Determining the Chemistry and Functionality of the *Caenorhabditis* Disordered Proteomes

McFadden, WM[1]; Buszczak, M[3]; Yanowitz, JL[1,2]

[1]Magee-Womens Research Institute, Pittsburgh, PA. [2]University of Pittsburgh School of Medicine, Pittsburgh, PA. [3]University of Texas Southwestern Medical Center, Dallas, TX

ABSTRACT

While structure has long guided our understanding of protein function, many proteins lack rigid secondary or tertiary structures are responsible for critical cellular functions. Certain amino acid compositions have an increased tendency to favor a disordered state rather than a highly-packed peptide. Over 40% of proteins in Caenorhabditis elegans are predicted to contain an intrinsically disordered region (IDR) of at least 30 amino acids in length. Furthermore, the genome encodes for many intrinsically disordered proteins (IDPs) that contain little-to-no structured domains. IDPs and IDRs have been previously implicated in many of biology's greatest questions such as the origin of life, emergence of multicellularity, and evolution of sexual reproduction (Kulkarni & Uversky, 2018). We sought to study IDRs and IDPs in C. elegans and its genus to further to understand the evolutionary history of these fascinating proteins.

We utilized previously described programs to identify the intrinsically disordered proteins within the *Caenorhabditis* genus. We further analyzed the chemical properties and the cellular functions of these IDPs/IDRs at a proteome-wide scale. An increased level of disorder correlated with an increased median isoelectric point (pI), which may be due to an enrichment of lysine. GO term analyses of the *C. elegans* and *C. briggsae* disordered proteome indicate that IDPs, which are considerably disordered along the entire peptide, were enriched in various nucleic-acid functions. This study will be extended to *Drosophilae* to determine if what we have seen in *Caenorhabditis* is conserved. These findings lay the groundwork for the continual progress of this study to investigate the *C. elegans* disordered proteome and assist future disorder-based studies.

OBJECTIVES

Utilize known features of IDPs to characterize evolutionary changes of the Intrinsically Disordered Proteome at the genus-level.

- Calculate the amino acid compositions and biochemical properties of IDPs and compare these to highly structured proteins.
- Use the comparisons of IDPs and structured proteins to identify long-term chemical conservation between genera.

Utilize bioinformatic tools to further elucidate IDP functionality based on chemical compositions.

MATERIALS

R Package	Use	Citation
dplyr	Data management and manipulation.	Wickham, H., et al. (2020). Version 0.8.5. https://CRAN.R-project.org/package=dplyr
ggalt	Data visualization (aesthetics).	Rudis, B., B. Bolker and J Schulz (2017). Version 0.4.0. https://CRAN.R-project.org/package=ggalt
ggplot2	Data visualization.	Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
ggpubr	Statistics (Wilcoxon test) and Data visualization.	Kassambara, A. (2020). Version 0.2.5. https://CRAN.R-project.org/package=ggpubr
ggtree	Visualization of phylogenies.	Yu, G., et al. Methods in Ecology and Evolution 2017, 8(1):28-36. & Yu, G., et al. Molecular Biology and Evolution 2018, 35(2):3041-3043.
RColorBrewer	Data visualization (aesthetics).	Neuwirth, E. (2014). Version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer
seqinr	Management of fasta files and sequence data.	Charif, D. and Lobry, J.R. (2007)
stringr	Data management and manipulation.	Wickham, H. (2019). Version 1.4.0. https://CRAN.R-project.org/package=stringr
tidyr	Management and cleaning of data sets	Wickham, H. and L. Henry (2020). tidyr: Tidy Messy Data. R version 1.0.2. https://CRAN.R-project.org/package=tidyr

Table 1: R Packages Used

R version 3.6.1 and RStudio version 1.2.1335 were used for sequence analysis, data management, and visualization (R Core Team, 2019; RStudio Team, 2018).

Sequences were retrieved from the UniProt Reference Proteomes, release 2020_01, 26-Feb-2020 (UniProt Consortium, 2019). One protein sequence per gene is represented in data the to reduce redundancy. Proteome ID and Species provided: *C. japonica* (DF5081) [UP000005237]; *C. elegans* (Bristol N2) [UP000001940]; *C. briggsae* (AF16) [UP000008549]; *C. remanei* (PB4641) [UP000008281]; *C. tropicalis* [UP000095282]; *C. brenneri* (PB2801) [UP000008068]; *C. latens* (PX534) [UP000216463]; *C. nigoni* (JU1422) [UP000230233]; *D. melanogaster* (Berkeley) [UP000000803]; *D. virilis* (Tucson 15010-1051.87) [UP000008792]; *D. simulans* (mosaic) [UP000000304]; *H. sapiens* [UP000005640]; *M. musculus* (C57BL/6J) [UP000000589].

Phylogenies from (Clark et al., 2007; Stevens et al., 2019).

METHODS AND RESULTS

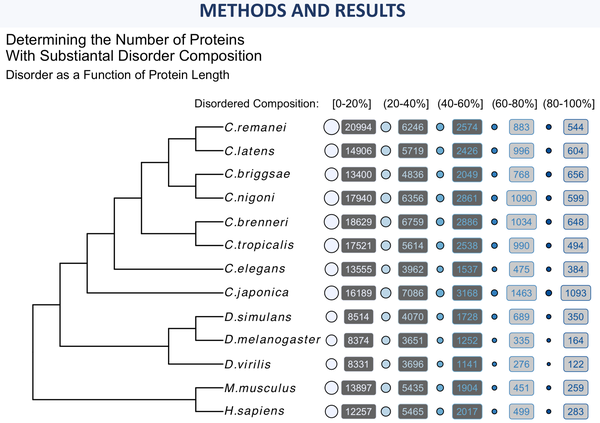

Figure 1: Caenorhabditis Species Have More Highly Disordered Proteins Than The Other Eukaryotes Analyzed.

RAPID was used to predict the composition of disorder for each sequence based on many known features of IDPs. Predictions are done on a proteome-wide scale and the output is a percentage of disordered residues within each peptide (Yan et al., 2013). Proteins were binned into 5 categories based on their RAPID score: [0-20%], (20-40%], (40-60%], (60-80%], (80-100%].

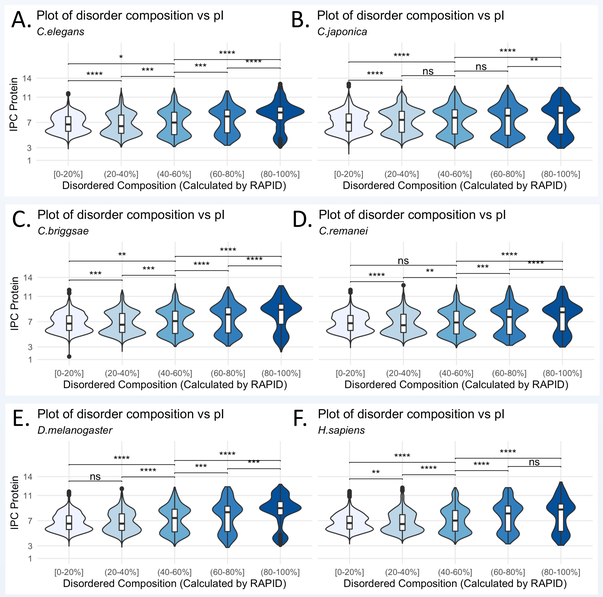

Figure 2: Disordered Proteins Tend to have an Increase of pI.

The Isoelectric Point Calculator was used to determine the pI for each protein using the IPC_Protein pKa values (Kozlowski, 2016). Violin plots are normalized for each bin to show density of pI frequency for each bin from Fig. 1. **A)** *C. elegans*, **B)** *C. japonica*, **C)** *C. briggsae,* **D)** *C. remanei,* **E)** D. *melanogaster,* **F)** *H. sapiens.*

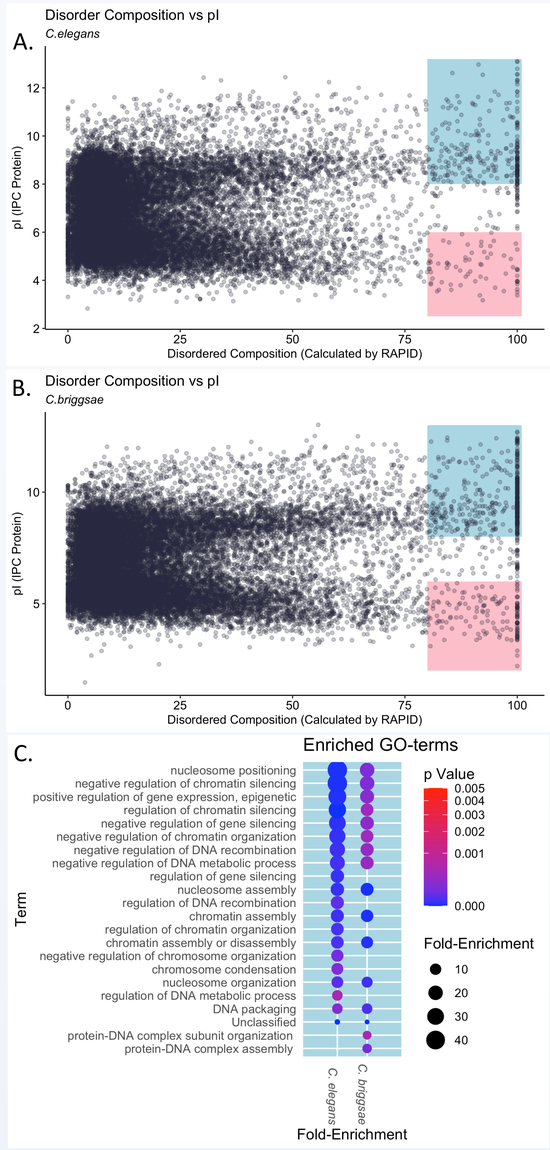

Figure 3: Basic IDPs are Enriched for DNA-related Processes.

A,B) Scatterplot of Disorder Composition, calculated by RAPID (Yan et al, 2013), and pI, calculated by IPC (Kozlowski, 2016), for the proteomes of *C. elegans* (A) and *C. briggsae* (B). Shaded regions represent the proteins used for GO-term enrichment analysis. Pink, acidic IDPs; light-blue, basic IDPs. **C)** Enriched GO-terms for basic IDPs are all involved in DNA-related processes. Acidic IDPs were only enriched in "Unclassified" proteins (not shown, enriched 1.66-fold in both *C. elegans* and *C. briggsae,* p = 1.63e-09 or 1.56e-27, respectively).

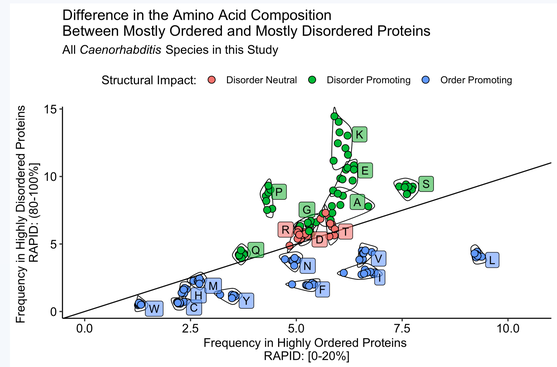

Figure 4: IDPs are Highly Enriched in Lysine for All *Caenorhabditis* Species.

A comparison of amino acid frequency in the most ordered vs the most disordered proteins (RAPID score 0-20% and 80-100%, respectively) for each *Caenorhabditis* species in this study. Diagonal line represents a 1:1 ratio, divergence from the line indicates enrichment within each bin. Residues colored for known structural tendencies (Kulkarni & Uversky, 2018).

CONCLUSIONS

- Increasing disorder composition trends toward increasing basicity. In the group of the most disordered proteins, lysine is the most enriched residue.
- While acidic IDPs are mostly uncharacterized, basic IDPs have many roles in DNA-related processes.
- Many Caenorhabditis proteins are IDPs. WormBase does not support disordered annotations, so Disordered Composition should be kept in mind forfuture protein studies.

FUTURE DIRECTIONS

- Validate data using additional disorder prediction programs.
- Further investigate the role of lysine in IDP function.
- Learn about evolutionary history of IDPs by looking at conservation across the *Caenorhabditis* genus.
- Use this as a set to identify uncharacterized proteins involved in fertility.

REFERENCES

Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., . . . *Broad Institute Genome Sequencing, P. (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature, 450(7167), 203-218. doi:10.1038/nature06341

Kozlowski, L. P. (2016). IPC–isoelectric point calculator. Biology direct, 11(1), 55.

Kulkarni, P., & Uversky, V. N. (2018). Intrinsically Disordered Proteins: The Dark Horse of the Dark Proteome. PROTEOMICS, 18(21-22), 1800061. doi:10.1002/pmic.201800061

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Research, 47(D1), D419-D426.

R Core Team. (2019). R: A language and environment for statistical computing (Version 3.6.1). Vienna, Austria: R Foundation for Statistical Computing, . Retrieved from https://www.R-project.org/.

RStudio Team. (2018). RStudio: Integrated Development for R. Boston, MA: RStudio, Inc.,. Retrieved from http://www.rstudio.com/.

Stevens, L., Félix, M.-A., Beltran, T., Braendle, C., Caurcel, C., Fausett, S., . . . Blaxter, M. (2019). Comparative genomics of 10 new Caenorhabditis species. Evolution Letters, 3(2), 217-236. doi:10.1002/evl3.110

The Gene Ontology Consortium. (2018). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research, 47(D1), D330-D338. doi:10.1093/nar/gky1055

UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research, 47(D1), D506-D515.

Uversky, V. N. (2019). Intrinsically disordered proteins and their "mysterious"(meta) physics. Frontiers in Physics, 7, 10.

Yan, J., Mizianty, M. J., Filipow, P. L., Uversky, V. N., & Kurgan, L. (2013). RAPID: fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1834(8), 1671-1680.

ACKNOWLEDGMENTS

Dr. Judith Yanowitz and the members of the Yanowitz Lab, Dr. Miguel Brieño-Enriquez and the members of the Brieño-Enriquez Lab. Dr. Mike Buszczak. Funded by NIGMS (R01GM127569).

CONTACT

Email: wmm27 at pitt.edu

Twitter: @mcfaddenwmj