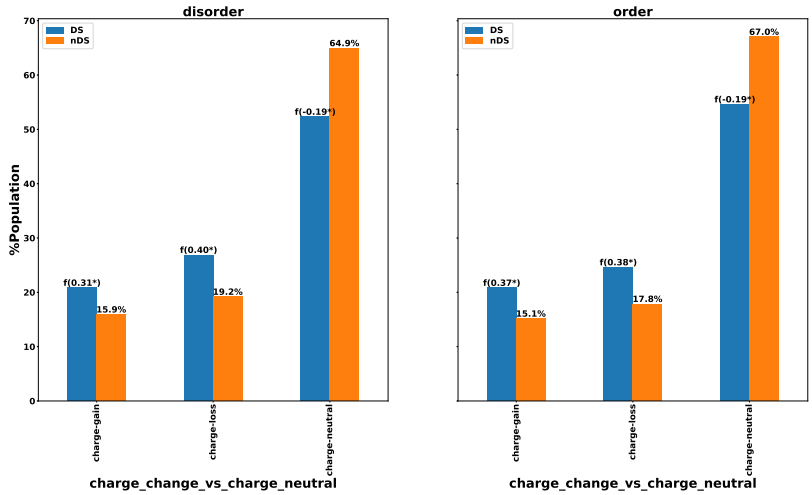

Fig 2. DS SNPs are more enriched in gain/loss of charge. We divided each SNP into three categories: charge-neutral SNPs, SNPs that lead to charge gain and SNPs that lead to charge loss. We then plotted the population for DS and nDS SNPS in ordered and disordered proteins in each category. We find that 50% of the disease causing SNPs in both ordered and disordered proteins are charge-neutral. Fold enrichment in DS SNPs when compared with not-disease-associated SNPs is annotated in the plot as well with f. If p-value from the binomial test is $< .005$, the enrichment is marked with star. DS SNPs are enriched in charge-gain or charge-loss but depleted in charge-neutral SNPs.

DS SNPs are enriched in hydrophobic regions

We looked at the hydrophobicity distribution of SNP residue and it's neighboring residue (window size 3,5 and 15). <H> at window size 3 is defined as the Hydropathy score of

$$\frac{H_i + H_{i-1} + H_{i+1}}{2} \tag{1}$$

where i is the SNP residue. We find that the DS SNP is more enriched in hydrophobic regions in both ordered and disordered proteins.

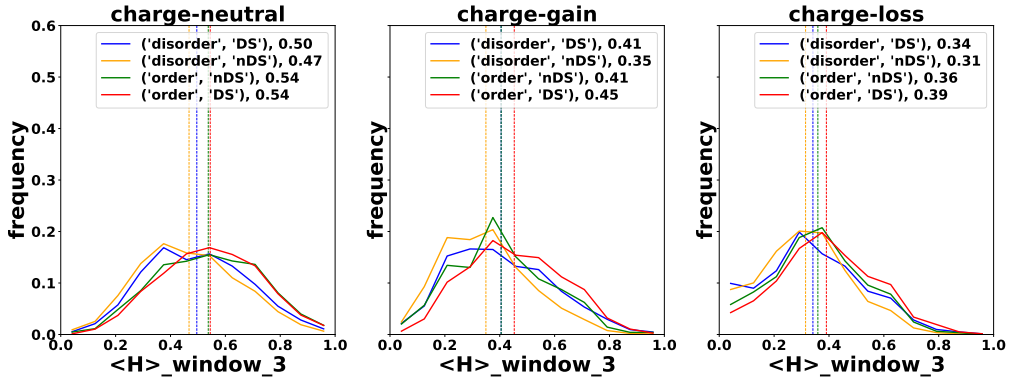

Fig 3. DS SNPs are more enriched in hydrophobic regions. We looked at the hydrophobicity distribution of SNP residue and it's neighboring residue. The mean of each histogram distribution is also reported in caption. We find that DS SNPs are found in hydrophobic regions when compared with nDS SNPs in both ordered and disordered proteins.

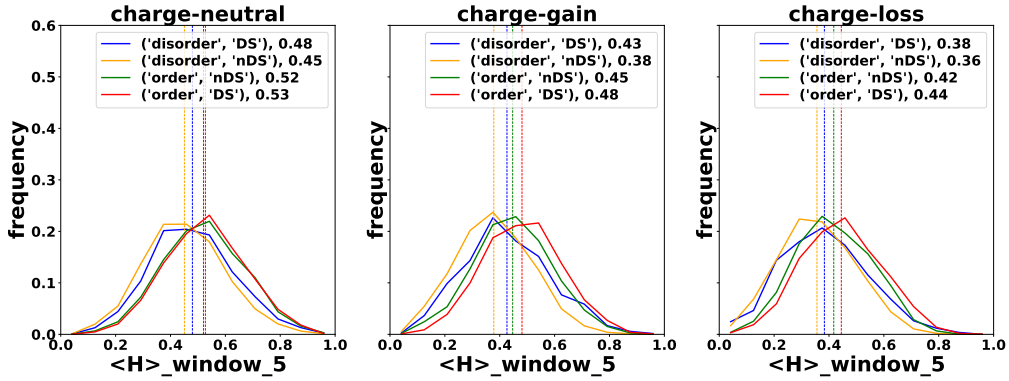

Fig 4. DS SNPs are more enriched in hydrophobic regions. Same as Fig 3 but with window size 5 instead of 3.

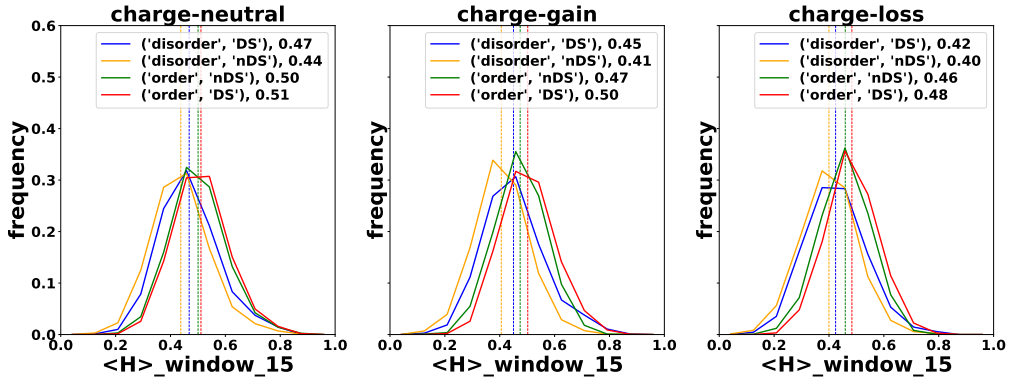

Fig 5. DS SNPs are more enriched in hydrophobic regions. Same as Fig 3 but with window size 15 instead of 3.

DS SNPs are enriched in hydrophobic domains

The sequence was divided into hydrophobic domains as described in Methods. We find that DS SNPs are enriched in h domains in both ordered and disordered proteins.

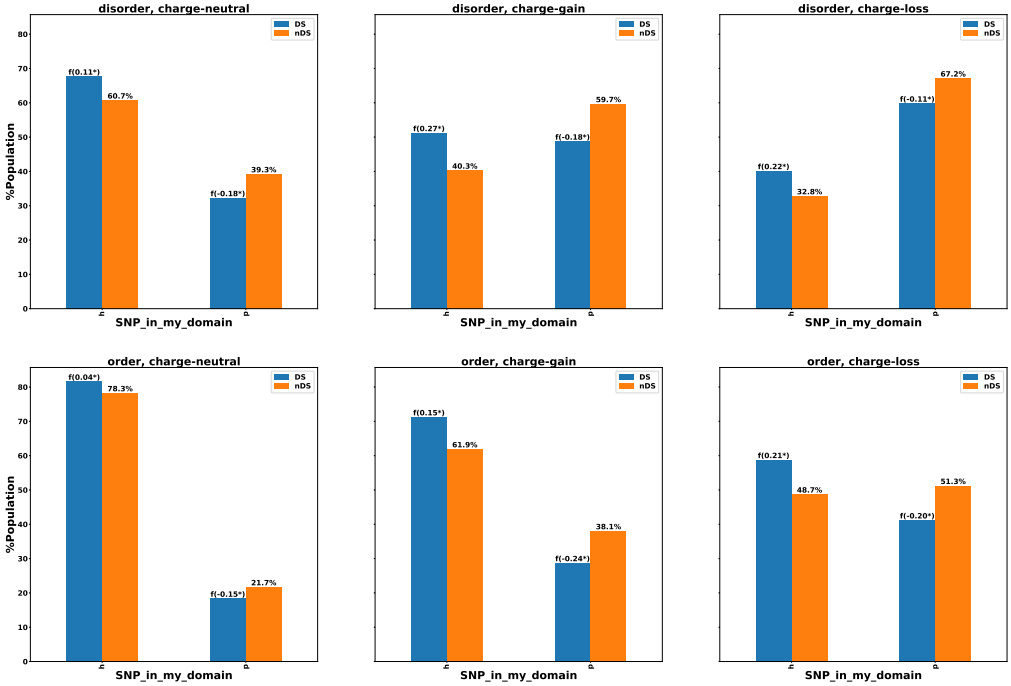

Fig 6. DS SNPs are enriched in hydrophobic domains. We looked at the population of SNPs in hydrophobic domains (h) and linker regions (p). The top and bottom panel is for disordered and ordered regions respectively. The sequence was divided into hydrophobic domains as described in Methods. We find that DS SNPs are enriched in h domains in both ordered and disordered proteins.Fold enrichment in DS SNPs when compared with not-disease-associated SNPs is annotated in the plot as well with f. If p-value from the binomial test is $< .005$, the enrichment is marked with star.

Mutations change hydrophobic regions to less hydrophobic regions/ domain transformation

We find that SNPs are enriched in conversion of a hydrophobic domain to a linker region in both ordered and disordered proteins.

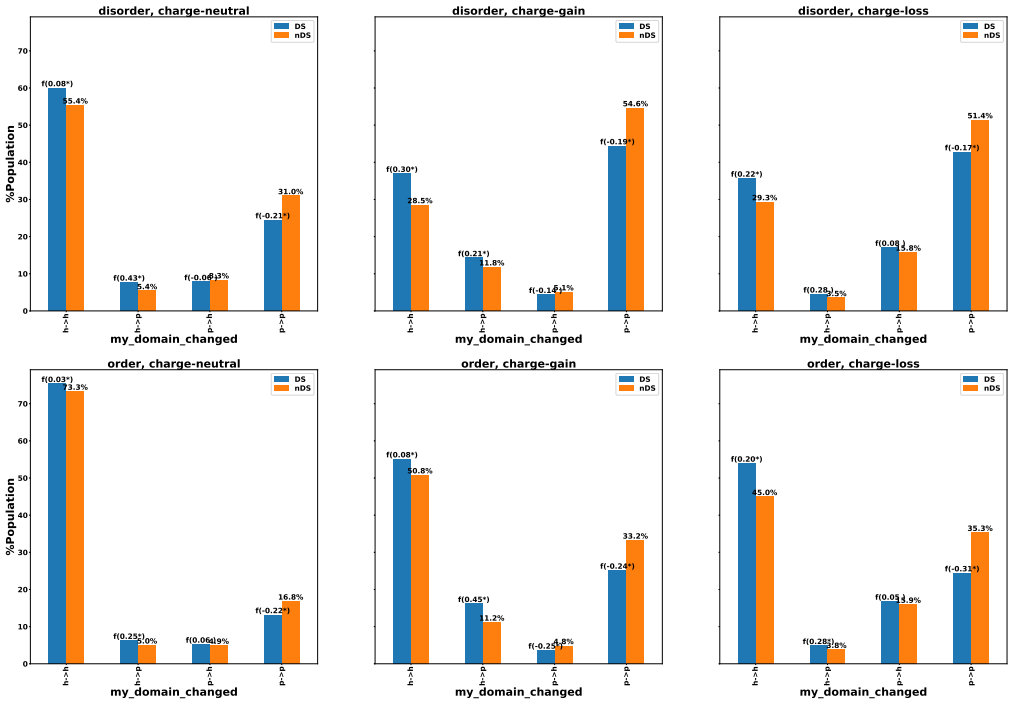

Fig 7. DS SNPs change hydrophobic regions to less hydrophobic regions. We looked at the populations of domain change associated with each SNPs in each category. DS SNPs in hydrophobic domains are enriched in changing these hydrophobic domains (h) to linker regions (p) in both disordered and ordered proteins. The top and bottom panel is for disordered and ordered regions respectively. Fold enrichment in DS SNPs when compared with not-disease-associated SNPs is annotated in the plot as well with f. If p-value from the binomial test is $< .005$, the enrichment is marked with star.

The addition of Methionine

We find that the addition of methionine is highly enriched with disease associations.

Few example of domain and SNP identification in proteins.

The online tool for analyzing your own protein sequence :)

Materials and Methods

Datasets

The list of all missense variants annotated in human UniProtKB/Swiss-Prot entries was obtained from http://beta.uniprot.org/docs/humsavar (last Release: 8th May 2019) [23]. This manually curated catalog contains missense mutations on the most common isoform of the given protein and does not contain frameshift and nonsense mutation. A variant is annotated as 'Disease' or 'Polymorphisms' depending on if it is implicated in disease or not according to literature reports. 'Polymorphisms' is also used to describe rare variants as well as polymorphisms that have an effect on protein function, but with no resulting clinical phenotype (functional polymorphisms) [23]. A total number of 78,678 missense mutations were found in the database, among which 30,597 (38.9%) are associated to diseases, 40,032 (50.1%) are polymorphisms, and 1,973 (10.2%) are still unclassified.

The initial set of mutations was filtered as follows: The proteins associated with disease mutations were clustered using UniRef50 [24]. Only one protein from each cluster was selected. 'Polymorphisms' dataset was analogously filtered as well. We further removed three proteins with an unusually high number of annotated disease mutations (UniProtKB: P35555, P35498, P00451).

Protein disorder for wild type sequences was obtained from Database of Disordered Protein Prediction (D2P2) (http://d2p2.pro). D2P2 has disorder predicted from nine disorder predictor including PONDR VL-XT, PONDR VSL2b, PrDOS, PV2, Espritz (all variants) and IUPred (all variants) [25]. We annotated any residue as disordered if at-least two disorder predictor predicts it be disordered.

Domain identification

Mean hydrophobicity ($\langle H \rangle$) at each residue is defined as the average Kyte-Dolittle [26] score with a window size of 3 residues, scaled to fit between 0 and 1. Any stretch of four or more residues with $\langle H \rangle > 0.37$ is classified as hydrophobic domain and from the remaining residues, stretch of four or more residues is classified as linker region.

Acknowledgments

Supporting Information

References

1. Uversky VN. Unusual biophysics of intrinsically disordered proteins. Biochim Biophys Acta. 2013;1834(5):932–51. doi:10.1016/j.bbapap.2012.12.008.

2. Panchenko AR, Babu MM. Editorial overview: Linking protein sequence and structural changes to function in the era of next-generation sequencing. Curr Opin Struct Biol. 2015;32:viii–x. doi:10.1016/j.sbi.2015.06.005.

3. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. J Mol Biol. 2004;337(3):635–645. doi:10.1016/j.jmb.2004.02.002.

4. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol. 2005;6(3):197–208. doi:10.1038/nrm1589.

5. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J. 2005;272(20):5129–5148. doi:10.1111/j.1742-4658.2005.04948.x.

6. Habchi J, Tompa P, Longhi S, Uversky VN. Introducing Protein Intrinsic Disorder. Chem Rev. 2014;114(13):6561–6588. doi:10.1021/cr400514h.

7. Babu MM, van der Lee R, de Groot NS, Gsponer J. Intrinsically disordered proteins: regulation and disease. Curr Opin Struct Biol. 2011;21(3):432–40. doi:10.1016/j.sbi.2011.03.011.

8. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, et al. DisProt: the Database of Disordered Proteins. Nucleic Acids Res. 2007;35(Database):D786–D793. doi:10.1093/nar/gkl893.

9. Das RK, Ruff KM, Pappu RV. Relating sequence encoded information to form and function of intrinsically disordered proteins. Curr Opin Struct Biol. 2015;32:102–112. doi:10.1016/j.sbi.2015.03.008.

10. Das RK, Pappu RV. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc Natl Acad Sci U S A. 2013;110(33):13392–7. doi:10.1073/pnas.1304749110.

11. Vacic V, Markwick PRL, Oldfield CJ, Zhao X, Haynes C, Uversky VN, et al. Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. PLoS Comput Biol. 2012;8(10):e1002709. doi:10.1371/journal.pcbi.1002709.

12. Larini L, Gessel MM, LaPointe NE, Do TD, Bowers MT, Feinstein SC, et al. Initiation of assembly of tau(273-284) and its ΔK280 mutant: an experimental and computational study. Phys Chem Chem Phys. 2013;15(23):8916. doi:10.1039/c3cp00063j.

13. Ganguly D, Chen J, Dyson H, Wright P, Uversky V, Oldfield C, et al. Modulation of the Disordered Conformational Ensembles of the p53 Transactivation Domain by Cancer-Associated Mutations. PLOS Comput Biol. 2015;11(4):e1004247. doi:10.1371/journal.pcbi.1004247.

14. Viet MH, Nguyen PH, Derreumaux P, Li MS. Effect of the English Familial Disease Mutation (H6R) on the Monomers and Dimers of Aβ40 and Aβ42. ACS Chem Neurosci. 2014;5(8):646–657. doi:10.1021/cn500007j.

15. Viet MH, Nguyen PH, Ngo ST, Li MS, Derreumaux P. Effect of the Tottori Familial Disease Mutation (D7N) on the Monomers and Dimers of Aβ 40 and Aβ 42. ACS Chem Neurosci. 2013;4(11):1446–1457. doi:10.1021/cn400110d.

16. Truong PM, Viet MH, Nguyen PH, Hu CK, Li MS. Effect of Taiwan Mutation (D7H) on Structures of Amyloid-β Peptides: Replica Exchange Molecular Dynamics Study. J Phys Chem B. 2014;118(30):8972–8981. doi:10.1021/jp503652s.

17. Zhan YA, Wu H, Powell AT, Daughdrill GW, Ytreberg FM. Impact of the K24N mutation on the transactivation domain of p53 and its binding to murine double-minute clone 2. Proteins Struct Funct Bioinforma. 2013;81(10):1738–1747. doi:10.1002/prot.24310.

18. Xu L, Shan S, Wang X. Single Point Mutation Alters the Microstate Dynamics of Amyloid β-Protein Aβ42 as Revealed by Dihedral Dynamics Analyses. J Phys Chem B. 2013;117(20):6206–6216. doi:10.1021/jp403288b.

19. Bah A, Forman-Kay JD. Modulation of Intrinsically Disordered Protein Function by Post-translational Modifications. J Biol Chem. 2016;291(13):6696–6705. doi:10.1074/jbc.R115.695056.

20. He Y, Chen Y, Mooney SM, Rajagopalan K, Bhargava A, Sacho E, et al. Phosphorylation-induced Conformational Ensemble Switching in an Intrinsically Disordered Cancer/Testis Antigen. J Biol Chem. 2015;290(41):25090–25102. doi:10.1074/jbc.M115.658583.

21. Conicella AE, Zerze GH, Mittal J, Fawzi NL, Alexander Conicella AE, Zerze GH, et al. ALS Mutations Disrupt Phase Separation Mediated by α-Helical Structure in the TDP-43 Low-Complexity C-Terminal Domain. Struct Des. 2016;24(9):1537–1549. doi:10.1016/j.str.2016.07.007.

22. Iešmantavičius V, Jensen MR, Ozenne V, Blackledge M, Poulsen FM, Kjaergaard M. Modulation of the Intrinsic Helix Propensity of an Intrinsically Disordered Protein Reveals Long-Range Helix–Helix Interactions. J Am Chem Soc. 2013;135(27):10155–10163. doi:10.1021/ja4045532.

23. Yip YL, Famiglietti M, Gos A, Duek PD, David FPA, Gateau A, et al. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. Hum Mutat. 2008;29(3):361–366. doi:10.1002/humu.20671.

24. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31(6):926–932. doi:10.1093/bioinformatics/btu739.

25. Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, et al. D2P2: database of disordered protein predictions. Nucleic Acids Res. 2012;41(D1):D508–D516. doi:10.1093/nar/gks1226.

26. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157(1):105–32.