10-3-2017

Potential functions of LEA proteins from the brine shrimp Artemia franciscana - Anhydrobiosis meets bioinformatics.

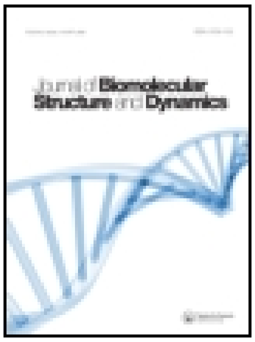
Brett Janis
University of Louisville

Vladimir N. Uversky
University of South Florida

Michael Menze
University of Louisville, michael.menze@louisville.edu

Original Publication Information

Janis, Brett, Vladimir N. Uversky and Michael A. Menze. "Potential Functions of LEA Proteins from the Brine Shrimp Artemia Franciscana - Anhydrobiosis Meets Bioinformatics." 2017. *Journal of Biomolecular Structure and Dynamics*, DOI: 10.1080/07391102.2017.1387177



Taylor & Francis
Taylor & Francis Group

Journal of Biomolecular Structure and Dynamics

ISSN: 0739-1102 (Print) 1538-0254 (Online) Journal homepage: http://www.tandfonline.com/loi/tbsd20

Potential Functions of LEA Proteins from the Brine Shrimp Artemia Franciscana - Anhydrobiosis Meets Bioinformatics

Brett Janis, Vladimir N. Uversky & Michael A. Menze

To cite this article: Brett Janis, Vladimir N. Uversky & Michael A. Menze (2017): Potential Functions of LEA Proteins from the Brine Shrimp Artemia Franciscana - Anhydrobiosis Meets Bioinformatics , Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2017.1387177

To link to this article: http://dx.doi.org/10.1080/07391102.2017.1387177

Accepted author version posted online: 03 Oct 2017.

Publisher: Taylor & Francis

Journal: *Journal of Biomolecular Structure and Dynamics*

DOI: http://doi.org/10.1080/07391102.2017.1387177

Potential Functions of LEA Proteins from the Brine Shrimp *ArtemiaFranciscana* - Anhydrobiosis Meets Bioinformatics

Brett Janis[1], Vladimir N. Uversky[2,3], and Michael A. Menze[1]

[1] Department of Biology, University of Louisville, Louisville, KY 40292

[2] Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, Florida 33612, USA

[3] Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Moscow Region, 142290, Russia

Address for reprint requests and other correspondence: Tel.: (502) 852-8962, Fax: 502-852-0725, Email: Brett.Janis@louisville.edu, vuversky@health.usf.edu, Michael.menze@louisville.edu

Abstract

Late embryogenesis abundant (LEA) proteins are a large group of anhydrobiosis-associated intrinsically disordered proteins (IDP), which are commonly found in plants and some animals. The brine shrimp *Artemiafranciscana* is the only known animal that expresses LEA proteins from three, and not only one, different groups in its anhydrobiotic life stage. The reason for the higher complexity in the *A. franciscana* LEA proteome (LEAome), compared with other anhydrobiotic animals, remains mostly unknown. To address this issue, we have employed a suite of bioinformatics tools to evaluate the disorder status of the *Artemia*LEAome and to analyze the roles of intrinsic disorder in functioning of brine shrimp LEA proteins. We show here that *A. franciscana*LEA proteins from different groups are more similar to each other than one originally expected, while functional differences among members of group 3 are possibly larger than commonly anticipated. Our data show that although these proteins are characterized by a large variety of forms and possible functions, as a general strategy, *A. franciscana* utilizes glassy matrix forming LEAs concurrently with proteins that more readily interact with binding partners. It is likely that the function(s) of both types, the matrix-forming and partner-binding LEA proteins, are regulated by changing water availability during desiccation.

Keywords

Water Stress, Desiccation, Anhydrobiosis, State Predictions, Phase Transitions

Abbreviations

LEA protein – Late Embryogenesis Abundant Protein

IDP – Intrinsically Disordered Protein

PONDR – Predictor of Native Disorder

CH Plot – Charge/Hydropathy Plot

CDF – Cumulative Distribution Function

MeDor – Metaserver of Disorder

HCA – Hydrophobic Cluster Analysis

Background

Late embryogenesis abundant (LEA) proteins constitute a large group of intrinsically disordered proteins (IDP)associated with anhydrobiosis, or 'life without water' (Hincha and Thalhammer, 2012; Kovacs, Agoston, & Tompa, 2008; Tompa and Kovacs, 2010). LEA proteins have been shown to improve desiccation tolerance in anhydrobiotic organisms (Gal, Glazer, & Koltai, 2004; Yu, Lai, Wu, Wu, & Guo, 2016) and in desiccation sensitive cell lines that ectopically express them (Marunde et al., 2013). Given the nature of anhydrobiosis, proteins that improve desiccation tolerance are difficult to characterize, because they likely remain mostly inactive in the fully hydrated state. Elucidation of LEA function(s) in the dried state represents another challenge since it excludes a variety of biochemical techniques commonly used to study proteins in solution. As a result, the functional structure of LEA proteins and their mechanisms of conferring desiccation tolerance have proven difficult to understand (Battaglia, Olvera-Carrillo, Garciarrubio, Campos, & Covarrubias, 2008; Hand, Menze, Toner, Boswell, & Moore, 2011; M.-D. Shih, Hoekstra, & Hsing, 2008; A. Tunnacliffe and Wise, 2007; Wise and Tunnacliffe, 2004). Due to the challenges in directly or indirectly observing LEA protein structure and function at distinct water levels, several hypotheses for their mechanism(s) of functionality have been presented, including molecular shielding (A. Tunnacliffe and Wise, 2007)(Chakrabortee et al., 2012), membrane stabilization (Steponkus, Uemura, Joseph, Gilmour, & Thomashow, 1998)(Tolleter et al., 2007)(Bremer, Wolff, Thalhammer, & Hincha, 2017; Moore, Hansen, & Hand, 2016; Thalhammer, Hundertmark, Popova, Seckler, & Hincha, 2010; Tolleter, Hincha, & Macherel, 2010) sequestration of divalent ions (Garay-Arroyo, Colmenero-Flores, Garciarrubio, & Covarrubias, 2000), increasing the glass transition temperature of sugar glasses (Shimizu et al., 2010) protection of proteins by prevention of protein aggregation (Goyal, Walton, & Tunnacliffe, 2005; Grelet et al., 2005; Popova, Rausch, Hundertmark, Gibon, & Hincha, 2015), and acting as hydration buffers (Mouillon, Gustafsson, & Harryson, 2006). Furthermore, functions of a given LEA protein may change with changes in hydration levels.

LEA proteins were first discovered in cotton seeds (L. Dure and Chlan, 1981; L. Dure and Galau, 1981; Leon Dure, Greenway, & Galau, 1981) and later were also found in seeds and vegetative tissues of several other plants (for review see (Hoekstra, Golovina, & Buitink, 2001; M.-D. Shih, et al., 2008)(Battaglia and Covarrubias, 2013; Graether and Boddington, 2014)(Pammenter, 1999)and, more recently, in some anhydrobiotic animals, such as nematodes

(Solomon, Salomon, Paperna, & Glazer, 2000)(Browne, Tunnacliffe, & Burnell, 2002), rotifers (Denekamp, Reinhardt, Kube, & Lubzens, 2010; Alan Tunnacliffe, Lapinski, & McGee, 2005), tardigrades (Schokraie et al., 2010), springtails (Clark et al., 2007), the chironomid*Polypedilumvanderplanki*(Kikawada et al., 2006), and the brine shrimp *Artemiafranciscana*(Hand, Jones, Menze, & Witt, 2007). LEA proteins have been proven to be difficult to conceptually organize, resulting in several different classification schemes that propose 6 to 12 different protein families (for overview see: (Hunault and Jaspard, 2010; Jaspard, Macherel, & Hunault, 2012)). Despite ongoing efforts to categorize LEA proteins into different functional groups, no classification method has been universally accepted. This lack of consensus may further illustrate the complex nature of these proteins and may resembles challenges associated with characterizing and classifying IDPs in general.

Depending on the amount of residual structure found in them, intrinsically disordered proteins (IDPs), at the whole molecule/domain level, can be organized into three distinct classes, such as native coils, native pre-molten globules, and native molten globules (V. N. Uversky, 2002; V. N. Uversky and Dunker, 2010). This categorization of whole molecule IDPs is based on their structural similarity to unfolded and different partially folded conformations detected for several globular proteins under various denaturing conditions (Ptitsyn, 1995; Uverskii, 1998; V.N. Uversky, 1997; V. N. Uversky and Ptitsyn, 1994, 1996). Therefore, it seems that structurally, functional proteins can be classified as intrinsically disordered (coils, pre-molten globules, molten globules), and ordered (globular). In reality, this picture is more complex, since different parts of a protein can be differently disordered, thereby forming a protein structure continuum (V. N. Uversky, 2013a, 2013b, 2016). Obviously, of these structural subtypes, only globular proteins are considered as ordered in the classic sense, typically serving as illustrations of the standard ‘lock and key’ model of the protein structure-function paradigm. Note that transmembrane and structural, e.g., fibrillar, proteins are intentionally excluded from this consideration.

Coil-like polypeptide chains, or almost entirely disordered proteins, can have a hydrodynamic radius dramatically exceeding that of ‘classic’ globular proteins (Uverskii, 1998; V. N. Uversky, 2002; V. N. Uversky and Dunker, 2010). The large hydrodynamic volume and the highly accessible structure of extended IDPs makes them especially susceptible to degradation. Importantly, highly disordered polypeptides are frequently found as spacers and linkers between functional domains in globular proteins might have additional functions. For example, ligand-binding elements linked together by random coils increase binding affinity through the chelate effect (Jencks, 1981). Extended IDPs and IDP regions (IDPRs) are highly susceptible to post-translational modifications, for example, containing up to 10 times as many phosphorylation sites as globular proteins (Xie et al., 2007). Pre-molten globular proteins (both IDPs or partially folded intermediates of globular proteins) contain of significant levels of secondary structure, but exhibit no globular tertiary structure and occupy about twice the volume of the molten globular proteins. Molten globular proteins are characterized by compact, globular conformations that contain high levels of defined secondary structure, but display limited tertiary features (V. N. Uversky, 2002; V. N. Uversky and Dunker, 2010).

IDPs/IDPRs display a variety of functions and functional mechanisms. They can show activity in their disordered state, often acting as chaperones, entropic chains, and recognition regions for interactions with a variety of partner molecules (Dunker and Obradovic, 2001).

Alternatively, IDPs/IDPRs can undergo disorder-to-order transitions, when their environment changes, such as during desiccation (Goyal et al., 2003; M. D. Shih, Hsieh, Lin, Hsing, & Hoekstra, 2010), or in response to recognition of binding partners (Dunker et al., 1998). In comparison with ordered proteins and domains, IDPs/IDPRs hold a variety of functional benefits (P. Lieutaud et al., 2016), such as conformational plasticity (Wright and Dyson, 1999), one-to-many and many-to-one signaling mechanisms (Romero et al., 1998), binding-site plasticity (Meador, Means, & Quiocho, 1992), thermodynamic regulation (Spolar and Record, 1994), and reduced cellular lifespan for transient expression patterns (Wright and Dyson, 1999). In the case of environmental conditions causing a conformational transition in the polypeptide chain, random coil-like regions tend to undergo disorder-to-order transitions more readily than pre-molten globular regions (Mészáros, Simon, & Dosztányi, 2009). These state transitions can occur due to target binding, changes in the chemical environment, or activation by post-translational modification, and several useful bioinformatics tools were developed to investigate potential biological functions of polypeptide chains with low structural complexity (Cheng et al., 2007; Disfani et al., 2012; Oldfield, Cheng, Cortese, Romero, et al., 2005; Z. Peng, Wang, Uversky, & Kurgan, 2017). In the study presented here, a variety of open access bioinformatics tools were employed to gain insights into the intrinsic disorder and potential function(s) of LEA proteins from the brine shrimp *Artemiafranciscana*.

While several biochemical methods can be applied to characterize IDPs (for reviews see Methods in Molecular Biology volumes 895 (Vladimir N. Uversky and Dunker, 2012a) and 896 (Vladimir N. Uversky and Dunker, 2012b)), great strides have been made in developing bioinformatics tools to explain and/or predict potential structural and functional elements of

IDPs/IDPRs, to guide future research, and to assist in data interpretation (Bracken, Iakoucheva, Romero, & Dunker, 2004; Radivojac et al., 2007; V. N. Uversky, Radivojac, Iakoucheva, Obradovic, & Dunker, 2007). Many of these programs have a high accuracy in predicting IDPs and the localization of IDPRs. In general, IDPs have an amino acid composition biased towards residues that promote disorder such as alanine (Ala), glycine (Gly), aspartic acid (Asp), methionine (Met), lysine (Lys), arginine (Arg), serine (Ser), glutamic acid (Glu), and proline (Pro) (Dunker et al., 2008). Additionally, certain motifs of physicochemical characteristics are common in amino acids sequences of IDPs, such as Pos(itive)-Pos-X-Pos, Neg(ative)-Neg-Neg, Glu-Glu-Glu, Lys-X-X-Lys-X-Lys, and Pro-X-Pro-X-Pro (Lemke, 2011; Lise and Jones, 2005). The amino acid composition may also be associated with different “flavors” of disorder (Lemke, 2011; Vucetic, Brown, Dunker, & Obradovic, 2003) that have weak but statistically significant associations with protein function. Given the relatively low complexity of IDP structures, amino acid sequence data has been used in bioinformatics programs in order to predict IDP function with some success (Lobley, Swindells, Orengo, & Jones, 2007).

The brine shrimp *Artemiafranciscana* is the only known animal expressing three different groups of LEA proteins (1, 3, and 6; for alternative classifications please refer to Tab. 1) in the anhydrobiotic life stage (Hand and Menze, 2015; MacRae, 2016; Warner, Chakrabortee, Tunnacliffe, & Clegg, 2012; Wu et al., 2011). The reason for the higher complexity in the *Artemia*LEA proteome compared with other anhydrobiotic animals that only express group 3 LEA proteins is unknown. We hypothesized that distinct functional differences among the three LEA groups may exist and offer additive or synergistically advantages to the anhydrobiotic stage in brine shrimp. To test this hypothesis, we employed a wide spectrum of bioinformatics tools.

[Table 1 here]

Methods

Predictor of Naturally Disordered Regions (PONDR)

The Predictor of Naturally Disordered Regions (PONDR) is a web server for intrinsic disorder prediction based on the input amino acid sequence of a query protein. PONDR server utilizes a combination of computational tools including several feedforward neural networks (PONDR® VLXT (Li, Romero, Rani, Dunker, & Obradovic, 1999; Romero et al., 2001), PONDR® VSL2 (K. Peng, Radivojac, Vucetic, Dunker, & Obradovic, 2006), and PONDR® VL3 (Obradovic et al., 2003; Radivojac, Obradovic, Brown, & Dunker, 2003)), and two binary disorder predictors that evaluate the probability of a query protein to be disordered as whole, Charge-Hydropathy (CH-plot) analysis (V. N. Uversky, Gillespie, & Fink, 2000), and a Cumulative Distribution Function (CDF) analysis (Dunker, Obradovic, Romero, Garner, & Brown, 2000; Oldfield, Cheng, Cortese, Brown, et al., 2005). CDF analysis is based on one of the outputs of PONDR® VLXT and summarizes the per-residue predictions by plotting PONDR scores against their cumulative frequency, which allows ordered and disordered proteins to be distinguished based on the distribution of prediction scores (Dunker, et al., 2000; Oldfield, Cheng, Cortese, Brown, et al., 2005). PONDR is freely available at http://www.pondr.com/.

Metaserver of Disorder (MeDor)

Metaserver of Disorder (MeDor) is a freely available platform that predicts the structure of a protein based on the input amino acid sequence, provides a hydrophobic cluster analysis

(HCA) plot that projects the protein in α-helical orientation, submits the sequence to several protein disorder and localization prediction servers (such as MeDor submits the amino acid sequence to IUPRED (Dosztanyi, Csizmok, Tompa, & Simon, 2005), PreLINK(Coeytaux and Poupon, 2005), RONN (Yang, Thomson, McNeil, & Esnouf, 2005), FoldUunfold(Garbuzynskiy, Lobanov, & Galzitskaya, 2004), DisEMBL1.5 (REM 465, loops, and hotloops) (R. Linding et al., 2003), FoldIndex(Prilusky et al., 2005), Globplot2.3 (Rune Linding, Russell, Neduva, & Gibson, 2003), PONDR® VL3 (Obradovic, et al., 2003), PONDR® VL3H (Obradovic, et al., 2003), PONDR® VSL2B (Obradovic, Peng, Vucetic, Radivojac, & Dunker, 2005), and Phobius(Kall, Krogh, & Sonnhammer, 2004)), and juxtaposes the results of each for ease of analysis (Philippe Lieutaud, Canard, & Longhi, 2008). MeDor offers secondary structure prediction using the Secondary Structure Predictor (SSP) Pred2ary (Chandonia and Karplus, 1999). The HCA plot is a useful tool for visual detection of disordered and potential binding regions by highlighting hydrophobic clusters and representing the characteristics of secondary structures by coloring residues based on their chemical properties (Callebaut et al., 1997). MeDor is freely available for download at http://www.vazymolo.org/MeDor/index.html.

IUPred and ANCHOR

IUPred is a disorder prediction server that uses pairwise energies of potential interactions between amino acid to predict the likelihood of disorder (Dosztányi, Csizmok, Tompa, & Simon, 2005; Dosztányi, Csizmók, Tompa, & Simon, 2005). IUPred predicts disorder based on two reading lengths, long regions of 30 or more amino acids and short regions of 25 or fewer amino acids. IUPred is freely available at http://iupred.enzim.hu/.

ANCHOR is a Molecular Recognition Feature (MoRF) prediction server that uses similar pairwise energies as IUPred employs, but combines them with characteristics of known MoRF regions (Dosztányi, Mészáros, & Simon, 2009; Mészáros, et al., 2009). ANCHOR, while not a trained algorithm, was tested on various data sets and predicted protein binding MoRF sites with 70% accuracy and a false-positive rate of <5% in globular protein datasets (Dosztányi, et al., 2009). ANCHOR specifically identifies protein-binding MoRF regions. ANCHOR is freely available at http://anchor.enzim.hu/.

DisEMBL1.5

DisEMBL is a disorder prediction server that utilized three artificial neural networks for structural analysis, Loops/Coils, Hot Loops, and Remark-465 (Bourhis, Canard, & Longhi, 2007; R. Linding, et al., 2003). The Loops/Coils predictor is based on proteins from the Dictionary of Secondary Structure of Proteins (DSSP) and contains ~57% of disordered residues (Kabsch and Sander, 1983; R. Linding, et al., 2003). It accurately predicts only ~50% of ordered sequences, but regions known to be disordered are extremely rarely predicted to be ordered. The Hot Loops predictor utilizes B-factors from the X-ray crystallography structures (R. Linding, et al., 2003). It also was trained on DSSP proteins with disordered residues and includes proteins representing members of each protein family listed in the database. The Remark465 neural network is trained on the stretches of amino acids with missing electron density in X-ray crystallography structures (R. Linding, et al., 2003). Remark465 has a false positive rate of ~16%, likely because missing electron density is only partly due to protein disorder. DisEMBL1.5 is freely available at http://dis.embl.de/.

GlobPlot2.3

GlobPlot is a propensity-based server for prediction of structural disorder and globular domains (Rune Linding, et al., 2003). GlobPlot utilized the Remark465 propensities, and its output may be adjusted. The default output is a sloped graph, in which negative slopes represent propensity for ordered domains and positive slopes indicate disorder predictions. Using the SMART server, coiled-coil regions and low complexity regions are highlighted as striped boxes or empty boxes, respectively. Along the bottom of the graph, GlobPlot gives a color-coded predictor of regional structure, with no color indicating uncertainty or structural flexibility. GlobPlot2.3 and all propensity sets are freely available at http://globplot.embl.de/.

Heliquest

Heliquest projects amino acid sequences as α-helices, calculates the physicochemical properties of these α-helices, and plots two superimposed graphs of hydropathy and hydrophobic moment at each amino acid position (Gautier, Douguet, Antonny, & Drin, 2008). Corresponding projections and graphs are derived from a sliding window, which the user can select to range from 11 to 54 residues. For each projection, an accompanying table includes the number of charged, polar, and uncharged residues, as well as special residues such as proline and cysteine. The table also includes standard hydropathy(Fauchere and Pliska, 1983), hydrophobic moment (Eisenberg, Weiss, & Terwilliger, 1982), and net charge at a pH 7.4. Heliquest is freely available at http://heliquest.ipmc.cnrs.fr/.

DISPHOS 1.3

DISPHOS 1.3 is an online phosphorylation prediction server specialized in identifying phosphorylation sites in the context of protein disorder (Iakoucheva et al., 2004). To assess potential phosphorylation sites, DISPHOS 1.3 predicts the surface exposure, electrostatic charge, hydropathy, and flexibility of amino acids that neighbor serine, threonine, and tyrosine. DISPHOS 1.3 is trained on specific data sets in the SWISS-PROT database, such as Eukaryotes, or specific model organisms to reduce mischaracterizations (e.g. *Caenorhabditiselegans*, *Drosophila melanogaster*, *Homo sapiens*, etc.). For the purposes of the analysis presented here, we used the predictor trained on proteins from *D. melanogaster* since both *A. franciscana* and *D. melanogaster*are arthropods. DISPHOS 1.3 is freely available at: http://www.dabi.temple.edu/disphos/.

CIDER/localCIDER

CIDER is a server that returns sequence-specific parameters such as the length, distribution of opposite charges (κ), the Frequency of Charged Residues (FCR), the Net Charge Per Residue (NCPR), hydropathy according to the Kyte& Doolittle scale (Kyte and Doolittle, 1982), the proportion of disorder promoting residues, and plots the protein on a diagram of states for a prediction of the structural qualities of a query protein (Holehouse, Ahad, Das, & Pappu, 2015). The distribution of opposite charges, represented as κ, is scaled between 0 and 1, where 0 represents a perfectly even distribution of charges across the protein and 1 indicates complete separation of charges. This measure is useful for identifying self-repulsion or attraction, especially in the desiccated state for LEA proteins. LocalCIDER is a high-performance software package that offers a more advanced analysis of protein sequences, including plotting parameters, such as NCPR for example, with a defined window size. Several other parameters may also be calculated or modified, such as calculating poly-proline helix propensity and changing the hydropathy or complexity. CIDER and localCIDER are freely available at http://pappulab.wustl.edu/CIDER/analysis/.

Results and Discussion

Within the last 14years, LEA proteins have been found to accumulate in some desiccation tolerant animals including the brine shrimp *A. franciscana* (for review see (Hand and Menze, 2015; MacRae, 2016)). However, *A. franciscana*expresses multiple LEA proteins from three classification groups (group 1, 3, and 6) in the desiccation tolerant embryo, making it unique among anhydrobiotic animals (Hand and Menze, 2015; Sharon, Kozarova, Clegg, Vacratsis, & Warner, 2009; Wu, et al., 2011). The reason(s) for the presence of a larger variety of LEA groups in *A. franciscana*, compared to other anhydrobiotic animals, is unknown. It seems reasonable to assume that proteins from different LEA groups may offer distinct or additive benefits to the animal if group specific differences in protein functions exist. However, even in the absence of group-specific functional differences, a large variety of LEA proteins might be necessary to confer desiccation tolerance in anhydrobiotic animals. The reasons for concurrent expression of multiple LEA proteins may include targeting different types of macromolecules (lipids, nucleic acids, proteins) or different members of the same macromolecular type, to serve different molecular functions (ion chaperones, molecular shields, structural reinforcement), and/or are to localize to different subcellular compartments.

Group 1

AfLEA1.1

A large number of highly similar group 1 LEA proteins has been described in *A. franciscana*(Sharon, et al., 2009) and two of them, *Af*LEA1.1 and *Af*LEA1.3, are almost identical, except that *Af*LEA1.3 contains an N-terminal signal sequence and localized to the mitochondria, whereas *Af*LEA1.1 lacks a signal sequence and is retained in the cytoplasm (Marunde, et al., 2013; Toxopeus, Warner, & MacRae, 2014; Warner, Guo, Moshi, Hudson, & Kozarova, 2016; Warner et al., 2010). Mitochondrial signal sequences are usually cleaved after incorporation of the protein into the mitochondrial matrix (Neupert and Herrmann, 2007)(Roise and Schatz, 1988). Therefore, the cytoplasmic protein *Af*LEA1.1 will be analyzed below as an illustrative representative for the other *A. franciscana*group 1 LEA proteins that basically differ only in the numbers of a repeat of a 20-amino acid long sequence motif (Warner, et al., 2016).

The first group 1 LEA protein in *A. franciscana* was described by Sharon and colleagues as a heat stable and highly hydrophilic 21-kDa protein (Sharon, et al., 2009). This protein contains a characteristic 20-amino acid motif (GGQTRREQLGEEGYSQMGRK), and several protein variants including 2 to 8 repeats of this motif have been discovered (Sharon, et al., 2009; Warner, et al., 2016; Warner, et al., 2010). The mean net charge and low hydropathy shown in the CH-plot (Fig. 1A) place *Af*LEA1.1 in the category of proteins with extended disorder, which is not surprising given the particularly high percentage of charged and polar residues (52.8%) in this protein. This is also in agreement with the output of CDF analysis (see Fig. 1B) which further supports the notion of a highly-disordered nature of *Af*LEA1.1. In fact, it was established earlier that seven boundary points located in the 12th through 18th bin provided the optimal separation of the ordered and disordered protein sets in the CDF plots and that classification of a query protein as wholly ordered or wholly disordered is based on whether a corresponding CDF curve was above or below a majority of boundary points, respectively (Oldfield, Cheng, Cortese, Brown, et al., 2005).According to these criteria, *Af*LEA1.1 is expected to be disordered as a whole.

In the CH-plot, *Af*LEA1.1 is closest to the group 3 LEA protein *Afr*LEA3m (Menze, Boswell, Toner, & Hand, 2009) whose secondary structure, along with that of *Afr*LEA2, has been characterized using circular dichroism (Boswell, Menze, & Hand, 2014). *Af*LEA1.1 most closely resembles *Afr*LEA2 in terms of its proportion of charged residues, but *Af*LEA1.1 has greater separation of its charged residues (Tab. 2), although both *Af*LEA1.1 and *Afr*LEA2 are being classified as Janus sequences by CIDER (Fig. 1C). In the desiccated state, electrostatic interactions likely hold greater impact on folding dynamics than in the hydrated state.

[Table 2 here]

[Figure 1 here]

Therefore, lower absolute mean net charges combined with higher κ values may become particularly influential in predicting secondary and tertiary structure motives in the dry state. The distribution of positive and negative charges alternates repetitively due to the 20-amino acid sequence motif, creating several points for favorable electrostatic interactions within this center region of the protein (Fig. 2). This separation of charges along the sequence likely cause *Af*LEA1.1 to adopt electrostatically-driven structures in the dry state that could be influenced by the presence or absence of ions.

[Figure 2 here]

We combined disorder predictions derived from applying several different algorithms to understand potential structural features in the hydrated and desiccated states of *Af*LEA1.1. DisEMBLE predicts *Af*LEA1.1 to be overall disordered (63.3%), with different likelihoods for ordered or disordered states at distinct regions within the polypeptide chain (see Supplementary Materials, Fig. S1). Stretches of the protein where ordered structure is predicted by MeDOR-based DisEMBL (Fig. 3A), show a strong tendency for β-strands in the hydrated state, a structure not as commonly found in group 1 LEA proteins as α-helices (Battaglia, et al., 2008). However, experimental analysis is needed to confirm this prediction. Our knowledge of secondary structure of group 1 LEA proteins is limited to plants, where most group 1 LEA proteins have been shown to be highly disordered, or to contain up to 47% of α-helices (for review see: (M.-D. Shih, et al., 2008)). Surprisingly, most of the predicted α-helices in *Af*LEA1.1 fall into regions that are likely to be disordered, suggesting that α-helices can only be formed in response to interactions with a binding partner or during desiccation. The β-strands, however, appear to more likely occur in the hydrated protein, with the potential for increased folding in less polar solvents or during desiccation.

[Figure 3 here]

Fig. 4A shows that *Af*LEA1.1 is predicted to have several regions that possess an ambiguous propensity for ordered and disordered structure that coincide with the positions of MoRF regions predicted by ANCHOR (Fig. 4A). Although ANCHOR predicts MoRF regions spaced relatively evenly across the protein, four MoRF regions with a likelihood greater than 80% are localized in pairs at the protein termini (amino acid positions 1-20 and 39-53 in the N-terminal region, and regions 140-154 and 163-180 and the C-tail). These terminal MoRF regions have distinct amino acid sequences not found in other regions. Sudden dips in the Globplot slope (see Supplementary Materials, Fig. S2) are predicted to form β-strands at positions 39-53 and 140-154 that separate the terminal MoRF pairs from the internal regions. These predicted β-strands are characterized by a specific clustering of hydrophobic residues, high glycine content, and complementary charges of basic and acidic amino acids. The finding that the primary amino acid sequence of these two 15 amino acid long MoRF regions are distinct from the other regions, while sharing an almost identical 13 amino acid overlap, suggests that they are either separating functional segments or are involved in orienting them. The two terminal MoRF regions (residues 1-20, 163-180) have pronounced structural differences, which is shown by ANCHOR as well as by sudden drops in PONDR® VLXT profile (Fig. 4A). Furthermore, PONDR® VL3 predictor weakly indicates that the N- and C-terminal MoRF sites might exhibit unique structural elements.

DISPHOS predicts the N-terminal region of the *Af*LEA1.1 to be heavily phosphorylated if translated in *D. melanogaster* (see Supplementary Materials, Fig. S3). The N-terminal region is highly enriched in positively charged residues (30%), serine residues (30%) predicted by DisPHOS to be phosphorylated, and contains a cluster of hydrophobic residues (25%). All eight serine residues within the N-terminus are predicted to be phosphorylated, with seven phosphorylation sites being located within the first 26 amino acids of the protein. Residues 4, 5, and 6 are consecutive serine residues resembling an α-helix cap, which may promote α-helical stability during desiccation (Aurora and Rose, 1998). This high concentration of likely phosphorylation sites further distinguishes the two N-terminal MoRF regions from the other regions of the protein.

[Figure 5 here]

Considering that α-helices may be important to LEA protein structure and function, Heliquest algorithm was used to evaluate the properties of any α-helices that might be formed in the hydrated and/or desiccated states (Fig. 5A). Interestingly, an α-helix within the N-terminal MoRF region would have a high hydrophobic moment, due to a small but concentrated hydrophobic face. An α-helix within the MoRF region at the C-terminus, on the other hand, would have a very low hydrophobic moment due to a relatively even distribution of hydrophobic residues. This means that, if *Af*LEA1.1 would interact with phospholipid membranes, this may occur at the N-terminus, but not at the C-terminus. The penultimate MoRF regions both exhibit a hydrophobic face composed of the six amino acids sequence "AMGGY", although the hydrophobic face of the MoRF in the 140-154 region is extended to "LMGAMGGY". This suggests that, if α-helices were to form in these two regions, then the formed structure would be an amphipathic α-helix with substantial flexibility due to the 2 or 3 glycine residues in this structure. However, given the helix-breaking propensity of glycine residues, the odds of these structures forming are low.

The Heliquest-based predictions of an α-helical region with a continuous hydrophobic stripe can be visualized on the HCA (Fig. 3A). This band becomes most pronounced at the predicted internal MoRF regions and less pronounced at the termini of the protein, again suggesting different functional behaviors for the termini compared to the internal regions of the protein. Another noteworthy observation is that the SMART server describes *Af*LEA1.1 as a protein containing quadruple repeat of LEA_5 (PF00477) domains, which are found in hydrophilic plant seed proteins. Furthermore, close homologies of the PFAM LEA5 domain to

EM-like proteins were found using NCBI tblastp (e.g. e-value of 8e^{-60} to GEA1 from *Camelina sativa* XP_010503885).

Group 3

AfrLEAI

*Afr*LEAImaintains a ratio of hydropathy to mean net charge of 0.094 which is similar to the group 6 LEA protein *Afr*LEA6, but higher than those of the other Artemia LEA proteins (Fig. 1A, Tab. 2). The overall charge of *Afr*LEAI is negative, and the CDF analysis predicts the protein to be pre-molten globular or to contain a mixture of coils and globular structures (Fig. 1B). CIDER predicts *Afr*LEAI to be a Janus sequence, similar to *Af*LEA1.1, and to undergo environmental conditions-dependent conformational transitions (Fig. 1C). *Afr*LEAI is predicted by the SMART server to be the most repetitive of the group 3 LEA proteins identified in *A. franciscana* with two distinct sets of repeating motifs. The first set of repeats spans amino acid position 5-47 and 56-98 (Fig. 3B). Further inspection of the sequence suggests that the physicochemical properties of the repeats are conserved for positions 5-58 and 60-118. These repeats are highly enriched in aromatic residues, which is a unique feature among the LEA proteins in *Afranciscana*. Furthermore, both repeats are enriched in alanine, which is well-established as an α-helix forming amino acid (Pace and Scholtz, 1998).

[Fig 1 here]

The second set of repeats spans amino acid positions 116-221 and 244-331. Each of these repeats consists of three highly conserved motifs composed of a hydrophobic cluster containing a proline-phenylalanine pair, which is followed by regions enriched in alanine, aromatic residues, and clusters of negative amino acids which are separated by three arginine residues (Fig. 3B). These repeats are predicted to contain coiled-coil regions at positions 98-125, 186-252, and 304-327. The hydrophobic cluster regions are, once again, enriched in aromatic residues, such as phenylalanine and tyrosine. Being enriched in alanine and complementary charges, these regions are predicted by MeDOR-based Pred2ary algorithm to readily form α-helices. Additionally, any α-helices in this region would have a hydrophobic face due to the linear alignment of hydrophobic residues on the helix surface. This face would be flanked on one side by an alternating negative-positive-negative stripe and a thin polar stripe, similar to *Af*LEA1.1 (Fig. 3B). The N-terminal domain has a consistently oscillating hydropathy, correlating to the charged, alanine rich regions, and aromatic hydrophobic clusters. Combining these amphipathic α-helical tendencies with the coiled-coil behavior predicted by the SMART server suggests that *Afr*LEAI may form a bundle of amphipathic α-helices capable of interacting with phospholipid bilayers and monolayers. An NCBI BLAST of *Afr*LEAI supports this interpretation considering the homologies found to perilipinproteins that are known to interact with phospholipid monolayers (e-value of 3e-12c to perilipin-4, XP_013194305).

ANCHOR and PONDR®both predict several different MoRF regions in *Afr*LEAI (Fig. 4B). The close agreement between these two programs suggests that *Afr*LEAI may undergo extensive conformational transitions either through the loss of water interactions or by contact with target molecules. Given the degree of shift in the PONDR®VLXT score, it may be that *Afr*LEAI binds to proteins or lipids under conditions of minimal water reduction, or even in the hydrated state, but considering the negative charge of the protein, it is unlikely that *Afr*LEAI will interact with nucleic acids. PONDR®VL3 profile suggests that the C-terminal repeats can be structurally segregated into separate domains, as well as the first two repeats of the N-terminal region (Fig. 4B). The second pair of repeats are combined in one domain, which correlates to a coiled-coil prediction by the SMART server (Fig. 3B). PONDR®VL3 predicts that the final two repeats of the N-terminus fall into one single structural domain, which is distinct from the first two domains. Both programs predict high MoRF potential in the hydrophobic, aromatic half of each repeat, which is separated by a proline residue from the more hydrophilic half. The conservation of aromatic residues in this region offers insights into potential binding partners, or points to aromatic stabilization of the structure (Lanzarotti, Biekofsky, Estrin, Marti, & Turjanski, 2011). The highly charged, alanine-rich halves of the N-terminal repeats are predicted to be disordered in the hydrated state by both IUPred and PONDR, but any conformational shifts during desiccation would favor α-helical conformations with a high capacity for tertiary structure due to alternating charges represented by a κ value of 0.145 (Tab. 2).

AfrLEA2

Compared to *Afr*LEAIthe protein *Afr*LEA2 has a substantially lower mean net charge over hydropathy ratio and is the second most hydrophobic LEA aside from *Afr*LEA6 (Fig. 1A, Tab. 2). *Afr*LEA2 has been shown to have protective effects on lipid vesicles (Moore, et al., 2016) and cytoplasmic and mitochondrial enzymes during desiccation, although the protection was not dramatically better than that conferred by bovine serum albumin (Boswell, et al., 2014). Many group 3 LEA proteins are characterized by repeating amino acid motifs that may fold into amphipathic α-helices during desiccation (A. Tunnacliffe and Wise, 2007), however *Afr*LEA2 does not contain a repeating sequence. Additionally, secondary structure data for *Afr*LEA2 using circular dichroism values at $[\theta]_{200}$ (-10205.4) and $[\theta]_{222}$ (-1509.88) suggest that the protein is most likely pre-molten globular in the hydrated state, with a net ensemble of ~19% β-sheets, ~4% α-helices, ~15% turns, and ~62% random coils. When desiccated, *Afr*LEA2 exhibited only ~5% β-pleated sheet structure, but the α-helical content increased from ~4% to ~50%, while turns remained at 15% (Boswell, et al., 2014), which agrees with CIDER prediction of Janus sequence-like structural plasticity (Fig. 1C). While this data sheds light on the degree of secondary structure adoption that *Afr*LEA2 undergoes during desiccated, the actual structure of any given polypeptide strand in the sample may vary substantially within the conformational ensemble, or may shift from one conformation to another in the hydrated state (V. N. Uversky and Ptitsyn, 1994). Furthermore, some LEA proteins have been observed to undergo different conformational transitions depending on the presence of monovalent or divalent ions (Furuki, Shimizu, Kikawada, Okuda, & Sakurai, 2011). However, even with structural plasticity and ion-interactions considered, the shift in the prevalence of ordered secondary structure during desiccation suggests a transition from a native pre-molten globular structure to a potentially active molten globule. This prediction is further supported by the CDF analysis, which places *Afr*LEA2 both above and slightly below the boundary for molten globular and globular proteins (Fig. 1B). Therefore, experimental evidence regarding structural uniformity or localization of structural motifs in the *Afr*LEA2 polypeptide is needed to gain further insight into the specific mechanisms by which this protein may increases desiccation tolerance in *A. franciscana*.

DisEMBLpredicts an overall degree of disorder of approximately 70.3% (see Supplementary Material, Fig. S4), which is fairly close to the circular dichroism data and the IUPred and GlobPlotoutputs according to which *Afr*LEA2 is expected to have 78.6% and 78.5%

disorder, respectively. The agreement among the predicted and experimental data is encouraging for our approach of combining the localized structural predictions with the circular dichroism data of *Afr*LEA2 to elucidate local structural propensities in the polypeptide chain.Given that these programs are trained to distinguish IDPs and IDPRs from globular proteins and domains, and they accurately predict the degrees of order in *Afr*LEA2, then the positions of these ordered regions might be reliable. Furthermore, the IUPred accuracy in determining *Afr*LEA2 structure is inspiring for the application of ANCHOR, which uses similar techniques (Dosztányi, et al., 2009).Based on this analysis, *Afr*LEA2 in the hydrated state is likely composed of a β-sheet in the first 81 amino acids of the structure and a highly-disordered, C-terminal tail with some α-helical tendency at amino acid position 280-300. Perhaps most notably, the Remark-465 predictions were the most accurate from GlobPlot and DisEMBLE, which suggests that the *Afr*LEA2 curve on the CDF suggests a combination of ordered structures and disordered regions rather than a cohesive molten globule in the hydrated state. It should be noted that the Pred2ary predictions from MeDOR significantly deviated from the experimental data, which suggests that these ordered regions are small and may interact with turns (Fig. 3C).

For *Afr*LEA2 to follow molten globular and globular folding patterns, it would need to be structurally distinct from the other group 3 LEA proteins in *A. franciscana*. This hypothesis is supported by the difference in both structure and conformational changes during drying observed for*Afr*LEA2 when compared to *Afr*LEA3m (Boswell, et al., 2014).From a bioinformatics perspective, the amino acid sequence of *Afr*LEA2 is indeed distinct from all other *A. franciscana* LEA proteins. As aforementioned, the net mean charge of *Afr*LEA2 is low, due to the positively and negatively charge residues being well balanced (58 negative and 53 positive residues), which make up approximately 30% of the protein. Furthermore, *Afr*LEA2 shows no signs of repeat sequences, whereas all other LEA protein contain several repeating sequences, sometimes making up almost the entire protein. The lack of repeating sequences is particularly surprising because *Afr*LEA2 is the largest known LEA protein in *A. franciscana*. This finding becomes even more noteworthy in the context of LEA proteins in general, which are characterized by the presence of specific repeating motifs that are typically used for the classification of LEA proteins (A. Tunnacliffe and Wise, 2007).

Several other unique features are observed in *Afr*LEA2. The protein shows an uneven distribution of proline and arginine residues throughout the polypeptide chain. Of the 12 proline residues in its sequence, 11 are observed after position 200 and six of them fall between amino acids positions 200 and 290 (Fig. 3C). Similarly, of the 12 arginine residues in of the protein, nine are observed after position 235, whereas the other charged residues appear to be relatively equally distributed throughout the protein. This suggests that in the region from amino acid 200 to 364, any secondary structure elements that may form under any condition would be interrupted by proline or glycine residues every 10-40 amino acids.DISPHOS predicts 18 phosphorylated serine residues in *Afr*LEA2 (see Supplementary Material, Fig. S5) and 13 fall between amino acid positions 200 and 290. These predicted phosphate groups may help to overcome electrostatic repulsion in the protein.

Also, contrasting to the other group 3 LEA proteins, *Afr*LEA2 does not show an even distribution of its predicted MoRF regions (see Fig. 4C). Aside from small MoRF regions with relatively low probability at positions 30-37 and 151-156, ANCHOR predicts the MoRFs to mainly occur downstream of a high probability MoRF at position 180 – 189. Following this 10 amino acid MoRF are three MoRF regions of nine amino acid that are spaced fairly evenly every 15 amino acids apart from each other. Unlike other LEA proteins, these MoRFs are not highly similar in sequence. After this region of small MoRFs follows a region containing three larger MoRFs ranging from 15 to 23 amino acids. These three MoRFs are quite different from each other except for a reoccurring small region of 3 hydrophobic amino acids flanked by charged and polar residues on either side.

PONDR®VLXT predicts a particularly ordered N-terminus, which suggests that its structure is mainly regulated by hydrophobic interactions and may explain the increase in α-helices observed by CD (Boswell, et al., 2014) (Fig. 4C). The stretch of amino acids 29-98, which is associated with high α-helical propensity and a high hydrophobic moment, has previously been predicted to form amphipathic α-helices (Moore, et al., 2016) (Fig. 5C). The N-terminal region of the protein up to amino acid position 180 correlates to the observed ~40% of α-helices in the desiccated state. This suggests that the C-terminus functions as either a functional domain that utilizes intrinsic disorder or functions as a targeting domain that undergoes a conformational transition when in contact with a binding partner rather than due to environmental factors.

The C-terminal domain is separated into two sub-domains by PONDR®VL3 (Fig. 4C). The first sub-domain spans from amino acid position 180 – 290 and contains a 10 residue-long MoRF and a cluster of three 9 residue-long MoRFs, which are simultaneously predicted by both PONDR®VLXT and ANCHOR. This region is enriched in serine residues which are likely to be phosphorylated and leucine, valine, phenylalanine, and lysine residues (Fig. 4C). The second C-terminal domain is composed mainly of three large MoRFs, enriched in isoleucine, methionine, leucine, and arginine residues.

Given the unique feature of the C-terminal region ranging from approximately amino acid position 180-364, it may be predicted that this the region is subjected to desiccation-induced folding. Expectedly, it appears that the length of this region directly correlates with the degree of secondary structure detected by circular dichroism in the dry as state (Boswell, et al., 2014). This is of particular importance considering the content of proline and glycine in the region that would break apart any α-helices that might be forming in this region. Furthermore, this region has an amino acid composition that is not conducive to form amphipathic α-helices. Heliquest predicts that possible α-helices in this region would have a lower hydrophobic moment than at any other position in the protein, except for a region spanning from about amino acid position 275 to about 300 (Fig. 5C). While it is unlikely that the CD detected secondary structure is exclusively located within this C-terminal region, it is reasonable to suggest that the degree of secondary structure in this region is higher than in the remainder of the polypeptide. This information can be highly useful for experiments regarding the function of *Afr*LEA2, such as ectopic expression of the C-terminal region and comparing effects of this region and full length *Afr*LEA2 on physiological properties of model cells under water stress, or using site-directed mutagenesis to remove the prolines separating the MoRF regions and observing the shift in secondary structure during desiccation of the protein via CD spectroscopy.

AfrLEA3m

*Afr*LEA3m has been shown to localize in the mitochondria (Boswell, et al., 2014; Menze, et al., 2009), which explains the peculiar cysteine residues near the N-terminus, which is most likely being cleaved off after the protein is incorporated into the mitochondrial matrix (Menze, et al., 2009). Therefore, the first 31 amino acids, which was predicted to serve as the signal sequence, are excluded from the bioinformatics analyses conducted in this study. *Afr*LEA3m is the least hydrophobic LEA protein known to occur in *A. franciscana* that belongs to group 3, falling very close to the group 1 protein *Af*LEA1.1 (Fig. 1A). Compared to the other group 3 members, this LEA protein contains the largest fraction of charged residues, making up approximately 38.8% of the sequence, but the distribution of charges is the most even observed for LEA proteins from *A. franciscana*, with a κ value of 0.072 (Tab. 2). CIDER predicts *Afr*LEA3m to be a strong polyampholyte, which, having such a low κ value, should be self-repulsive unless the charges are aligned via the adoption of secondary structure (Fig. 1C).

The protein is predicted by CDF analysis to be mainly intrinsically disordered, making it the only group 3 LEA protein to fall below the boundary of the CDF (Fig. 1B). This further suggests a somewhat structured protein with high self-repulsion in the hydrated state. Its proximity to the group 1 protein *Af*LEA1.1 on the CH-plot is of particular interest given its sequence length and its classification as a member of group 3 LEA proteins. DisEMBL disorder prediction for *Afr*LEA3m suggest that this protein is that about 89.5% disordered in the hydrated state, although this percentage drops to 69.1% if Remark-465 is not being considered (see Supplementary Material, Fig. S6). The observed degree of disorder for *Afr*LEA3m by CD spectroscopy (Boswell, et al., 2014) is approximately 74% in the hydrated state and reduces to approximately 60% during desiccation. The predictions by DisEMBL, after removing the consideration of missing electron density in structures of globular protein domains, falls closely between the hydrated and dry states measured experimentally. This is particularly interesting because it implies that *Afr*LEA3m may fulfill some functions in the hydrated state, that only a few key regions are regulated by desiccation, or perhaps that its secondary structure is not as important to its function as previously hypothesized. This is not to say that tertiary structure, such as the predicted coiled-coil region, may not be regulated by desiccation and be crucial for function, but the methods currently employed do not adequately address these possibilities.

The Smart Server predicts two 46-48 residue-long repeats at positions 116-163 and positions 191-236 separated by a coiled-coil region spanning amino acids 157-185. These repeats and the coiled-coil region each coincide with α-helices predicted by MEDOR (Fig. 3D) and GlobPlot (see Supplementary Material, Fig. S7). Furthermore, each of the α-helices is predicted to be amphipathic in nature by Heliquest, implying helical interactions among the three regions (Fig. 5D). These higher-order folding patterns may be relevant to interactions with lipids and/or membranes. In this way, *Afr*LEA3m resembles *Afr*LEAI, although the former protein is potentially less ordered in the desiccated state. This may offer support to the hypothesis that tertiary structure is relevant to *Afr*LEA3m function in the desiccated state.Combined with the potential relevance of *Afr*LEAI tertiary structure to its function, it may be that group 3 LEA proteins adopt more tertiary structure during desiccation compared to members from groups 1 and 6.

The mature protein most likely spans from amino acid position 31-307 based on the indications from IUPred, GlobPlot, and a review of signal peptides from *D. melanogaster*. ANCHOR predicts a similar MoRF region pattern as observed for the cytoplasmic *Af*LEA1.1 and *Afr*LEAI proteins in that two distinct and different MoRF regions are found around the protein termini, and the appearance of internal MoRF regions correlates with repeating amino acid patterns (Fig. 4D). The PONDR®VLXT plot shows several peaks and troughs with extreme slopes spanning the entirety of the protein, suggesting that the majority of folding should be regulated by some binding partner (Fig. 4D). The PONDR®VL3 predictor also shows three distinct domains, which correlates with the arrangement of MoRF sites predicted by ANCHOR. It appears that *Afr*LEA3m, like *Afr*LEAI, may be associated with membranes or other lipids due to the amphipathic coiled-coil region predicted to occur roughly in the middle of the protein. Perhaps a unique role for *Afr*LEA3m might be to undergo a conformational shift exclusively in the presence of a membrane to orientate its hydrophobic face. The distribution of charges may also allow *Afr*LEA3m to interact in some way with others of itself, forming some kind of loosely associating matrix with nanogel like properties, even the proteins will only interact among each other via non-covalent bonding.

Group 6

AfrLEA6

*Afr*LEA6 is unique due to its position on the CH-plot being well within the region where most globular proteins fall (Fig. 1A). While it is flanked by two well-characterized IDPs, α-synuclein and γ-synuclein, its location is right on the edge of where such exceptions are observed. The mean net charge to hydropathy ratio of 0.094 is comparable to the one observed for *Afr*LEAI. *Afr*LEA6 is classified as a group 6 LEA protein, which is the most recently defined group that shows, compared to other LEA groups, unusual characteristics and hydropathy is not considered a major characteristic of this LEA group. CIDER predicts *Afr*LEA6 to be a weak polyampholyte, potentially forming a tadpole or globular structure (Fig. 1C). This globular structure would seem to agree with the CH-plot. Despite being predicted to be globular by the

CH-plot, the overall DisEMBL prediction of disorder for *Afr*LEA6 is 80.9% (see Supplementary Material, Fig. S8) and CDF analysis places *Afr*LEA6 well below the boundary (Fig. 1B). This may be indicative that it is another exceptional IDP, but the predictions from each program, and even within the same program, may offer additional insight into the structure and behavior of this protein.

Algorithms using missing electron density from x-ray crystallography data tend to suggest that *Afr*LEA6 is a globular protein with less than 40% disorder, including DisEMBL Remark-465 prediction. Algorithms that predict disorder using secondary structure propensity such as Pred2ary from MEDOR, Loops/Coils from DisEMBL, and IUPred predict that the degree of disorder for *Afr*LEA6 ranges from about 50%-75%. The Hot-Loops predictor, which is based on the B-factor, predicts that only 32.3% of the polypeptide is disordered, and therefore agrees with the results of GlobPlot analysis.While the programs appear to disagree on whether or not the disorder propensity breaches an appropriate threshold, they are quite consistent in showing the locations of possible disordered regions and domains. Each predictor suggests that there are regions with a high likelihood of order juxtaposed to regions with a high propensity for disorder. Programs that smooth the data appear to favor an ordered interpretation, whereas programs with smaller windows or less smoothing tend to favor disorder, implying that there are small, defined regions of order and disorder scattered throughout the protein.

The SMART server predicts that *Afr*LEA6 has two Pfam-SMP domains, one at position 9-55 and the other at position 90-137 (Fig. 3E). Pfam-SMP, or seed maturation proteins, are associated with desiccation tolerance in seeds, but have not been characterized in animals, with *Artemia* being the only animal known to express a protein containing Pfam-SMP domains. In contrast to the second SMP domain, the first domain is recognized by NCBI BLAST, although the second region has a very high sequence similarity to the first domain. Both domains appear to be parts of a larger repeat, spanning from amino acid positions 2-70 and 81-155.At position 140 to 184, appears to be a large region with a very high concentration of proline residues. Half of the proline residues in the entire protein are concentrated into this relatively short region, spanning approximately 11% of the sequence. Given the nature of proline as an α-helix and β-sheet disruptor, it is unlikely that defined secondary structures fall within this region. The prolines are also spaced in such a way as to make a poly-proline helix unlikely, which suggests that this region remains disordered at any hydration level. In addition to the high content of prolines, this region contains several hydrophobic residues, making it exceptionally hydrophobic for a disordered region. Aromatic residues such as tyrosine and phenylalanine are disproportionately included in this region as well. This may also explain the problems that missing electron density programs have for predicting secondary structure features in this location.Following the proline-rich region is again a region with similarity to the Pfam-SMP domain, although it is somewhat more degenerated from the two aforementioned domains. Given the length of each repeat, it appears that the protein is composed of 3 repeats, with one region of the last repeat being less conserved and enriched in proline residues than the other two regions. This may indicate that the third region has evolved from an SMP domain into a distinct domain with unknown functions. The C-terminal region exhibits a unique staggering of positive and negative charges separated by proline and glycine residues, potentially allowing folding in the desiccated state (Fig. 3E).

ANCHOR predicts two conserved MoRF regions within the N-terminal SMP domains, which fall in a region of relatively low disorder-propensity (Fig. 4E). The second half of the second SMP domain has a large MoRF region ranging from amino acid 105 to 134, which is not shared with the first SMP domain. PONDR®VL-XT predicts weak potential binding capacity shared between the last 20 amino acids of the second SMP domain, the proline-rich region, and the first half of the C-terminal region (Fig. 4E). An N-terminal disordered region correlates with the disorder prediction of IUPred (Fig. 4E), and the C-terminus has a disordered region with limited binding capacity that coincides with the MoRF region predicted by ANCHOR. PONDR®VL3 predicts three distinct domains, separated as an N-terminal domain at the point where the SMP domains meet, a large domain spanning the combined MoRF regions described above including the proline-rich region and the neighboring regions, and a C-terminal domain downstream of the MoRF region. Due to the occurrence of charges in the internal region that may be complementary to charges at N-terminal and C-terminal regions the desiccated protein likely forms a structure resembling a bio-glass.

Conclusions

We have utilized a broad suite of open source bioinformatics tools to gain insights into the dynamic structures of LEA proteins from the brine shrimp *A. franciscana*. Results of our analysis were used to refine current hypotheses regarding the function of LEA proteins in animals. Our analysis indicates that LEA proteins from different groups are more similar than we originally hypothesized, while functional differences among members of group 3 are possibly larger than commonly anticipated. Each of the LEA proteins analyzed, except for *Afr*LEAI, had three distinct domains; one at each terminus with potential binding sites connected by an intermediary domain. We predict that *Afr*LEA1.1 is a highly-disordered protein with coil-like structure that appears to have two distinct MoRF domains on either side of a repeating internal spacer domain and is predicted to be a Janus sequence that exists as a mostly random coil in the hydrated state. The internal domain may undergo a conformational transition during water loss, pulling the terminal MoRF sites, and potentially attached binding partners, closer together during desiccation.

The group 3 LEA proteins all showed domains with amphipathic α-helix propensities, but otherwise showed substantial differences among each other. *Afr*LEAI, as previously noted, is the only LEA protein with just two distinct domains, an N-terminal domain with more even distribution of hydrophobic and charged residues, and a C-terminal domain with six repeats of a coiled region that may form amphipathic, potentially self-interactive, α-helices, which could form a perilipin-like bundle. *Afr*LEAI also appears to be the most readily protein-binding LEA protein found in *A. franciscana*, potentially interacting with multiple partners, and is one of the two LEA proteins that appears to be molten globular in the hydrated state. *Afr*LEAI is predicted to function as a Janus sequence which should undergo conformational changes during desiccation.

*Afr*LEA2 is more hydrophobic than the other group 3 LEA proteins and has no detectable internal repeats in its sequence. It has a uniquely stable intermediary domain that likely includes the observed α-helical MoRFs found in CD spectra (Boswell, et al., 2014). This increase in orderly structure supports our prediction that *Afr*LEA2 functions as a Janus sequence, and bolsters our confidence in similar results for the other the proteins not yet characterized by CD spectroscopy. The relatively small N- and C-terminal domains likely interact with binding partners and *Afr*LEA2 appears to be natively either molten globular or to contain globular regions in the hydrated state. *Afr*LEA3m uniquely categorizes as a strong polyampholyte of low mean net charge with a low κ value, which suggests that it should maintain a relatively high degree of disorderdespite desiccation. The termini appear to have MoRFs, which are separated by an intermediate spacer region. The distribution of charges may be overcome by folding into an α-helical conformation in this region, but not at the termini. Staggering of two or more of this protein might also facilitate favorable protein interactions, rather than gaining substantial structure on its own. Most certainly, *Afr*LEA3m will need a compatible binding partner before it undergoes a conformational transition, instead of being regulated more readily by desiccation as the other group 3 LEA proteins appear to be.

*Afr*LEA6 is the most distinct LEA protein compared to the other LEAs in *A. franciscana*. It is by far the most hydrophobic and the protein contains two SMP domains, which appear to function only when they interact with another sequence. *Afr*LEA6 has a tremendously proline-enriched intermediate domain that may either function as a highly flexible spacer or as a very unique binding site. The N-terminal domain is composed of a proline- and isoleucine-rich region flanked by two SMP domains, which begin with low PONDR score and transition suddenly to a high score. This slope does not strictly indicate a binding site, but may points to the potential of self-interaction between the SMP domains. The juxtaposition of SMP domains upstream of proline regions indicates that this pattern might be important for its function, which has yet to be elucidated. The C-terminus has a distinct separation of charges that makes it very susceptible to binding other proteins in the desiccated state, contributing to the model of a weak polyampholyte tadpole. In such a model, the N-terminus might act as a globular “head” whereas the C-terminus would act as a sticky “tail” which coil to form some kind of glassy or gel-like matrix.

Overall, our investigation indicates a variety of differences in form and potential function(s) of LEA proteins expressed in *A. franciscana* during anhydrobiosis, but indicates that as a general strategy the animal utilizes glassy matrix forming LEAs concurrently with proteins that more likely interact with more specific binding partners. Nevertheless, the function(s) of both types, the matrix-forming and partner-binding LEA proteins, are likely regulated by changing water availability during desiccation.

Acknowledgments

This research was funded by the National Science Foundation under grant number IOS-1659970/IOS-1457061.

References

Aurora, R., & Rose, G. D. (1998). Helix capping. *Protein Sci, 7*(1), pp. 21-38. doi:10.1002/pro.5560070103

Battaglia, M., & Covarrubias, A. (2013). Late Embryogenesis Abundant (LEA) proteins in legumes. [Review]. *Frontiers in Plant Science, 4*(190)doi:10.3389/fpls.2013.00190

Battaglia, M., Olvera-Carrillo, Y., Garciarrubio, A., Campos, F., & Covarrubias, A. A. (2008). The Enigmatic LEA Proteins and Other Hydrophilins. *Plant Physiology, 148*(1), pp. 6-24. doi:10.1104/pp.108.120725

Boswell, L. C., Menze, M. A., & Hand, S. C. (2014). Group 3 late embryogenesis abundant proteins from embryos of Artemia franciscana: structural properties and protective abilities during desiccation. *Physiol Biochem Zool, 87*(5), pp. 640-651. doi:10.1086/676936

Bourhis, J., Canard, B., & Longhi, S. (2007). Predicting protein disorder and induced folding: from theoretical principles to practical applications. *Curr Protein Pept Sci, 8*doi:10.2174/138920307780363451

Bracken, C., Iakoucheva, L. M., Romero, P. R., & Dunker, A. K. (2004). Combining prediction, computation and experiment for the characterization of protein disorder. *Curr Opin Struct Biol, 14*(5), pp. 570-576. doi:10.1016/j.sbi.2004.08.003

Bremer, A., Wolff, M., Thalhammer, A., & Hincha, D. K. (2017). Folding of intrinsically disordered plant LEA proteins is driven by glycerol-induced crowding and the presence of membranes. *The FEBS Journal, 284*(6), pp. 919-936. doi:10.1111/febs.14023

Browne, J., Tunnacliffe, A., & Burnell, A. (2002). Anhydrobiosis: plant desiccation gene found in a nematode. *Nature, 416*(6876), p 38. doi:10.1038/416038a

Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., . . . Mornon, J. P. (1997). Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cellular and Molecular Life Sciences CMLS, 53*(8), pp. 621-645. doi:10.1007/s000180050082

Chakrabortee, S., Tripathi, R., Watson, M., Schierle, G. S., Kurniawan, D. P., Kaminski, C. F., . . . Tunnacliffe, A. (2012). Intrinsically disordered proteins as molecular shields. *Mol Biosyst, 8*(1), pp. 210-219. doi:10.1039/c1mb05263b

Chandonia, J. M., & Karplus, M. (1999). New methods for accurate prediction of protein secondary structure. *Proteins, 35*doi:3.0.co;2-l

Cheng, Y., Oldfield, C. J., Meng, J., Romero, P., Uversky, V. N., & Dunker, A. K. (2007). Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry, 46*(47), pp. 13468-13477. doi:10.1021/bi7012273

Clark, M. S., Thorne, M. A., Purać, J., Grubor-Lajšić, G., Kube, M., Reinhardt, R., & Worland, M. R. (2007). Surviving extreme polar winters by desiccation: clues from Arctic springtail (Onychiurus arcticus) EST libraries. [journal article]. *BMC Genomics, 8*(1), p 475. doi:10.1186/1471-2164-8-475

Coeytaux, K., & Poupon, A. (2005). Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics, 21*(9), pp. 1891-1900. doi:10.1093/bioinformatics/bti266

Denekamp, N. Y., Reinhardt, R., Kube, M., & Lubzens, E. (2010). Late embryogenesis abundant (LEA) proteins in nondesiccated, encysted, and diapausing embryos of rotifers. *Biol Reprod, 82*(4), pp. 714-724. doi:10.1095/biolreprod.109.081091

Disfani, F. M., Hsu, W. L., Mizianty, M. J., Oldfield, C. J., Xue, B., Dunker, A. K., . . . Kurgan, L. (2012). MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics, 28*(12), pp. i75-83. doi:10.1093/bioinformatics/bts209

Dosztanyi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics, 21*doi:10.1093/bioinformatics/bti541

Dosztányi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics, 21*(16), pp. 3433-3434. doi:10.1093/bioinformatics/bti541

Dosztányi, Z., Csizmók, V., Tompa, P., & Simon, I. (2005). The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *J Mol Biol, 347*(4), pp. 827-839. doi:http://doi.org/10.1016/j.jmb.2005.01.071

Dosztányi, Z., Mészáros, B., & Simon, I. (2009). ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics, 25*(20), pp. 2745-2746. doi:10.1093/bioinformatics/btp518

Dunker, A. K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., . . . Villafranca, J. E. (1998). Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput*, pp. 473-484.

Dunker, A. K., & Obradovic, Z. (2001). The protein trinity – linking function and disorder. *Nat Biotechnol, 19*doi:10.1038/nbt0901-805

Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., & Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform, 11*, pp. 161-171.

Dunker, A. K., Oldfield, C. J., Meng, J., Romero, P., Yang, J. Y., Chen, J. W., . . . Uversky, V. N. (2008). The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics, 9 Suppl 2*, p S1. doi:10.1186/1471-2164-9-s2-s1

Dure, L., & Chlan, C. (1981). Developmental Biochemistry of Cottonseed Embryogenesis and Germination : XII. PURIFICATION AND PROPERTIES OF PRINCIPAL STORAGE PROTEINS. *Plant Physiol, 68*(1), pp. 180-186.

Dure, L., & Galau, G. A. (1981). Developmental Biochemistry of Cottonseed Embryogenesis and Germination : XIII. REGULATION OF BIOSYNTHESIS OF PRINCIPAL STORAGE PROTEINS. *Plant Physiol, 68*(1), pp. 187-194.

Dure, L., Greenway, S. C., & Galau, G. A. (1981). Developmental biochemistry of cottonseed embryogenesis and germination: changing messenger ribonucleic acid populations as shown by in vitro and in vivo protein synthesis. *Biochemistry, 20*(14), pp. 4162-4168. doi:10.1021/bi00517a033

Eisenberg, D., Weiss, R. M., & Terwilliger, T. C. (1982). The helical hydrophobic moment: a measure of the amphiphilicity of a helix. [10.1038/299371a0]. *Nature, 299*(5881), pp. 371-374.

Fauchere, J.-L., & Pliska, V. (1983). Hydrophobic parameters pi of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem, 18*(3), pp. 369-375.

Furuki, T., Shimizu, T., Kikawada, T., Okuda, T., & Sakurai, M. (2011). Salt Effects on the Structural and Thermodynamic Properties of a Group 3 LEA Protein Model Peptide. *Biochemistry, 50*(33), pp. 7093-7103. doi:10.1021/bi200719s

Gal, T. Z., Glazer, I., & Koltai, H. (2004). An LEA group 3 family member is involved in survival of C. elegans during exposure to stress. *FEBS Letters, 577*(1–2), pp. 21-26. doi:https://doi.org/10.1016/j.febslet.2004.09.049

Garay-Arroyo, A., Colmenero-Flores, J. M., Garciarrubio, A., & Covarrubias, A. A. (2000). Highly hydrophilic proteins in prokaryotes and eukaryotes are common during conditions of water deficit. *J Biol Chem, 275*(8), pp. 5668-5674.

Garbuzynskiy, S. O., Lobanov, M. Y., & Galzitskaya, O. V. (2004). To be folded or to be unfolded? *Protein Sci, 13*(11), pp. 2871-2877. doi:10.1110/ps.04881304

Gautier, R., Douguet, D., Antonny, B., & Drin, G. (2008). HELIQUEST: a web server to screen sequences with specific α-helical properties. *Bioinformatics, 24*(18), pp. 2101-2102.

doi:10.1093/bioinformatics/btn392

Goyal, K., Tisi, L., Basran, A., Browne, J., Burnell, A., Zurdo, J., & Tunnacliffe, A. (2003). Transition from natively unfolded to folded state induced by desiccation in an anhydrobiotic nematode protein. *J Biol Chem, 278*(15), pp. 12977-12984. doi:10.1074/jbc.M212007200

Goyal, K., Walton, L. J., & Tunnacliffe, A. (2005). LEA proteins prevent protein aggregation due to water stress. *Biochem J, 388*(Pt 1), pp. 151-157. doi:10.1042/BJ20041931

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1186703/pdf/bj3880151.pdf

Graether, S. P., & Boddington, K. F. (2014). Disorder and function: a review of the dehydrin protein family. [Review]. *Frontiers in Plant Science, 5*(576)doi:10.3389/fpls.2014.00576

Grelet, J., Benamar, A., Teyssier, E., Avelange-Macherel, M. H., Grunwald, D., & Macherel, D. (2005). Identification in pea seed mitochondria of a late-embryogenesis abundant protein able to protect enzymes from drying. *Plant Physiol, 137*(1), pp. 157-167. doi:10.1104/pp.104.052480

Hand, S. C., Jones, D., Menze, M. A., & Witt, T. L. (2007). Life without water: expression of plant LEA genes by an anhydrobiotic arthropod. *J Exp Zool A Ecol Genet Physiol, 307*(1), pp. 62-66. doi:10.1002/jez.a.343

Hand, S. C., & Menze, M. A. (2015). Molecular approaches for improving desiccation tolerance: insights from the brine shrimp Artemia franciscana. *Planta, 242*(2), pp. 379-388. doi:10.1007/s00425-015-2281-9

Hand, S. C., Menze, M. A., Toner, M., Boswell, L., & Moore, D. (2011). LEA proteins during water stress: not just for plants anymore. *Annu Rev Physiol, 73*, pp. 115-134. doi:10.1146/annurev-physiol-012110-142203

Hincha, D. K., & Thalhammer, A. (2012). LEA proteins: IDPs with versatile functions in cellular dehydration tolerance. *Biochem Soc Trans, 40*(5), pp. 1000-1003. doi:10.1042/bst20120109

Hoekstra, F. A., Golovina, E. A., & Buitink, J. (2001). Mechanisms of plant desiccation tolerance. *Trends Plant Sci, 6*(9), pp. 431-438.

Holehouse, A. S., Ahad, J., Das, R. K., & Pappu, R. V. (2015). CIDER: Classification of Intrinsically Disordered Ensemble Regions. *Biophysical Journal, 108*(2), p 228a. doi:10.1016/j.bpj.2014.11.1260

Hunault, G., & Jaspard, E. (2010). LEAPdb: a database for the late embryogenesis abundant proteins. *BMC Genomics, 11*, p 221. doi:10.1186/1471-2164-11-221

Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., & Dunker, A. K. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res, 32*(3), pp. 1037-1049. doi:10.1093/nar/gkh253

Jaspard, E., Macherel, D., & Hunault, G. (2012). Computational and Statistical Analyses of Amino Acid Usage and Physico-Chemical Properties of the Twelve Late Embryogenesis Abundant Protein Classes. *PLoS One, 7*(5), p e36968. doi:10.1371/journal.pone.0036968

Jencks, W. P. (1981). On the attribution and additivity of binding energies. *Proc Natl Acad Sci U S A, 78*(7), pp. 4046-4050.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers, 22*(12), pp. 2577-2637. doi:10.1002/bip.360221211

Kall, L., Krogh, A., & Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *J Mol Biol, 338*(5), pp. 1027-1036. doi:10.1016/j.jmb.2004.03.016

Kikawada, T., Nakahara, Y., Kanamori, Y., Iwata, K., Watanabe, M., McGee, B., . . . Okuda, T. (2006). Dehydration-induced expression of LEA proteins in an anhydrobiotic chironomid. *Biochem Biophys Res Commun, 348*(1), pp. 56-61. doi:10.1016/j.bbrc.2006.07.003

Kovacs, D., Agoston, B., & Tompa, P. (2008). Disordered plant LEA proteins as molecular chaperones. *Plant Signal Behav, 3*(9), pp. 710-713.

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol, 157*(1), pp. 105-132.

Lanzarotti, E., Biekofsky, R. R., Estrin, D. A., Marti, M. A., & Turjanski, A. G. (2011). Aromatic–Aromatic Interactions in Proteins: Beyond the Dimer. *Journal of Chemical Information and Modeling, 51*(7), pp. 1623-1633. doi:10.1021/ci200062e

Lemke, E. A. (2011). Structure and Function of Intrinsically Disordered Proteins. By Peter Tompa. *ChemBioChem, 12*(8), pp. 1280-1280. doi:10.1002/cbic.201100142

Li, X., Romero, P., Rani, M., Dunker, A. K., & Obradovic, Z. (1999). Predicting Protein Disorder for N-, C-, and Internal Regions. *Genome Inform Ser Workshop Genome Inform, 10*, pp. 30-40.

Lieutaud, P., Canard, B., & Longhi, S. (2008). MeDor: a metaserver for predicting protein disorder. [journal article]. *BMC Genomics, 9*(2), p S25. doi:10.1186/1471-2164-9-s2-s25

Lieutaud, P., Ferron, F., Uversky, A. V., Kurgan, L., Uversky, V. N., & Longhi, S. (2016). How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe. *Intrinsically Disord Proteins, 4*(1), p e1259708. doi:10.1080/21690707.2016.1259708

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., & Russell, R. B. (2003). Protein disorder prediction: implications for structural proteomics. *Structure (Camb), 11*doi:10.1016/j.str.2003.10.002

Linding, R., Russell, R. B., Neduva, V., & Gibson, T. J. (2003). GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Research, 31*(13), pp. 3701-3708.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC169197/pdf/gkg519.pdf

Lise, S., & Jones, D. T. (2005). Sequence patterns associated with disordered regions in proteins. *Proteins, 58*(1), pp. 144-150. doi:10.1002/prot.20279

Lobley, A., Swindells, M. B., Orengo, C. A., & Jones, D. T. (2007). Inferring function using patterns of native disorder in proteins. *PLoS Comput Biol, 3* doi:10.1371/journal.pcbi.0030162

MacRae, T. H. (2016). Stress tolerance during diapause and quiescence of the brine shrimp, Artemia. *Cell Stress & Chaperones, 21*(1), pp. 9-18. doi:10.1007/s12192-015-0635-7

Marunde, M. R., Samarajeewa, D. A., Anderson, J., Li, S., Hand, S. C., & Menze, M. A. (2013).

Improved tolerance to salt and water stress in Drosophila melanogaster cells conferred by late embryogenesis abundant protein. *J Insect Physiol, 59*(4), pp. 377-386. doi:10.1016/j.jinsphys.2013.01.004

Meador, W. E., Means, A. R., & Quiocho, F. A. (1992). Target enzyme recognition by calmodulin: 2.4 A structure of a calmodulin-peptide complex. *Science, 257*(5074), pp. 1251-1255.

Menze, M. A., Boswell, L., Toner, M., & Hand, S. C. (2009). Occurrence of mitochondria-targeted Late Embryogenesis Abundant (LEA) gene in animals increases organelle resistance to water stress. *J Biol Chem, 284*(16), pp. 10714-10719. doi:10.1074/jbc.C900001200

Mészáros, B., Simon, I., & Dosztányi, Z. (2009). Prediction of Protein Binding Regions in Disordered Proteins. *PLoS Computational Biology, 5*(5), p e1000376. doi:10.1371/journal.pcbi.1000376

Moore, D. S., Hansen, R., & Hand, S. C. (2016). Liposomes with diverse compositions are protected during desiccation by LEA proteins from Artemia franciscana and trehalose. *Biochimica et Biophysica Acta (BBA) - Biomembranes, 1858*(1), pp. 104-115. doi:http://doi.org/10.1016/j.bbamem.2015.10.019

Mouillon, J. M., Gustafsson, P., & Harryson, P. (2006). Structural investigation of disordered stress proteins. Comparison of full-length dehydrins with isolated peptides of their conserved segments. *Plant Physiol, 141*(2), pp. 638-650. doi:10.1104/pp.106.079848

Neupert, W., & Herrmann, J. M. (2007). Translocation of proteins into mitochondria. *Annu Rev Biochem, 76*, pp. 723-749. doi:10.1146/annurev.biochem.76.052705.163409

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., & Dunker, A. K. (2003). Predicting intrinsic disorder from amino acid sequence. *Proteins, 53*doi:10.1002/prot.10532

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., & Dunker, A. K. (2005). Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins, 61 Suppl 7*, pp. 176-182. doi:10.1002/prot.20735

Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., & Dunker, A. K. (2005). Comparing and combining predictors of mostly disordered proteins. *Biochemistry, 44*(6), pp. 1989-2000. doi:10.1021/bi047993o

Oldfield, C. J., Cheng, Y., Cortese, M. S., Romero, P., Uversky, V. N., & Dunker, A. K. (2005). Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry, 44*(37), pp. 12454-12470. doi:10.1021/bi050736e

Pace, C. N., & Scholtz, J. M. (1998). A helix propensity scale based on experimental studies of peptides and proteins. *Biophysical Journal, 75*(1), pp. 422-427.

Pammenter, N. B., P. (1999). A review of recalcitrant seed physiology in relation to desiccation-tolerance mechanisms. *Seed Science Research, 9*(1), pp. 13-37.

Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., & Obradovic, Z. (2006). Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics, 7*, p 208. doi:10.1186/1471-2105-7-208

Peng, Z., Wang, C., Uversky, V. N., & Kurgan, L. (2017). Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind. *Methods Mol Biol, 1484*, pp. 187-203. doi:10.1007/978-1-4939-6406-2_14

Popova, A. V., Rausch, S., Hundertmark, M., Gibon, Y., & Hincha, D. K. (2015). The intrinsically disordered protein LEA7 from Arabidopsis thaliana protects the isolated enzyme lactate dehydrogenase and enzymes in a soluble leaf proteome during freezing and drying. *Biochim Biophys Acta, 1854*(10 Pt A), pp. 1517-1525. doi:10.1016/j.bbapap.2015.05.002

Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E. H., Man, O., Beckmann, J. S., . . . Sussman, J. L. (2005). FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics, 21*doi:10.1093/bioinformatics/bti537

Ptitsyn, O. B. (1995). Molten globule and protein folding. *Adv Protein Chem, 47*, pp. 83-229.

Radivojac, P., Iakoucheva, L. M., Oldfield, C. J., Obradovic, Z., Uversky, V. N., & Dunker, A. K. (2007). Intrinsic disorder and functional proteomics. *Biophys J, 92*(5), pp. 1439-1456. doi:10.1529/biophysj.106.094045

Radivojac, P., Obradovic, Z., Brown, C. J., & Dunker, A. K. (2003). Prediction of boundaries between intrinsically ordered and disordered protein regions. *Pac Symp Biocomput*, pp. 216-227.

Roise, D., & Schatz, G. (1988). Mitochondrial presequences. *J Biol Chem, 263*(10), pp. 4509-4511.

Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E., Garner, E., Guilliot, S., & Dunker, A. K. (1998). Thousands of proteins likely to have long disordered regions. *Pac Symp Biocomput*, pp. 437-448.

Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., & Dunker, A. K. (2001). Sequence complexity of disordered protein. *Proteins, 42*(1), pp. 38-48.

Schokraie, E., Hotz-Wagenblatt, A., Warnken, U., Mali, B., Frohme, M., Forster, F., . . . Schnolzer, M. (2010). Proteomic analysis of tardigrades: towards a better understanding of molecular mechanisms by anhydrobiotic organisms. *PLoS One, 5*(3), p e9502. doi:10.1371/journal.pone.0009502

Sharon, M. A., Kozarova, A., Clegg, J. S., Vacratsis, P. O., & Warner, A. H. (2009). Characterization of a group 1 late embryogenesis abundant protein in encysted embryos of the brine shrimp Artemia franciscana. *Biochem Cell Biol, 87*(2), pp. 415-430. doi:10.1139/o09-001

Shih, M.-D., Hoekstra, F. A., & Hsing, Y.-I. C. (2008). Late Embryogenesis Abundant Proteins. In K. Jean-Claude & D. Michel (Eds.), *Advances in Botanical Research* (Vol. Volume 48, pp. 211-255): Academic Press.

Shih, M. D., Hsieh, T. Y., Lin, T. P., Hsing, Y. I., & Hoekstra, F. A. (2010). Characterization of two soybean (Glycine max L.) LEA IV proteins by circular dichroism and Fourier transform infrared spectrometry. *Plant Cell Physiol, 51*(3), pp. 395-407. doi:10.1093/pcp/pcq005

Shimizu, T., Kanamori, Y., Furuki, T., Kikawada, T., Okuda, T., Takahashi, T., . . . Sakurai, M. (2010). Desiccation-induced structuralization and glass formation of group 3 late embryogenesis abundant protein model peptides. *Biochemistry, 49*(6), pp. 1093-1104. doi:10.1021/bi901745f

Solomon, A., Salomon, R., Paperna, I., & Glazer, I. (2000). Desiccation stress of entomopathogenic nematodes induces the accumulation of a novel heat-stable protein.

Parasitology, 121 (Pt 4), pp. 409-416.

Spolar, R. S., & Record, M. T., Jr. (1994). Coupling of local folding to site-specific binding of proteins to DNA. *Science, 263*(5148), pp. 777-784.

Steponkus, P. L., Uemura, M., Joseph, R. A., Gilmour, S. J., & Thomashow, M. F. (1998). Mode of action of the COR15a gene on the freezing tolerance of Arabidopsis thaliana. *Proc Natl Acad Sci U S A, 95*(24), pp. 14570-14575. doi:10.1073/pnas.95.24.14570

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC24414/pdf/pq014570.pdf

Thalhammer, A., Hundertmark, M., Popova, A. V., Seckler, R., & Hincha, D. K. (2010). Interaction of two intrinsically disordered plant stress proteins (COR15A and COR15B) with lipid membranes in the dry state. *Biochim Biophys Acta, 1798*(9), pp. 1812-1820. doi:10.1016/j.bbamem.2010.05.015

Tolleter, D., Hincha, D. K., & Macherel, D. (2010). A mitochondrial late embryogenesis abundant protein stabilizes model membranes in the dry state. *Biochim Biophys Acta, 1798*(10), pp. 1926-1933. doi:10.1016/j.bbamem.2010.06.029

Tolleter, D., Jaquinod, M., Mangavel, C., Passirani, C., Saulnier, P., Manon, S., . . . Macherel, D. (2007). Structure and function of a mitochondrial late embryogenesis abundant protein are revealed by desiccation. *Plant Cell, 19*(5), pp. 1580-1589. doi:10.1105/tpc.107.050104

Tompa, P., & Kovacs, D. (2010). Intrinsically disordered chaperones in plants and animals. *Biochem Cell Biol, 88*(2), pp. 167-174. doi:10.1139/o09-163

Toxopeus, J., Warner, A. H., & MacRae, T. H. (2014). Group 1 LEA proteins contribute to the desiccation and freeze tolerance of Artemia franciscana embryos during diapause. *Cell Stress Chaperones, 19*(6), pp. 939-948. doi:10.1007/s12192-014-0518-3

Tunnacliffe, A., Lapinski, J., & McGee, B. (2005). A Putative LEA Protein, but no Trehalose, is Present in Anhydrobiotic Bdelloid Rotifers. *Hydrobiologia, 546*(1), pp. 315-321. doi:10.1007/s10750-005-4239-6

Tunnacliffe, A., & Wise, M. J. (2007). The continuing conundrum of the LEA proteins. *Naturwissenschaften, 94*(10), pp. 791-812. doi:10.1007/s00114-007-0254-y

Uverskii, V. N. (1998). [How many molten globules states exist?]. *Biofizika, 43*(3), pp. 416-421.

Uversky, V. N. (1997). Diversity of compact forms of denatured globular proteins. *Protein & Peptide Letters, 4*, pp. 355-367.

Uversky, V. N. (2002). Natively unfolded proteins: A point where biology waits for physics. *Protein Science : A Publication of the Protein Society, 11*(4), pp. 739-756.

Uversky, V. N. (2013a). A decade and a half of protein intrinsic disorder: biology still waits for physics. *Protein Sci, 22*(6), pp. 693-724. doi:10.1002/pro.2261

Uversky, V. N. (2013b). Unusual biophysics of intrinsically disordered proteins. *Biochim Biophys Acta, 1834*(5), pp. 932-951. doi:10.1016/j.bbapap.2012.12.008

Uversky, V. N. (2016). Dancing Protein Clouds: The Strange Biology and Chaotic Physics of Intrinsically Disordered Proteins. *J Biol Chem, 291*(13), pp. 6681-6688. doi:10.1074/jbc.R115.685859

Uversky, V. N., & Dunker, A. K. (2010). Understanding protein non-folding. *Biochim Biophys*

Acta, 1804(6), pp. 1231-1264. doi:10.1016/j.bbapap.2010.01.017

Uversky, V. N., & Dunker, A. K. (2012a). *Intrinsically disordered protein analysis Volume 1, Volume 1* New York: Springer.

Uversky, V. N., & Dunker, A. K. (2012b). *Intrinsically disordered protein analysis Volume 2, Volume 2* New York: Springer.

Uversky, V. N., Gillespie, J. R., & Fink, A. L. (2000). Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Bioinformatics, 41*(3), pp. 415-427. doi:10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7

Uversky, V. N., & Ptitsyn, O. B. (1994). "Partly folded" state, a new equilibrium state of protein molecules: four-state guanidinium chloride-induced unfolding of beta-lactamase at low temperature. *Biochemistry, 33*(10), pp. 2782-2791.

Uversky, V. N., & Ptitsyn, O. B. (1996). Further evidence on the equilibrium "pre-molten globule state": four-state guanidinium chloride-induced unfolding of carbonic anhydrase B at low temperature. *J Mol Biol, 255*(1), pp. 215-228. doi:10.1006/jmbi.1996.0018

Uversky, V. N., Radivojac, P., Iakoucheva, L. M., Obradovic, Z., & Dunker, A. K. (2007). Prediction of intrinsic disorder and its use in functional proteomics. *Methods Mol Biol, 408*, pp. 69-92.

Vucetic, S., Brown, C., Dunker, K., & Obradovic, Z. (2003). Flavors of protein disorder. *Proteins, 52*doi:10.1002/prot.10437

Warner, A. H., Chakrabortee, S., Tunnacliffe, A., & Clegg, J. S. (2012). Complexity of the heat-soluble LEA proteome in Artemia species. *Comp Biochem Physiol Part D Genomics Proteomics, 7*(3), pp. 260-267. doi:10.1016/j.cbd.2012.04.002

Warner, A. H., Guo, Z. H., Moshi, S., Hudson, J. W., & Kozarova, A. (2016). Study of model systems to test the potential function of Artemia group 1 late embryogenesis abundant (LEA) proteins. *Cell Stress Chaperones, 21*(1), pp. 139-154. doi:10.1007/s12192-015-0647-3

Warner, A. H., Miroshnychenko, O., Kozarova, A., Vacratsis, P. O., MacRae, T. H., Kim, J., & Clegg, J. S. (2010). Evidence for multiple group 1 late embryogenesis abundant proteins in encysted embryos of Artemia and their organelles. *J Biochem, 148*(5), pp. 581-592. doi:10.1093/jb/mvq091

Wise, M. J., & Tunnacliffe, A. (2004). POPP the question: what do LEA proteins do? *Trends Plant Sci, 9*(1), pp. 13-17. doi:10.1016/j.tplants.2003.10.012

Wright, P. E., & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol, 293*(2), pp. 321-331. doi:10.1006/jmbi.1999.3110

Wu, G., Zhang, H., Sun, J., Liu, F., Ge, X., Chen, W. H., . . . Wang, W. (2011). Diverse LEA (late embryogenesis abundant) and LEA-like genes and their responses to hypersaline stress in post-diapause embryonic development of Artemia franciscana. *Comp Biochem Physiol B Biochem Mol Biol, 160*(1), pp. 32-39. doi:10.1016/j.cbpb.2011.05.005

Xie, H., Vucetic, S., Iakoucheva, L. M., Oldfield, C. J., Dunker, A. K., Obradovic, Z., & Uversky, V. N. (2007). Functional Anthology of Intrinsic Disorder. III. Ligands, Postranslational Modifications and Diseases Associated with Intrinsically Disordered Proteins. *Journal of proteome research, 6*(5), pp. 1917-1932. doi:10.1021/pr060394e

Yang, Z. R., Thomson, R., McNeil, P., & Esnouf, R. M. (2005). RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics, 21*doi:10.1093/bioinformatics/bti534

Yu, J., Lai, Y., Wu, X., Wu, G., & Guo, C. (2016). Overexpression of OsEm1 encoding a group I LEA protein confers enhanced drought tolerance in rice. *Biochem Biophys Res Commun, 478*(2), pp. 703-709. doi:https://doi.org/10.1016/j.bbrc.2016.08.010

Figure Legends

Figure 1. Global analysis of intrinsic disorder predispositions of LEA proteins from *A. franciscana*. **A**. CH-plot including LEA proteins from *A. franciscana* (diamonds) that are plotted together with a set of known IDPs (red circles), and globular proteins (blue squares). **B**. CDF analysis of LEA proteins from *A. franciscana*. The order-disorder boundary is shown by bold black line. **C**. CIDER state predictions of each LEA proteins based on their FCRs, separated into positively and negatively charged residues. *Afr*LEA6 and *Afr*LEA3m are the only two LEA proteins that fall into their own distinct regions of the plot as weak and strong polyampholytes, respectively. *Afr*LEA1.1, *Afr*LEAI*,* and *Afr*LEA2 are predicted to be Janus sequences with independent conformational transitions.

Figure 2. NCPR distribution in *Af*LEA1.1 with a window size of five. The protein displays a distinct separation of charges based on the region of the protein. The N-terminus has a strongly positively charges region, whereas the C-terminus has two adjacent positive and negative regions.

Figure 3. MeDor-based analysis of LEA proteins from *A. franciscana*. For *Af*LEA1.1 (**A**), Pred2ary predicts β-sheets separating the termini from the central protein domain, which are shown within the boxes. The HCA shows series of small hydrophobic clusters embedded inside the regions enriched in charged and polar residues. **B**. In*Afr*LEAI, the two N-terminal internal repeats (red boxes), contain several hydrophobic clusters enriched in tyrosine, followed by a proline. The six C-terminal repeats (blue boxes) are composed of a hydrophobic cluster enriched in phenylalanine and is interrupted by a proline as well as a stretch of alternating charges enriched in a hydrophobic face of alanine residues. SMART server predicts coiled coil regions throughout the protein (black bar). **C**. The*Afr*LEA2 protein has three distinct domains (black boxes). The N-terminal domain has a likely amphipathic α-helix propensity due to the arrangement of polar and nonpolar residues and enrichment in alanine. The second domain is enriched in leucine and valine residues, with little likely structure due to enrichment of regularly spaced proline residues. The third domain begins with a hydrophobic cluster enriched in glycine. The domain is enriched in isoleucine and methionine. **D**. The*Afr*LEA3m protein has two internal repeats from positions 116 – 236 (black boxes) which are separated by a coiled-coil region predicted by SMART server (black bar). **E**. The*Afr*LEA6 protein has two internal SMP domains towards the N-terminus (black boxes) and a proline-rich intermediary domain (red box) connecting a C-terminal domain (blue box).

Figure 4. Analysis of LEA proteins from A. *franciscana* (*Afr*LEA1.1 (**A**), *Afr*LEAI (**B**), *Afr*LEA2 (**C**), and *Afr*LEA3m (**D**), *Afr*LEA6 (**E**)) by a set of per-residue disorder predictors, such as PONDR® VL3 (red), PONDR® VLXT (black), PONDR® VSL2 (green), PONDR® FIT (pink), IUPred_short (yellow), and IUPred-long (blue). Bold dashed cyan lines show the mean disorder propensity calculated by averaging disorder profiles of individual predictors. Light pink shadow around the PONDR® FIT shows error distribution. In these analyses, the predicted intrinsic disorder scores above 0.5 are considered to correspond to the disordered residues/regions, whereas regions with the disorder scores between 0.2 and 0.5 are considered flexible. The plots also include the results of functional analysis of these proteins by ANCHOR to evaluate the MoRF probability (dark pink).

Figure 5. Heliquest output of local hydropathy (red) and hydrophobic moment (blue) for *Afr*LEA1.1 (**A**), *Afr*LEAI (**B**), *Afr*LEA2 (**C**), and *Afr*LEA3m (**D**), *Afr*LEA6 (**E**).

Tables

Table 1: Classifications of LEA proteins found in the brine shrimp *Artemiafranciscana**.

Protein	**Tunnacliffe& Wise**	**Dure et al.**	**Hundertmark&Hincha**	**LEApb**	**PFAM**
***Af*LEA1.1**	Group 1	D19, D132	LEA_5	Class 5	PF00477
***Afr*LEAI**	Group 3	D7	LEA_4	Class 6	PF02987
***Afr*LEA2**	Group 3	D7	LEA_4	Class 6	PF02987
***Afr*LEA3m**	Group 3	D7	LEA_4	Class 6	PF02987
***Afr*LEA6**	Group 6	D34	SMP	Class 11	PF04927

*In this manuscript, we are using the classification scheme proposed by Tunnacliffe and Wise.

Table 2: CIDER and PONDR Parameters* of LEA Protein Sequences from *A. franciscana*.

Protein	**κ**	**FCR**	**κ/FCR**	**|MNC|**	**MNH**	**|MNC|/MNH**
***Af*LEA1.1**	0.194264	0.283333	1.458491	0.0278	0.3490	0.0797
***Afr*LEAI**	0.145081	0.29972	2.065885	0.0364	0.3858	0.0943
***Afr*LEA2**	0.079765	0.304945	3.098031	0.0137	0.4017	0.0341
***Afr*LEA3m**	0.072713	0.387681	0.187558	0.0109	0.3388	0.0322
***Afr*LEA6**	0.142528	0.206226	1.446918	0.0428	0.4536	0.0945

*The κ, FCR, and the fraction of both values. As κ increases, the likelihood of self-interaction increases, whereas if κ decreases, then the protein becomes self-repelling. Men net charge (MNC) and mean hydrophathy (MNH) were calculated based on PONDR. For more information please refer to text.

Supplemental Figure Legends

FigureS1: DisEMBL disorder predictions for *Af*LEA1.1 by loops/coil (blue), Remark465 (Green), and HotLoops (red) predictors, with dotted line thresholds for disorder with the correlating colors.

FigureS2: GlobPlot disorder prediction for AfLEA1.1 using the Remark465 propensity set. Positive slopes denote propensity towards disorder and a blue bar at the bottom of the figure denotes structural disorder prediction.

FigureS3: DISPHOS 1.3 phosphorylation prediction of *Af*LEA1.1 based on phosphorylation patterns in D. melanogaster. The phosphorylation propensity of serine residues (red triangles) and tyrosine residues (green squares) are shown for all residues above a 50% threshold. AfLEA1.1 has 100% serine phosphorylation, 14.3% tyrosine phosphorylation, and 0% threonine phosphorylation.

FigureS4: DisEMBL disorder predictions for *Afr*LEA2 by loops/coil (blue), Remark465 (Green), and HotLoops (red) predictors, with dotted line thresholds for disorder with the correlating colors.

FigureS5: DISPHOS 1.3 phosphorylation prediction of *Afr*LEA2 based on phosphorylation patterns in D. melanogaster. The phosphorylation propensity of serine residues (red triangles) and tyrosine residues (green squares) are shown for all residues above a 50% threshold.

AfrLEA2 has 43.9% serine phosphorylation, 0% tyrosine phosphorylation, and 0% threonine phosphorylation.

FigureS6: DisEMBL disorder predictions for *Afr*LEA3m by loops/coil (blue), Remark465 (Green), and HotLoops (red) predictors, with dotted line thresholds for disorder with the correlating colors.

FigureS7: GlobPlot disorder prediction for *Afr*LEA3m using the Remark465 propensity set. Positive slopes denote propensity towards disorder and a blue bar at the bottom of the figure denotes structural disorder prediction. The yellow bar at the top depicts a low-complexity region and the striped bar indicates a coiled-coil region.

FigureS8: DisEMBL disorder predictions for *Afr*LEA6 by loops/coil (blue), Remark465 (Green), and HotLoops (red) predictors, with dotted line thresholds for disorder with the correlating colors.

Sequences used for Analysis

>AfLEA1.1
MELSSSKLNRSIFKRRSKMSEQGKLSRQEAGQRGGQARAEQLGHEGYVEMGRKGGQA
RAEQLGHEGYQEMGQKGGQARAEQLGTEGYQEMGQKGGQKRAEQLGHEGYQEIGQK
GGQTRAEQLGTEGYQEMGQKGGQTRAEQLGHEGYVQMGKMGGEARKQQMSPEDYA
AMGQKGGLARQK

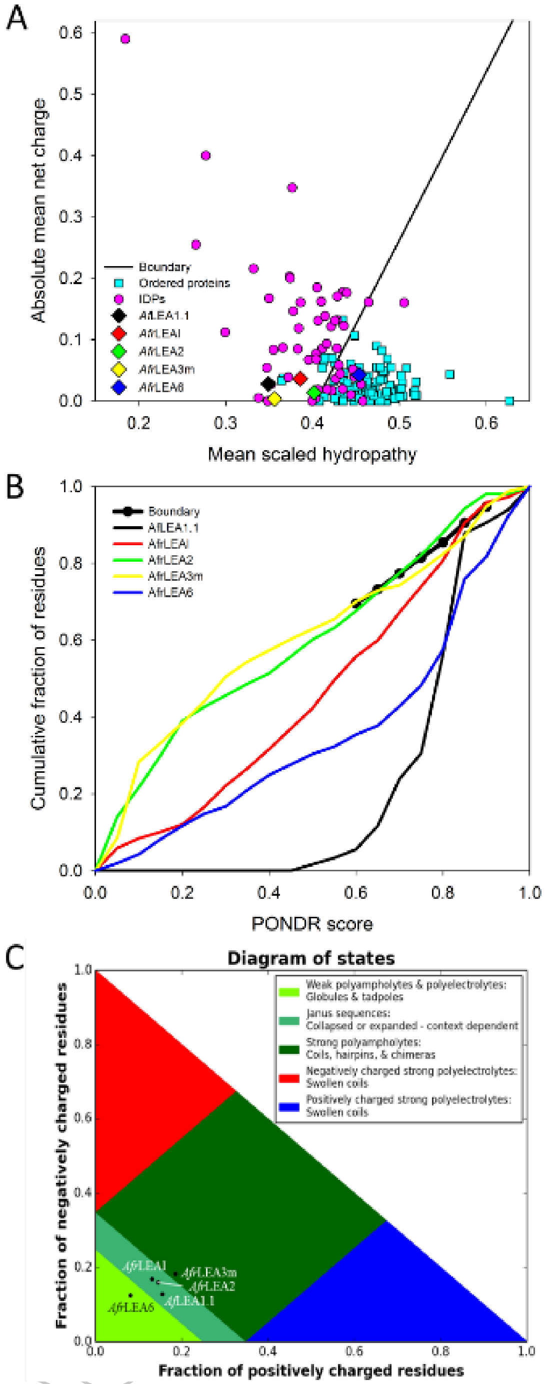
>AfrLEAI
MAEPEEPPGIYEKVKSAFVSAPDRAQEAYNQAYESARSVFDDAVRSARKMKNTAAEQA

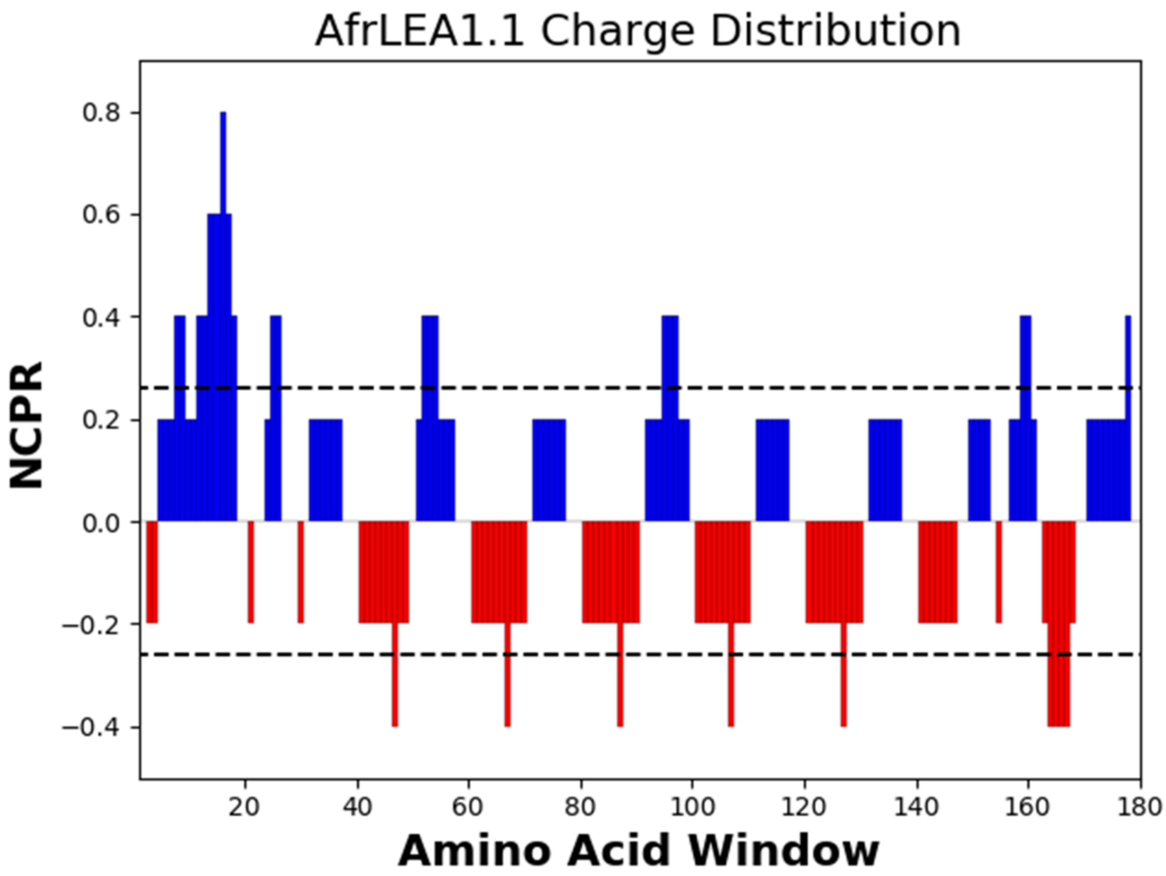
QGAYEGLKESPENLQRVTRDIYHQAQDTGKGAYETVAGSADDAYRRAQETAQAAQEQ
SKGFLNRVKDTLTAPFSSSSDQAKETYDRTKDEAQYRAQQAADAGQGFFGKVKDTITA
PFTSGYDQTQEGYERARRSAEEAAQQAADQGQTLFERAKDTITSPFSSGSEQAQESFERA
KRAAEEQVEQSKGMFQNIKGTITSPFNSAADTAKEAGQRAKKQAEEAADQSQGFMQK
VKDTVASPFLSAGEESQEAIERTKREAEEARHQGEGFLHRVADTIMHPFQSSSEQVGEA
ADRIKRGA

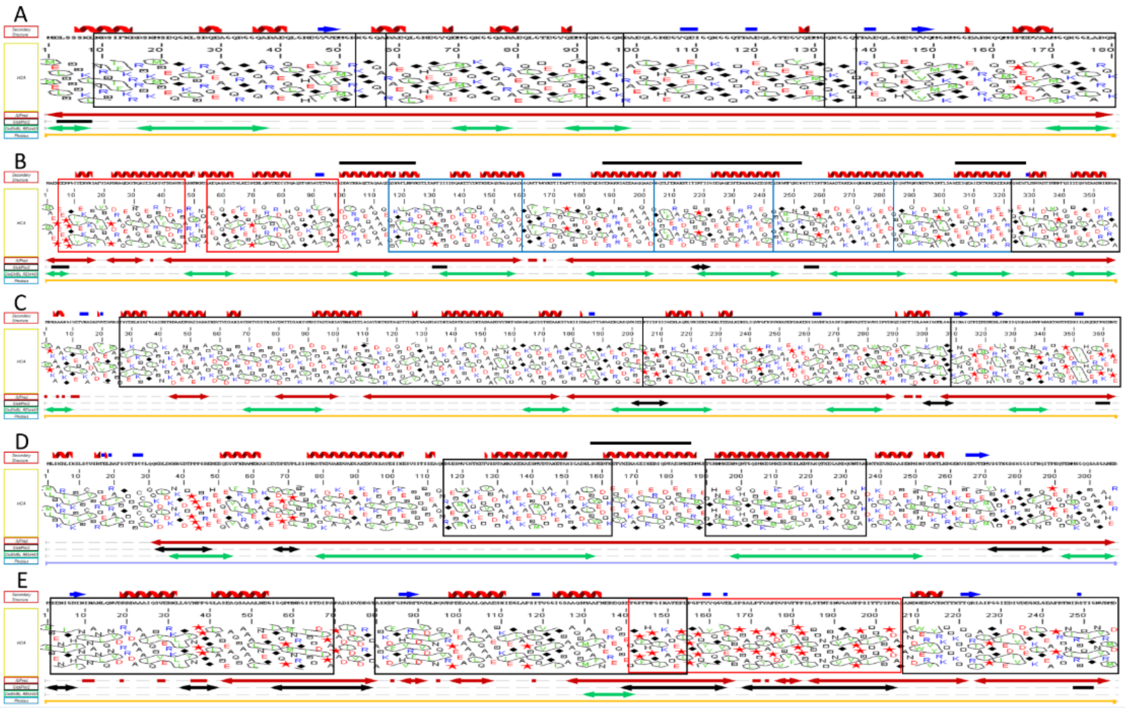
>AfrLEA2
MPKAAAKGIGETVKADADVVEGMASTGYEKLKSAFGIASNKTKDAAENVAESARATK
DYTVDSAKSAYDKTVDSTKSAYDKTTDSAKSVHDSTADTAKSAYNKATETLGSAYDK
TKDTAQSTYDQVTGAAHSAYDKTAEATKSAYDKTADAAHSVYNKTGDAGKQAYDST
KEAARSTGKSISDAAYFTGKGAERQGDQVKSELPSYSPSSSGEKLAQHLVKSEKEGKKL
TEEALKDRDLSQVPGFRSVKKAHEPDAKEDISAVDFASASPSQRKVADTEGVWSSPVDR
QESRFFSDLAGKIGDMLGGGKINAIQTPEEMDHERLIHKSSQSQVAGNVPGRAKTAWTP
EDRIILHQERFPKENPE

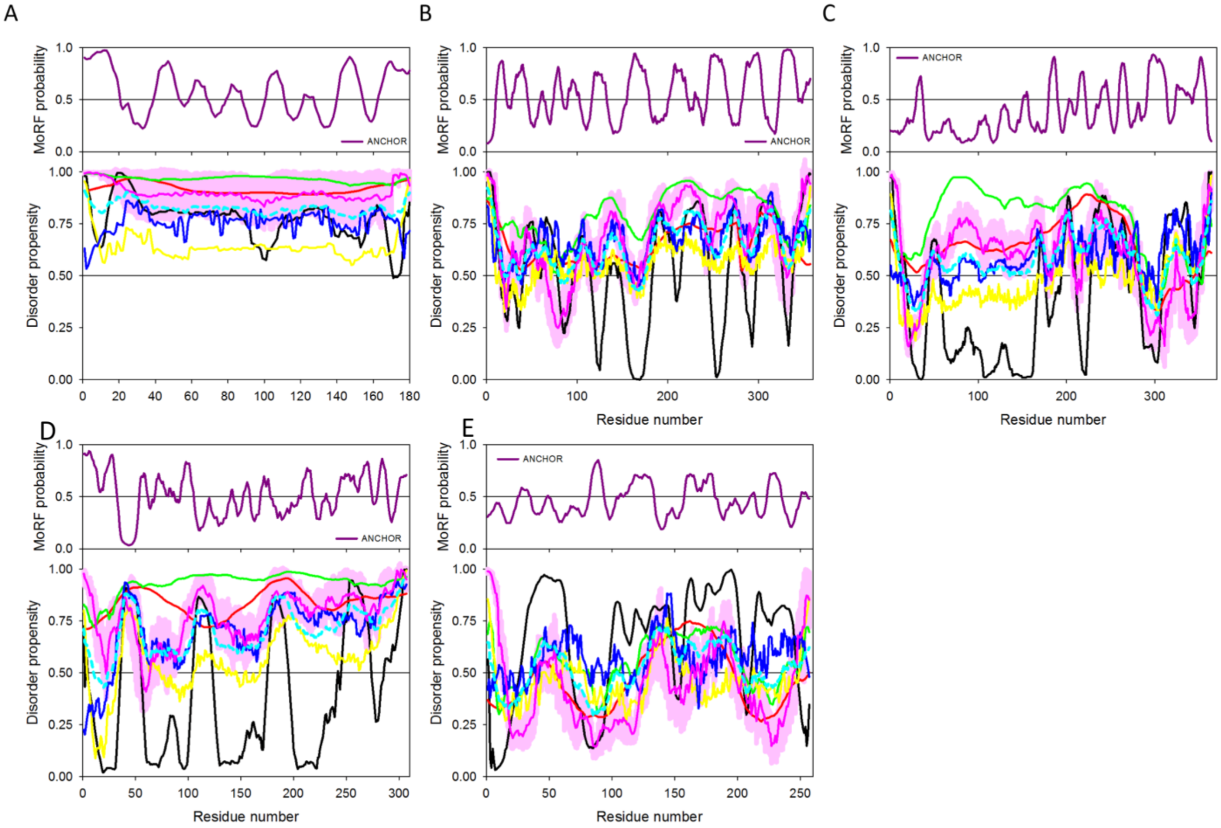
>AfrLEA3m
MLSKRLIKSLSCVSRTELRAFSGTTSCCLQQKDLDKNKGDTPPPSREHEEQEGVFKRAM
EKAKGEYDPEYPLSSSMKATKDVAKDVAEGAKEKVKSAYESIKESVSSTSSEAQNRGES
MYGKTKETVSDTANKAKEKAESMYDTAKETAKSGADKLSWEDTKETYKEKAGEIKER
IQDTAESMKERMGETGHNMKEKMQHTGQSMKEGMKESWESLKDTAKQTKEGAHDQ
WNTAKDKTKEVKDAASEKMSNSVDKTLKRGEKVSERVTEMYSGTKGDSKGGSGFNQI
TPEQTENMKGQQSASGAHER

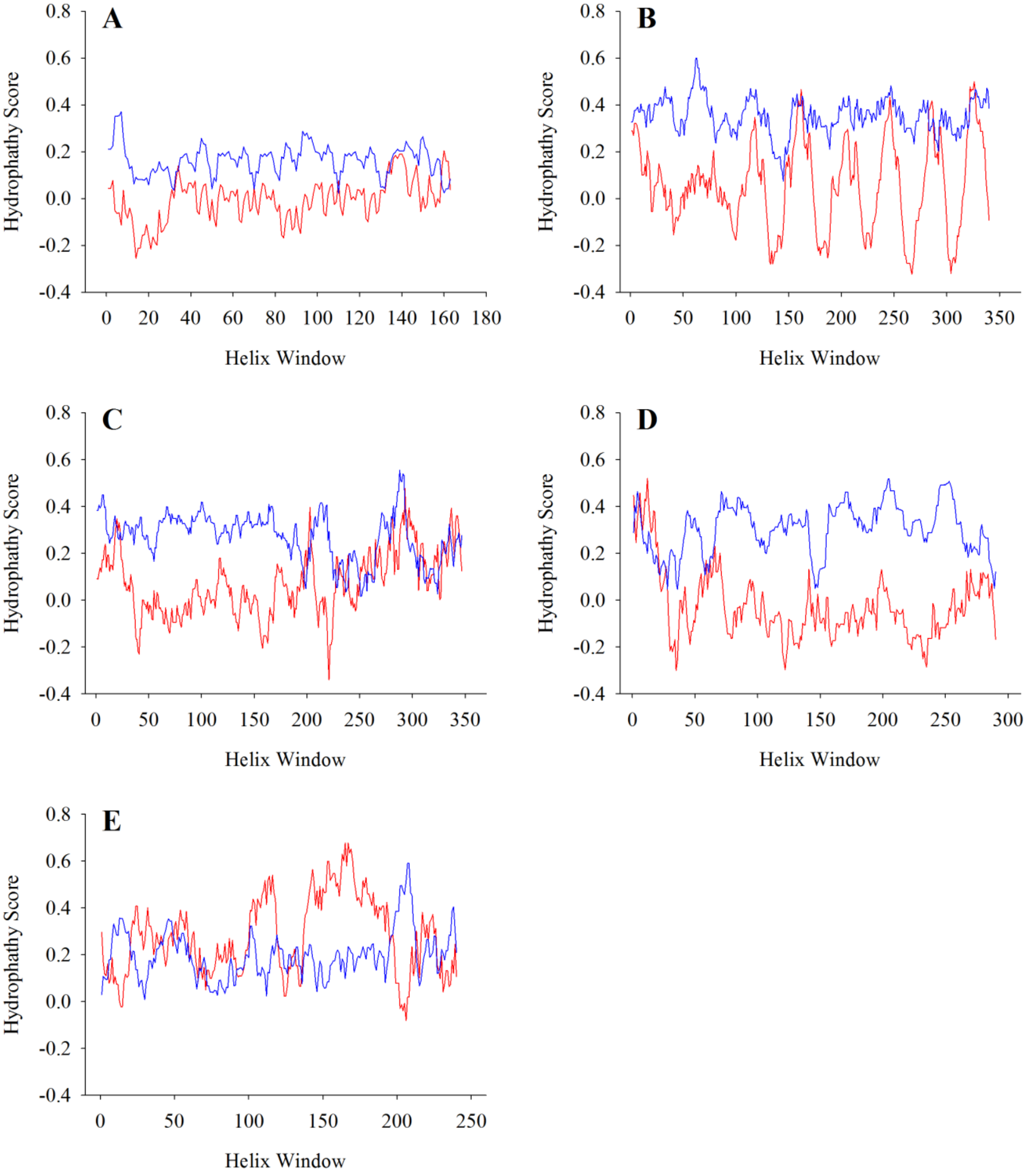
>AfrLEA6
MSENIGHININANLQNVDRRDAAAIQSVERKLLGYNPPGGLASEAQSAAALNEGIGQPM
NRGISTDIPAPADIDVDRGTASKDFGHVRFDVDLNQVRPEEAAALQAAESKIEGLAPSIT
VGGIGSAAQSMAAFNEREQSETGPFHPGIKATEPLPGPTYYQGVELSPSALPTYAPDVSV
FPPSLSTNTSNVGAVPPSITTYSPDAGANDWERVYRKTTKTTQRIAIPGGIEDIVDEGKLG
EAPRTNIRS

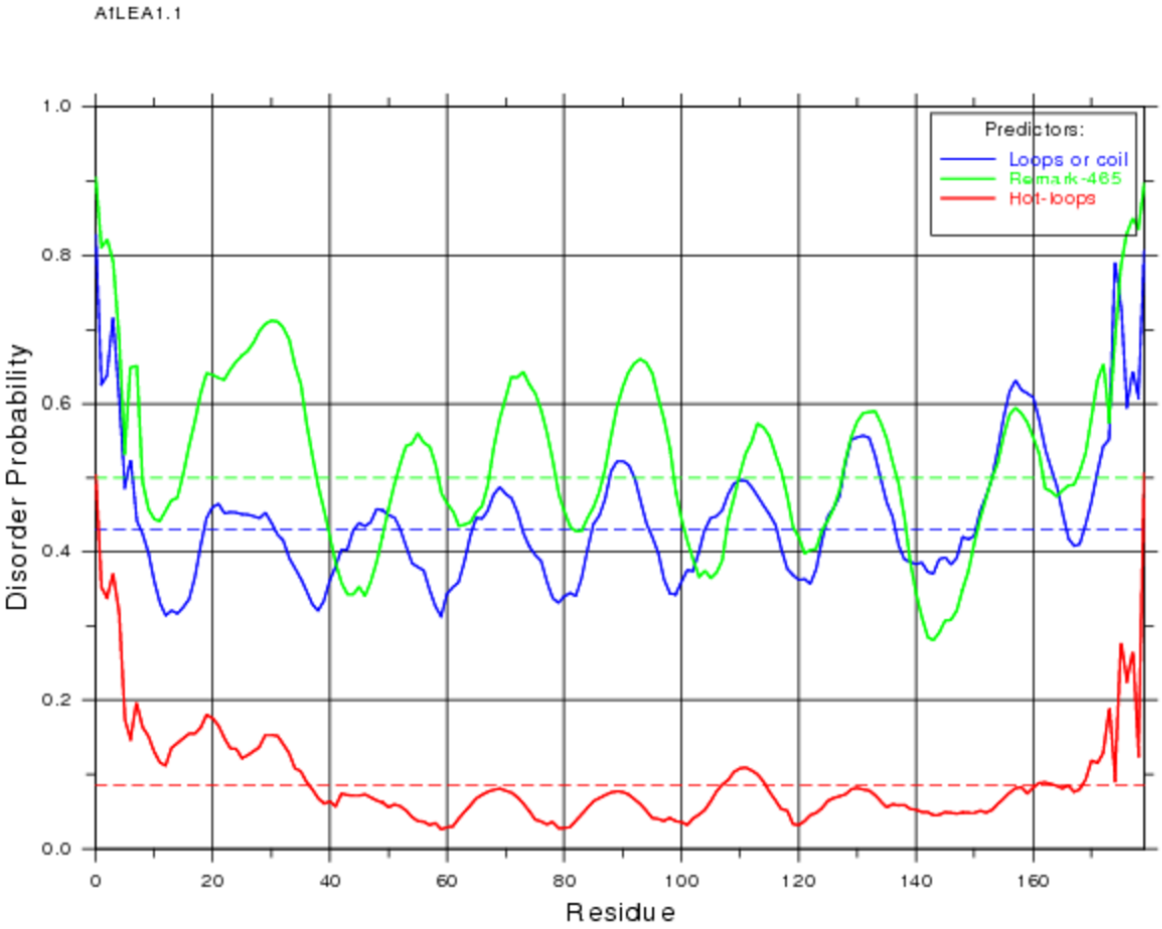

A
Absolute mean net charge
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.2
0.3
0.4
0.5
0.6
Mean scaled hydropathy
Boundary
Ordered proteins
IDPs
AfLEA1.1
AfrLEAI
AfrLEA2
AfrLEA3m
AfrLEA6
B
Cumulative fraction of residues
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
PONDR score
Boundary
AfLEA1.1
AfrLEAI
AfrLEA2
AfrLEA3m
AfrLEA6
C
Diagram of states
Fraction of negatively charged residues
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of positively charged residues
Weak polyampholytes & polyelectrolytes: Globules & tadpoles
Janus sequences: Collapsed or expanded - context dependent
Strong polyampholytes: Coils, hairpins, & chimeras
Negatively charged strong polyelectrolytes: Swollen coils
Positively charged strong polyelectrolytes: Swollen coils
AfrLEA1
AfrLEA3m
AfrLEA2
AfLEA1.1
AfrLEA6

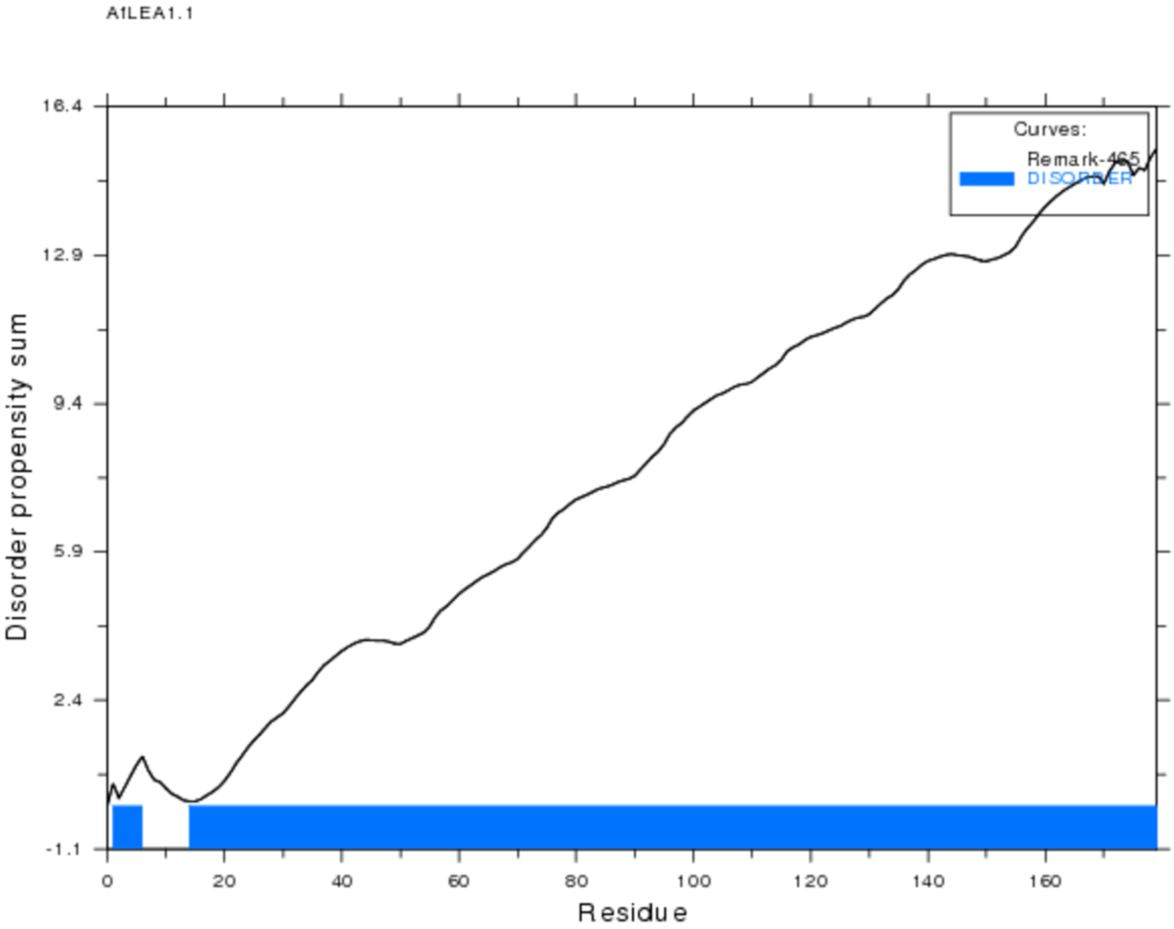

AfrLEA1.1 Charge Distribution
NCPR
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
20
40
60
80
100
120
140
160
180
Amino Acid Window

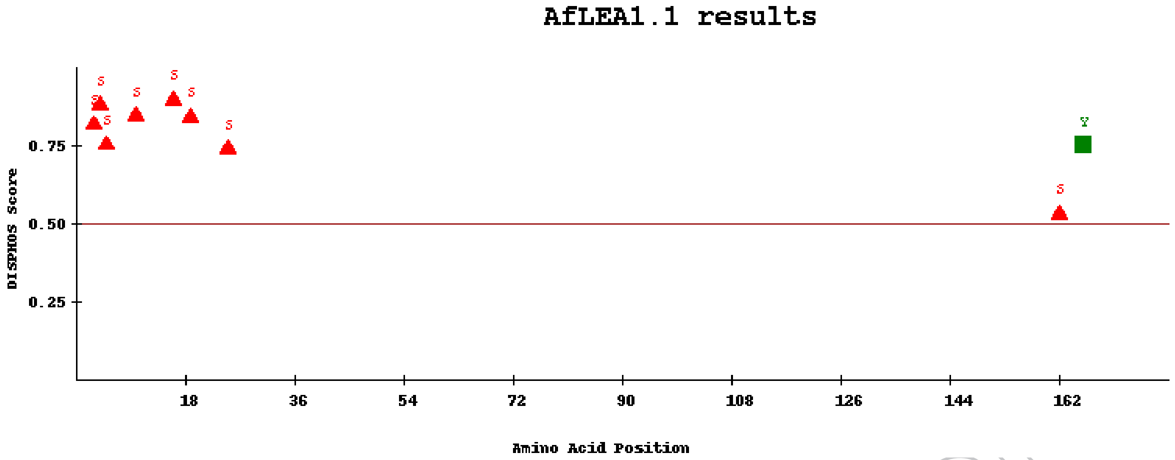

A
B
C
D
E


A
B
C
D
E
MoRF probability
1.0
0.5
0.0
ANCHOR
Disorder propensity
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
100
120
140
160
180
100
200
300
50
150
250
Residue number

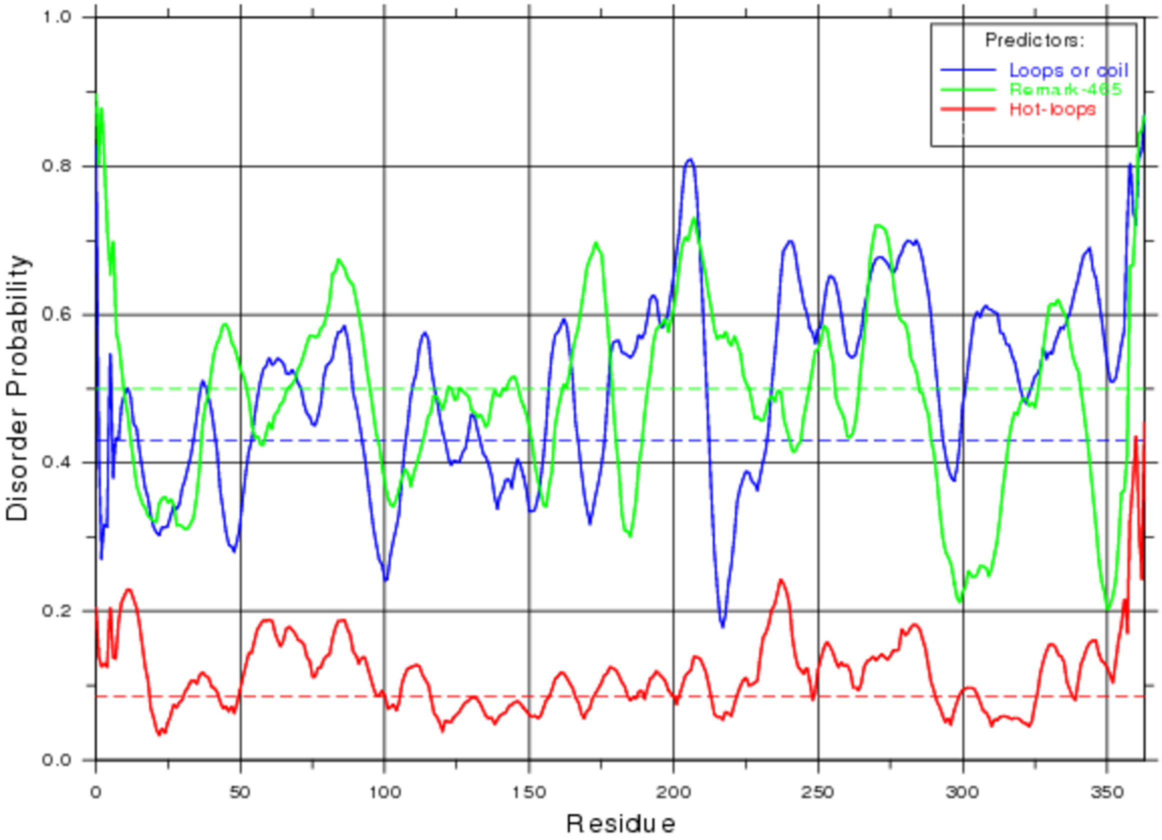

A
Hydropathy Score
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0
20
40
60
80
100
120
140
160
180
Helix Window
B
Hydropathy Score
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
200
250
300
350
Helix Window
C
Hydropathy Score
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
200
250
300
350
Helix Window
D
Hydropathy Score
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
200
250
300
Helix Window
E
Hydropathy Score
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
200
250
Helix Window

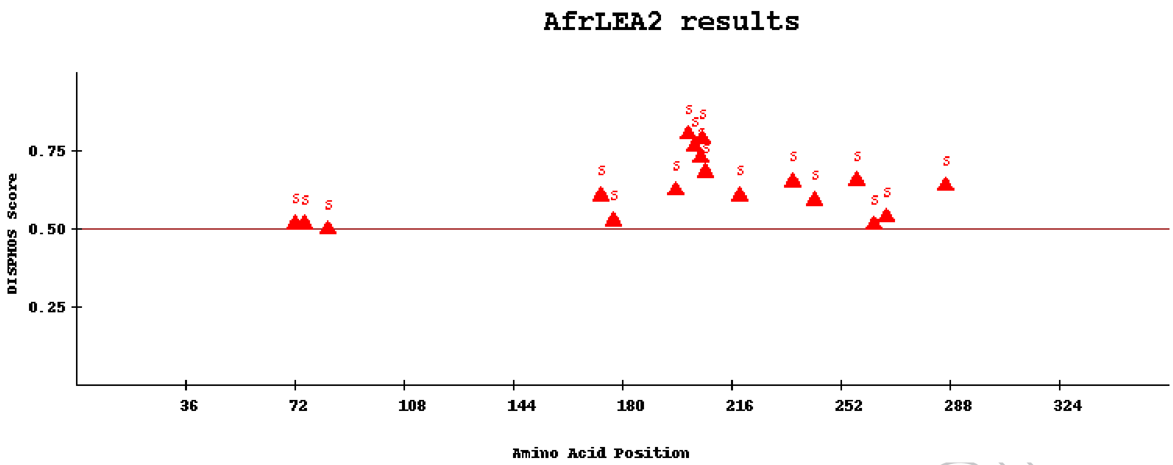

AfLEA1.1
Disorder Probability
1.0
0.8
0.6
0.4
0.2
0.0
Residue
0
20
40
60
80
100
120
140
160
Predictors:
Loops or coil
Remark-465
Hot-loops

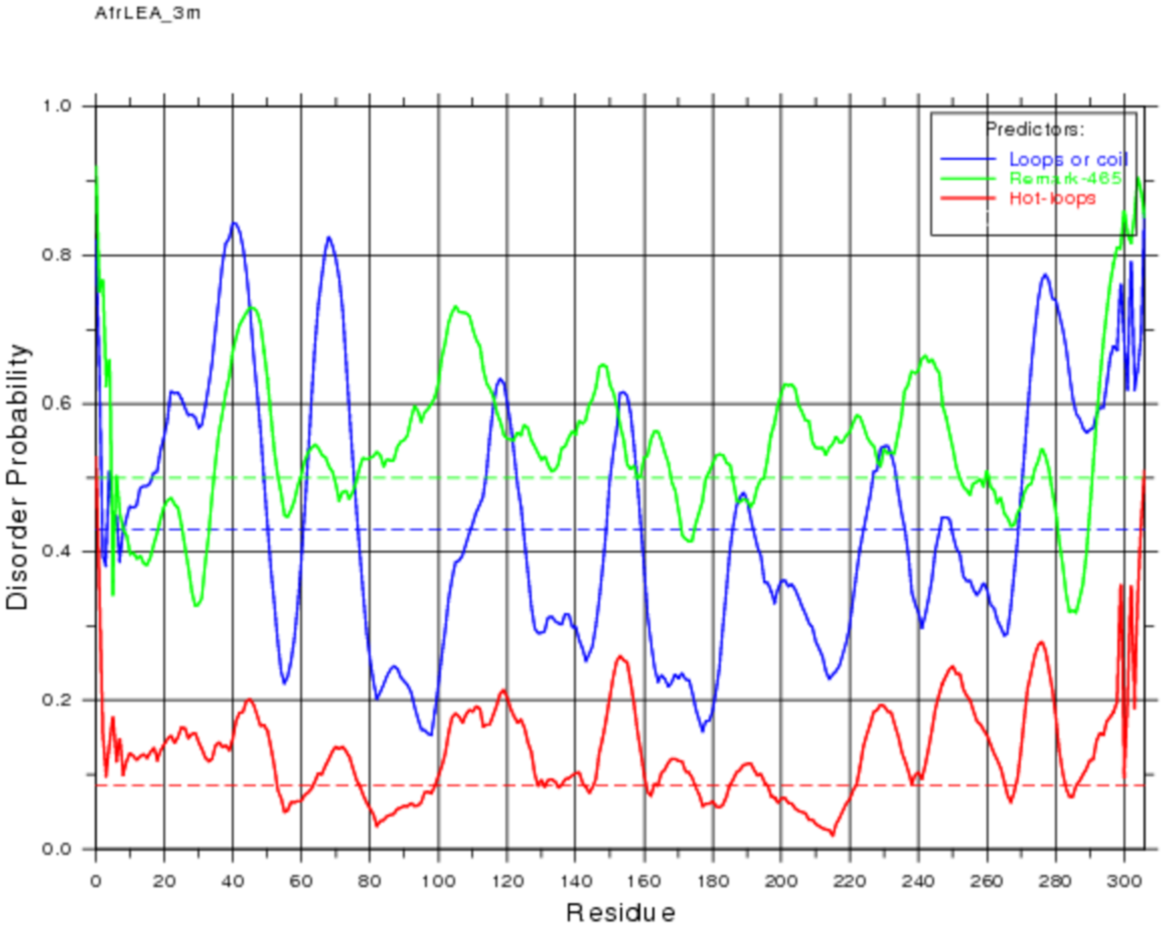

AtLEA1.1
Curves:
Remark-465
DISORDER
Disorder propensity sum
16.4
12.9
9.4
5.9
2.4
-1.1
0
20
40
60
80
100
120
140
160
Residue

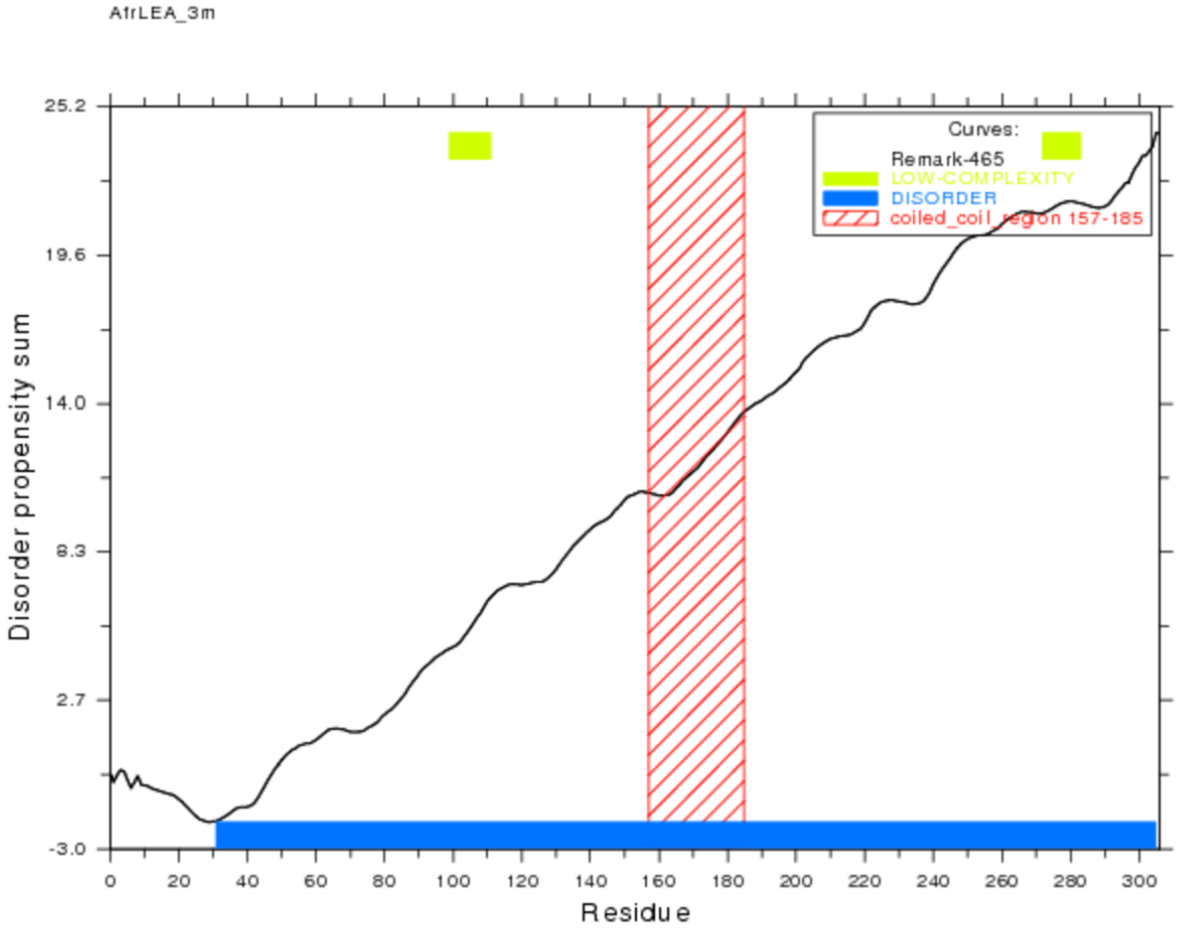

AfLEA1.1 results
DISPHOS Score
0.75
0.50
0.25
S
S
S
S
S
S
S
S
Y
18
36
54
72
90
108
126
144
162
Amino Acid Position

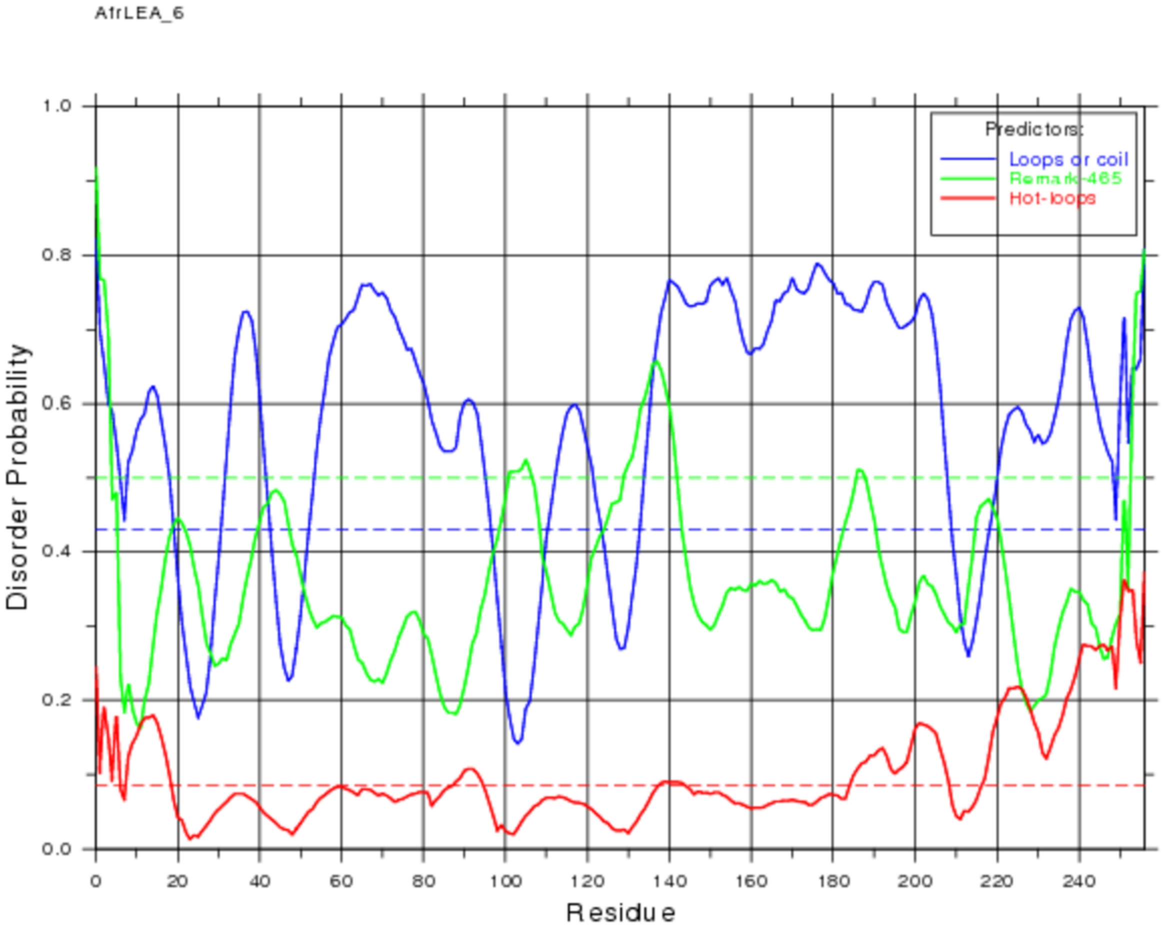
AfrLEA_2


1.0
0.8
0.6
0.4
0.2
0.0
Disorder Probability
0
50
100
150
200
250
300
350
Residue
Predictors:
Loops or coil
Remark-465
Hot-loops


AfrLEA2 results
DISPHOS Score
0.75
0.50
0.25
36
72
108
144
180
216
252
288
324
Amino Acid Position
S


AfrLEA_3m
Predictors:
Loops or coil
Remark-465
Hot-loops
Disorder Probability
1.0
0.8
0.6
0.4
0.2
0.0
Residue
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300


AfrLEA_3m
Curves:
Remark-465
🟩 LOW-COMPLEXITY
🟦 DISORDER
▨ coiled_coil_region 157-185
Disorder propensity sum
25.2
19.6
14.0
8.3
2.7
-3.0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
Residue


AfrLEA_6
Predictors:
Loops or coil
Remark-465
Hot-loops
Disorder Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
140
160
180
200
220
240
Residue