IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 2, February 2022

DOI: 10.17148/IARJSET.2022.9271

The Disorder percentage of Late Embryogenesis Abundant Proteins and Transcription factors is higher than that of other proteins

Ayush Gowda D A[1], I Md Matheen[2], Pragathi Gowda C[3], Tejaswini N[4], Vaishali H J[5], Manohar G M[6], Nagamani T S[7*]

[1-7]Department of PG studies in Biotechnology, Nrupathunga University, Nrupathunga road, Bangalore, Karnataka

[7]halliLabs, Ragihalli, Anekal Tq, Bangalore , Karnataka

Abstract: Disordered Proteins are active proteins, some part of which may not have a specific three-dimensional structure. They are known to adopt a specific structure in conjunction with its binding partner. They lack hydrophobic amino acids that help in the process of protein folding. Late Embryogenesis Abundant proteins help in protecting higher plants from damages caused by environmental stresses. We analyzed randomly selected LEA proteins and transcription factors of plants and animals for various parameters of disorderness and compared them with proteins of *Arabidopsis thaliana* and *Caenorhabditis elegans*. The disordered percentage in each of these proteins was calculated using PONDR. Their amino acid composition, hydropathicity values were obtained by Protparam. The Ramachandran plot was used to obtain favorable and unfavorable regions. Some of the structures which were not available in structural databases were modeled using Alpha Fold 2. Our results show that the disordered percentages in LEA proteins transcription factors are statistically higher compared to normal proteins in *Arabidopsis thaliana* and *Caenorabhditis elegans*. We speculate that the necessity of LEA proteins to function in stressful developmental conditions and the necessity of the transcription factors to bind to multiple partners (both DNA elements and other factors) may be the reason why these groups of proteins exhibit such high percentages of disorderness.

Keywords: *Disordered proteins, Late Embryogenesis Abundant protein, Transcription factor, Hydropathicity, Alpha fold 2.*

1. INTRODUCTION:

Disordered Proteins are active proteins which lack rigid three dimensional structure in the absence of macromolecular interaction and they lack hydrophobic amino acids which help them in the folding process [1]. After attaching to other macromolecules, many disordered proteins can acquire a stable three-dimensional structure. In general, disorder proteins differ from structural proteins in a number of areas, including function, structure, sequence, and interactions [2].

During late embryogenesis, the Late Embryogenesis Abundant proteins are expressed in roots, stems, and other organs throughout the plant's growth phase [3]. Late Embryogenesis Abundant proteins are a type of highly hydrophilic, glycine-rich protein that have antioxidant, metal ion binding properties. It is also involved in membrane and protein stabilization, hydration buffering, and DNA and RNA interactions [3]. They are required for abiotic stress protection as well as proper plant growth and development. Abiotic stressors like cold, drought, or high salinity usually promote LEA expression. LEA proteins have been found in a wide variety of organisms like prokaryotes, invertebrates, and plants ranging from algae to angiosperms [3]. Transcription factors are proteins required for the expression of a gene [4]. They bind to other factors and specific DNA binding sites [5].

Because usually proteins that have functional roles under high abiotic stresses like the LEA have a propensity for disorderness and also that transcription factors in virtue of binding to multiple protein and DNA partners, we hypothesized that these proteins inherently may have a higher percentage of disorderness compared to random normal proteins. PONDR was used to calculate the disordered percentages in each of the types of proteins such as disorder proteins LEA and Transcription factors and other proteins of *Arabidopsis thaliana* [6] and *Caenorhabditis elegans* [7]. Using Protparam we determined their amino acid composition and hydropathicity values. The Ramachandran plot was used to determine which regions of the protein structure were highly preferred and which were in the questionable region. Alpha Fold 2 was used to model several structures that were not available in structural databases. Our findings reveal that the disordered percentages in LEA proteins, transcription factors are statistically higher than normal proteins. Our finding supports a speculation that the requirement for LEA proteins to function in stressful conditions under which maintaining an established three dimensional structure is challenging. Also the requirement of transcription factors to bind to numerous protein partners and regulatory DNA sequences, may be the basis for the high levels of disorder in these protein families.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 2, February 2022

DOI: 10.17148/IARJSET.2022.9271

2. MATERIALS AND METHODS:

2.1 Protein data: LEA protein and transcription factor protein structures were retrieved from PDB [8] (https://www.rcsb.org/) and Fasta sequences were retrieved from UniprotKB [9] and PDB Database (Table S1,S2,S3). List of other proteins of *Arabidopsis thaliana* and *Caenorhabditis elegans* were obtained from UniprotKB (Table S4,S5). (https://www.uniprot.org/)
2.2 PONDR : Disorder percentage of the proteins based on multiple predictors (VLXT,XL1_XT,CAN_XT,VL3,VSL2) were obtained from PONDR (Predictor of Natural Disordered Regions) [10]. (http://pondr.com/)
2.3 PROTPARAM : Disordered regions of proteins, hydropathicity GRAVY values and amino acid composition were obtained from protparam [11]. (https://web.expasy.org/protparam/)
2.4 RAMACHANDRAN PLOT: Ramchandran plot was plotted to identify the fully allowed part (favored) where the beta sheets and the right handed alpha helix occur, outer limit (allowed) with smaller van der Waals radii which corresponds to the left-handed alpha-helix and disallowed part (these residues are glycine can also be asparagine or aspartate) [12][13]. (https://zlab.umassmed.edu/bu/rama/)
2.5 Alpha Fold 2: Using the Fasta sequence of LEA proteins and transcription factor protein structures were modeled using Alpha Fold 2 [14] whose structures were not available in the PDB database. 12 proteins of random organisms were modeled with Alpha Fold 2 (list of the proteins modeled using Alphafold2 is given in Table S6) (colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb#scrollTo=33g5IIegij5R)
2.6 Statistical analysis : Kruskal-Wallis statistical test was done on the hydropathicity GRAVY values obtained from Protparam and also for highly preferred, preferred and questionable regions obtained from Ramachandran plot. All data was examined for homogeneity of variance and normality (transformation), as well as non parametric analysis. To examine the significant differences between the groups, the Kruskal Wallis test [15] was used. $P > 0.05$ was considered significant for all tests. SPSS software version 20 for Windows was used to analyze the data (IBM Corp).

3. RESULTS:

Late embryogenesis-abundant (LEA) proteins play important role in normal plant growth and development and also in abiotic stress protection.We calculated the disorder percentage of the 25 LEA proteins from plants and 5 from animals, Transcription factors from 10 different organisms and also other proteins which were not known as disordered proteins, 25 each of *Arabidopsis thaliana* and *Caenorhabditis elegans* through PONDR (Table S7-S11). The disordered percentages in LEA proteins in plants (orange) and in animals (gray) ,transcription factors (blue) are higher compared to other proteins in *Arabidopsis thaliana* (yellow) and *Caenorhabditis elegans* (green) **(Fig 1).**

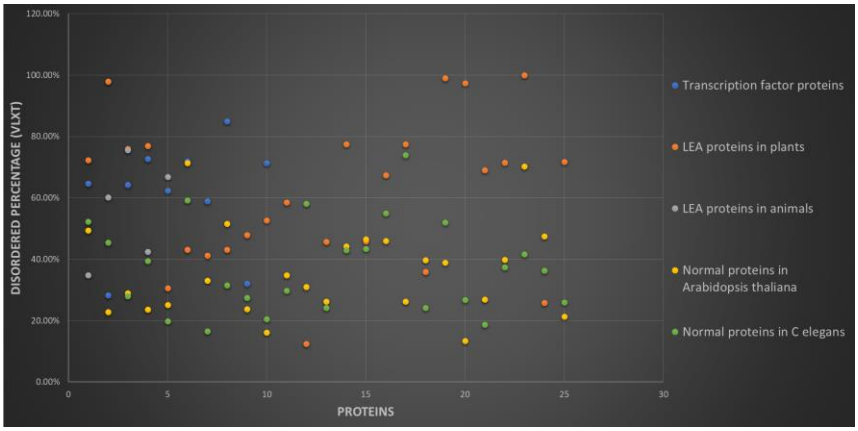


Fig. 1. The disordered percentages in LEA proteins in plants (orange) and in animals (gray) ,transcription factors (blue) are higher compared to other proteins in *Arabidopsis thaliana* (yellow) and *Caenorhabditis elegans* (green)

A negative hydropathicity (GRAVY) values from protparam of LEA proteins, transcription factors indicate that they are mostly hydrophilic proteins (a positive free energy change of the surrounding solvent indicates hydrophobicity, whereas a negative free energy change implies hydrophilicity) **(Fig 2.1)**. The PROTPARAM results show that the hydropathicity values in LEA proteins ,transcription factors are statistically higher (P value 0.013) compared to other proteins in *Arabidopsis thaliana* and *Caenorhabditis elegans* **(Fig 2.2)**.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 2, February 2022

DOI: 10.17148/IARJSET.2022.9271

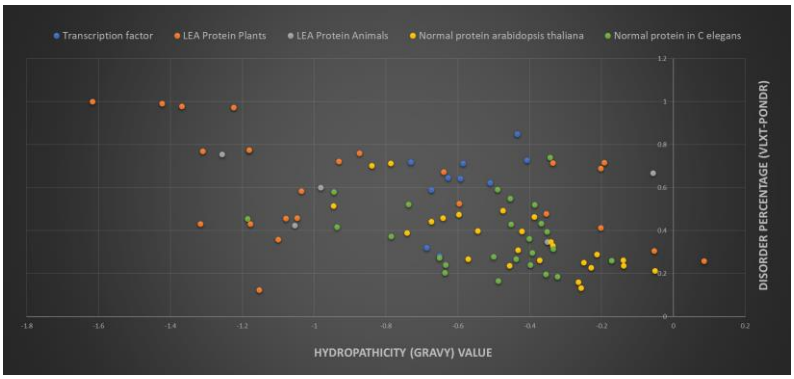
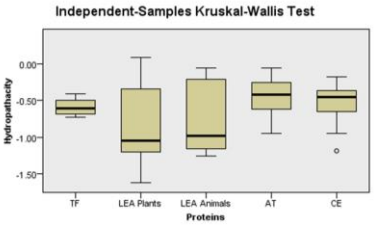


Fig. 2.1 LEA proteins have more negative value of hydropathicity than the other proteins



Total N	90
Test Statistic	12.636
Degrees of Freedom	4
Asymptotic Sig. (2-sided test)	.013

1. The test statistic is adjusted for ties.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
LEA Plants-LEA Animals	-6.620	12.798	-.517	.605	1.000
LEA Plants-TF	8.320	9.775	.851	.395	1.000
LEA Plants-CE	-18.000	7.389	-2.436	.015	.148
LEA Plants-AT	-24.580	7.389	-3.327	.001	.009
LEA Animals-TF	1.700	14.309	.119	.905	1.000
LEA Animals-CE	-11.380	12.798	-.889	.374	1.000
LEA Animals-AT	-17.960	12.798	-1.403	.161	1.000
TF-CE	-9.680	9.775	-.990	.322	1.000
TF-AT	-16.260	9.775	-1.663	.096	.962
CE-AT	6.580	7.389	.891	.373	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

TF - Transcription factors, AT - *Arabidopsis thaliana*, CE - *Caenorhabditis elegans*

Fig. 2.2 LEA proteins have more negative value of hydropathicity than the other proteins, the Kruskal Wallis test was performed on LEA proteins in plants and animals, and transcription factors hydropathicity values obtained from protparam

The secondary structures of proteins including the LEA proteins, transcription factors and representatives of the normal proteins using the pdb structures were analyzed using the Ramachandran plot as they lack three dimensional structures. Other proteins of *Arabidopsis thaliana* and *Caenorhabditis elegans*, have more *highly preferred regions*. (P value 0.049) The analysis of *preferred regions* for LEA and other proteins is around 5-20% (P value is 0.070). Also LEA proteins and transcription factors have more *questionable regions* (P value 0.010) than other proteins. **(Fig 3, 4 & 5)**.

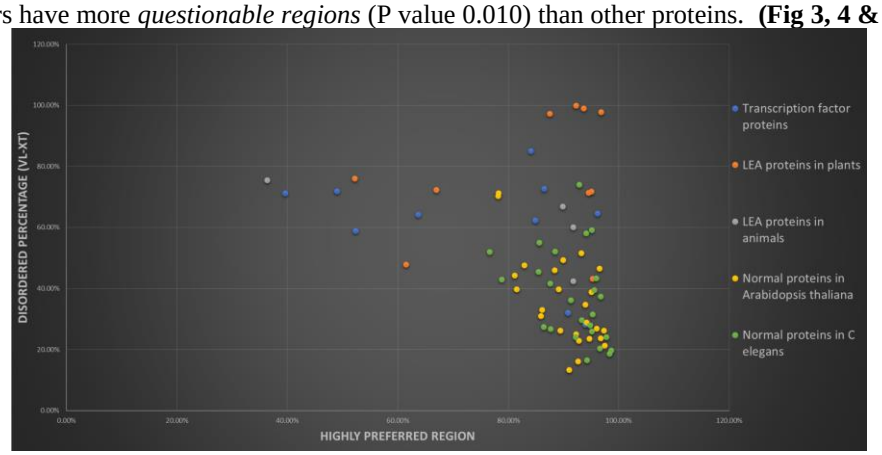


Fig. 3. Other proteins of *Arabidopsis thaliana* and *Caenorhabditis elegans* have more *highly preferred regions (*P value 0.049) than the LEA proteins and Transcription factors.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 7.105 ⁜ Vol. 9, Issue 2, February 2022

DOI: 10.17148/IARJSET.2022.9271

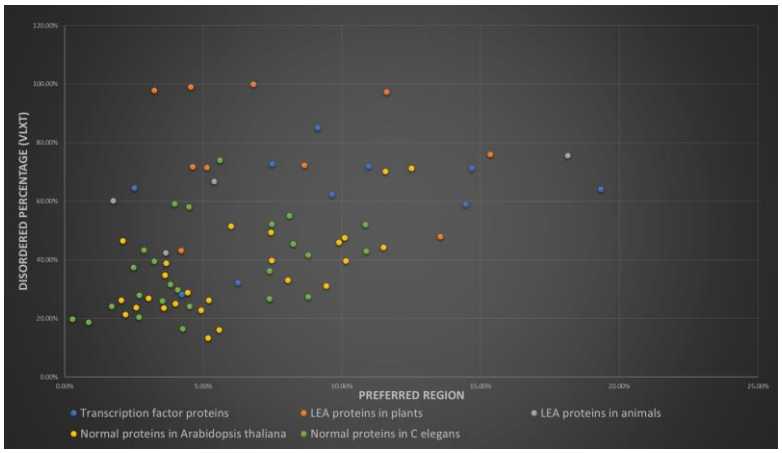


Fig. 4. LEA proteins and Transcription factors have more *preferred regions* (P value is 0.070) than the other proteins of *Arabidopsis thaliana* and *Caenorhabditis elegans* and the LEA proteins and Transcription factors values are scattered.

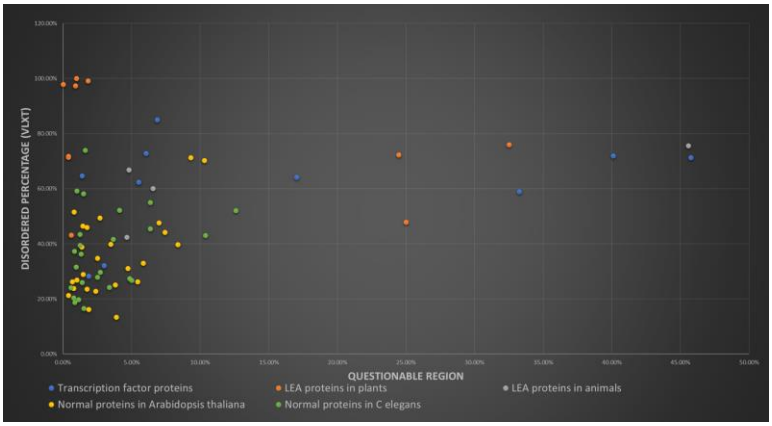


Fig. 5. LEA proteins and Transcription factors have more *questionable regions* (P value 0.049) than the other proteins of *Arabidopsis thaliana* and *Caenorhabditis elegans.* Order is entropically unfavorable for glycine since it has no side chain

The Kruskal Wallis test was used for hydropathicity values of LEA proteins in plants and animals, as well as transcription factor's values acquired from protparam (P value 0.013).
The test on highly preferred protein regions, the beta sheets and right-handed alpha helix are the highly preferred regions (P value 0.049) And questionable regions which have steric barriers between the side chain C-beta methylene group and the main chain atoms, also have low P value (0.010) whereas the significant P value is 0.05. The P value is 0.070 for the preferred region, that correlates to the left-handed alpha helix, is greater than the significant P value of 0.05.

DISCUSSION:

Due to the intrinsically unstructured nature of LEA Proteins they appear to be an incredibly versatile protein family. When exposed to various stress factors, including as drought, high-salinity stress, low-temperature stress, heavy-metal stress, and biotic stresses, they exhibit a variety of functions (e.g., chaperone, cryoprotective, antifreeze, ion-binding functions).
To investigate the disorderness in the structures of LEA proteins we analyzed different parameters associated with disorderness, LEA, in normal proteins of *Arabidopsis thaliana* and *Caenorhabditis elegans* and also transcription factors all of which were randomly chosen. Through PONDR (VL–XT) the disorder percentage of the proteins were calculated and LEA proteins and transcription factors were found to have a statistically higher percentage of disorderness compared to other proteins of *Arabidopsis thaliana* and *Caenorhabditis elegans*. And hydropathicity values of LEA proteins are more negative indicating that they are hydrophilic as they lack hydrophobic amino acids. A Ramchandran plot was used to determine the highly preferred region, which are the beta sheets and right-handed alpha helix , the preferred protein regions which correspond to the left-handed alpha helix. The steric barrier between the side chain C-beta methylene group and the main chain atoms is the most common reason for questionable regions. Because glycine

DOI: 10.17148/IARJSET.2022.9271

lacks a side chain, it can form phi and psi angles in all four quadrants of the Ramachandran plot. LEA proteins and transcription factors have more questionable regions where these regions are unstructured, as they are disordered proteins and lack three dimensional structure. The requirement of LEA proteins are necessary for plant growth and development as well as protection from abiotic stress, as well as the requirement of transcription factors to bind to numerous partners (both DNA elements and other factors), mediators, and regulatory areas in addition to regulating many genes, may be the basis for the high levels of disorder in these protein families.

ACKNOWLEDGEMENT:

We deeply appreciate the help and mentoring from **Srikanth L**, Research Scholar from University of Bristol, U K for regular inputs for this research**. Dr. Hariprasad TPN**, Assistant Professor, Bangalore University kindly helped us with the statistical analysis.

REFERENCES:

[1] van der Lee, Robin et al. “Classification of intrinsically disordered regions and proteins.” *Chemical reviews* vol. 114,13 (2014): 6589-631. doi:10.1021/cr400525m

[2] van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, Kim PM, Kriwacki RW, Oldfield CJ, Pappu RV, Tompa P, Uversky VN, Wright PE, Babu MM (2014). "Classification of intrinsically disordered regions and proteins". *Chemical Reviews*. **114** (13): 6589–631. doi:10.1021/cr400525m. PMC 4095912. PMID 24773235.

[3] Chen, Yongkun et al. “The Role of the Late Embryogenesis-Abundant (LEA) Protein Family in Development and the Abiotic Stress Response: A Comprehensive Expression Analysis of Potato (*Solanum Tuberosum*).” *Genes* vol. 10,2 148. 15 Feb. 2019, doi:10.3390/genes10020148

[4] Cooper, John A.. "transcription factor". Encyclopedia Britannica, 11 Sep. 2018, Accessed 20 February 2022.

[5] Latchman DS (December 1997). "Transcription factors: an overview". The International Journal of Biochemistry & Cell Biology. 29 (12): 1305–12. doi:10.1016/S1357-2725(97)00085-X. PMC 2002184. PMID 9570129

[6] Clay K.Clay K. & Putten, W. H. V. d. (1999) in Life History Evolution in Plants, eds. Vuorisalo, T. & Mutikainen, P. (Kluwer, Dordrecht, The Netherlands), pp. 275–301.Google

[7] Calahorro, Fernando, and Manuel Ruiz-Rubio. “Caenorhabditis elegans as an experimental tool for the study of complex neurological diseases: Parkinson's disease, Alzheimer's disease and autism spectrum disorder.” Invertebrate neuroscience : IN vol. 11,2 (2011): 73-83. doi:10.1007/s10158-011-0126-1

[8] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242.

[9] Morgat A, Lombardot T, Coudert E, Axelsen K, Neto TB, Gehant S, Bansal P, Bolleman J, Gasteiger E, de Castro E, Baratin D, Pozzato M, Xenarios I, Poux S, Redaschi N, Bridge A, UniProt Consortium. Enzyme annotation in UniProtKB using Rhea Bioinformatics 36(6):1896-1901 (2019)

[10] Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A.;*Protein Identification and Analysis Tools on the ExPASy Server;* (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press (2005).pp. 571-607

[11] Amara, I. , Zaidi, I. , Masmoudi, K. , Ludevid, M. , Pagès, M. , Goday, A. and Brini, F. (2014) Insights into Late Embryogenesis Abundant (LEA) Proteins in Plants: From Structure to the Functions. American Journal of Plant Sciences, 5, 3440-3455. doi: 10.4236/ajps.2014.522360.

[12] SC Lovell et al. (2003). Structure Validation by Cα Geometry: ϕ, ψ and Cβ Deviation. Proteins 50, 437-450. (PMID: 12557186, website)

[13] BK Ho and R Brasseur (2005). The Ramachandran plots of glycine and pre-proline. BMC Structural Biology 5(14) (2005-09-09)

[14] Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

[15] Kruskal; Wallis (1952). "Use of ranks in one-criterion variance analysis". Journal of the American Statistical Association. 47 (260): 583–621. doi:10.1080/01621459.1952.10483441

SUPPLEMENTARY SECTION:

https://docs.google.com/document/d/1F1MwEhDwlIn1kiTd7aK5zxngu4KRxYUyQLDxEXH13hQ/edit?usp=sharing

