LARKSdb: a tool for proteome-wide identification of functional low-complexity domains of proteins

Michael Hughes (mphughes@ucla.edu)[1]* and **David Eisenberg**[1]

[1]Department of Biological Chemistry and Department of Chemistry and Biochemistry, University of California Los Angeles (UCLA), Howard Hughes Medical Institute (HHMI), UCLA-Department of Energy (DOE) Institute for Genomics and Proteomics, Los Angeles, CA 90095, USA

Project Goals: We have previously identified structures, LARKS, that correlate with and explain labile amyloid associated with protein hydrogels and membraneless organelles. LARKS promote formation of reversible amyloid-like fibrils. Here we present LARKSdb, an online tool to predict LARKS forming segments and domains in proteins, and present preliminary results of LARKS predictions in non-human proteomes.

Membraneless organelles are assemblies of nucleic acids and proteins. Examples in eukaryotic cells include P-bodies, the nucleolus, nuclear gems, and stress granules. All of these are dynamic, rapidly rearranging bodies that perform functions for the cell. These assemblies are self-organizing in the absence of an enveloping membrane. How are such membraneless organelles organized in the absence of a physical barrier?

Low-complexity domains (LC) of proteins are important in organizing membraneless organelles (Kato *et al*., 2012). LC domains are regions of protein sequences with a biased composition of amino acids. In humans, LC domains often over represent glycine, serine, and alanine. *In vitro*, these domains phase separate and form labile hydrogels. These phase transitions mirror some of the behavior of membraneless organelles, so we sought structure of proteins that form labile hydrogels to better understand organization of membraneless organelles.

Closer inspection of hydrogels formed by the LC domains of FUS, hnRNPA1, and TDP43 revealed that they form protein fibrils that give diffraction indicative of amyloid fibrils (Kato *et al*., 2012). This was unexpected because pathogenic amyloid is typically irreversible, unlike the behavior of the hydrogels and membraneless organelles. We found that adhesive segments of LC domains form β-rich fibrils as in typical amyloid, but that the protein segments have sharp kinks in their peptide backbones. These kinks prevent extensive interfaces forming between mated β-sheets (Hughes *et al*., 2018). A consequence of a smaller interface is weakened interactions allowing for labile amyloid-like fibrils (Guenther *et al*., 2018). Reflecting their novel structure, we named them Low-complexity, Amyloid-like Reversible Kinked Segments (LARKS).

In previous work we computationally predicted the location of LARKS throughout the human proteome and found them to be enriched in proteins that form membraneless organelles (Hughes *et al*., 2018). We are launching an online server and database called LARKSdb enabling

researchers to predict LARKS within their proteins of interest. In humans we find that LARKS are enriched in LC domains. We present the following findings in our preliminary search for LARKS across kingdoms of life:

1) Organisms have variable LC coding regions in their proteomes; E. coli has 1.7% of proteins with LC domains, the corresponding values are 8.8% for yeast, 10.0% for tuberculosis, 42.0% for malaria, 26.0% for tetrahymena, and 15.9% for humans.

2) The number of LARKS rich proteins does not correlate with abundance of LCRs:

	E. coli	Yeast	Tuberculosis	Human
# proteins	4309	6049	3983	20396
# LC domain proteins	65	484	398	3243
# proteins > 10% LARKS	0	3	56	73

3) Functions of LARKS rich proteins varies by organism. In humans most LARKS-rich proteins are intracellular nucleic acid binding proteins and keratins (Hughes *et al.*, 2018). In tuberculosis the proteins are extracellular, possibly to interact with intracellular proteins from host eukaryotes. Nature seems to have co-opted LCR proteins and LARKS rich proteins to different functions.

LARKSdb enables scientists to identify LARKS in proteins of interest as an aid in understanding protein function. Therefore we believe that LARKSdb will be an important tool moving forward for scientists to interrogate the role of proteins in membranless organelles throughout the kingdoms of life.

References

E.L. Guenther, Q. Cao, H. Trinh, J. Lu, M.R. Sawaya, D. Cascio, D.R. Boyer, J.A. Rodriguez, M.P. Hughes, D.S. Eisenberg. Atomic structures of TDP-43 LCD segments and insights into reversible or pathogenic aggregation. Nat. Struct. Mol. Biol. (6) (2018), pp. 463-47.

M. Kato, T. W. Han, S. Xie, K. Shi, X. Du, L. C. Wu, H. Mirzaei, E. J. Goldsmith, J. Longgood, J. Pei, N. V. Grishin, D. E. Frantz, J. W. Schneider, S. Chen, L. Li, M. R. Sawaya, D. Eisenberg, R. Tycko, S. L. McKnight, Cell-free formation of RNA granules: Low complexity sequence domains form dynamic fibers within hydrogels. Cell 149, 753–767 (2012).

M.P. Hughes, M.R. Sawaya, D.R. Boyer, L. Goldschmidt, J.A. Rodriguez, D. Cascio, L. Chong, T. Gonen, D.S. Eisenberg. Atomic structures of low-complexity protein segments reveal kinked β sheets that assemble networks. Science, 359 (2018), pp. 698-701.

This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under contract no. DE-AC02-06CH11357. We thank NSF MCB-1616265, NIH AG-054022, DOE, and HHMI for support