A comprehensive BIAPSS platform for the physicochemical featurization of phase separating proteins

Vladimir Uversky (✉ vuversky@health.usf.edu)

University of South Florida <https://orcid.org/0000-0002-4037-5857>

Aleksandra Badaczewska-Dawid

Iowa State University <https://orcid.org/0000-0003-1832-4364>

Davit Potoyan

Iowa State University

Resource

Keywords:

Posted Date: December 22nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1053840/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

The liquid-liquid phase separation (LLPS) of biological macromolecules has emerged as a foundational mechanism underlying the formation of a myriad of membraneless organelles (MLOs), such as stress granules, transcription factor condensates, and chromatin compartments. A molecular grammar of sequences, which would enable a quantitative prediction and understanding of protein phase separation from first principles is currently missing. A major challenge in the field is the sparsity of bioinformatics data and the lack of computational, data-driven tools for biophysical and statistical analysis of proteins capable of phase separation. Here we present the utility of web applications framed within a novel open-source platform for BioInformatic Analysis of liquid-liquid Phase-Separating protein Sequences, <https://biapss.chem.iastate.edu/>. BIAPSS combines high-throughput interactive data analytics of physicochemical and evolutionary features with a comprehensive repository of bioinformatic data for on-the-fly research of the sequence-dependent properties of proteins with known LLPS behavior. To facilitate exploration of the services and provide the interpretation guideline, we present two attention-getting case studies of FUS and hnRNPD. This should help the LLPS community uncover the nature of interactions driving the formation of membraneless organelles.

Introduction

In the past few years, the liquid-liquid phase separation (LLPS) of biomolecules has become a unifying physical mechanism for understanding principles of intracellular compartmentalization, formation of membraneless organelles, and gene regulation^{1–5}. The ability of proteins to phase separate appears to be encoded primarily in the peculiarities of their primary sequences, which often contain low complexity regions and are enriched in charged and multivalent interaction centers^{6–8}. While some general sequence trends have emerged, the quantitative aspects of how amino acid sequences encode and decode phase separation still remain largely unknown^{9–11}. This is because many different combinations of relevant interactions seem to be contributing to phase separation without anyone being universally necessary¹². As a consequence (with a few exceptions^{13–16}) mostly case-by-case studies of different sequences are performed, with the broader context of many findings, including their statistical significance remaining unknown.

In recent years, several databases have emerged which collect LLPS-related protein sequence data and metadata, with prominent examples being PhaSepDB¹⁴, PhaSePro¹⁵, LLPSeDB¹⁶, and DrLLPS¹⁷. These databases collect and annotate partially overlapping sets of phase-separating protein sequences including data on experimental conditions and significant annotations. In particular, PhaSePro, LLPSeDB, and a subset of PhaSepDB contain manually curated proteins, which are recognized for driving the formation of subcellular compartments.

Accumulation of high-quality datasets is certainly a necessary condition for making progress towards uncovering driving forces of protein phase separation. However, one needs a biophysically motivated computational infrastructure to be able to harness the data from carefully and manually curated sets of

phase separating proteins for revealing molecular features that determine protein phase separation. To this end, we have developed a novel web platform named BIAPSS: BioInformatic Analysis of liquid-liquid Phase-Separating protein Sequences, available online at <https://biapss.chem.iastate.edu/>. BIAPSS combines a high-throughput interactive deep sequence analysis with a comprehensive pre-parsed bioinformatics database containing a wide array of physicochemical and evolutionary features relevant for low-complexity, disordered, and ordered proteins. This platform provides scientists working in the field of biomolecular condensates with a versatile tool for rapid and on-the-fly deep statistical analysis of LLPS-driver protein sequences.

Here we present the utility of SingleSEQ applications implemented within the BIAPSS web platform, which employ a residue-resolution biophysical analyzer for interrogating individual protein sequences. In addition to usage guidelines and analytic purposes of individual applications, we discuss in detail case studies for two attention-getting LLPS proteins, FUS and hnRNPD, respectively.

Results

The overview of the BIAPSS web platform

BIAPSS is a novel web platform, which has two major components (i) A comprehensive bioinformatics database of phase separating proteins, which includes a large number of deeply parsed biophysically motivated features, robust predictions of structural properties, detection of Pfam functional domains, and the other shorter sequential and structural motifs. (ii) A collection of high-throughput, interactive data analysis and visualization applications that enable rapid explorative inferences and comparisons of the data from the web browser of convenience. We note that the database of BIAPSS encompasses a superset of over 500 unique LLPS sequences extracted and parsed from manually curated primary LLPS databases. The layout and main functionalities of the SingleSEQ module of BIAPSS services are summarized in Figure 1.

The BIAPSS web application consists of five tabs. The Home provides a high-level overview of the features of the BIAPSS services. The SingleSEQ tab houses the applications for the exploration of single sequence characteristics. There, users can correlate regions implicated in phase separation with a large array of compositional and physicochemical attributes, structural properties (e.g., secondary structure, solvent accessibility, contacts), low-complexity, disorder regions, evolutionary context, location of domains, and various sequential or structural motifs. The MultiSEQ tab provides the user with a set of web applications for a broad array of statistics on a superset of LLPS sequences. Through MultiSEQ users can explore the regularities and specific trends in the sequence universe of phase separating proteins, for instance, an overall sequence composition compared to various reference datasets, including sequence complexity with highlighted specific amino acid enrichments. Besides showing the overall content of polar/hydrophobic/charged/aromatic residues, MultiSEQ also allows visualizing charge

decoration parameters that recently emerged as an informative measure for electrostatic interactions of intrinsically disordered proteins (IDPs) ¹⁸. All of the pre-processed bioinformatics data that are passed to BIAPSS visualizers are available for download (these options are included in the DOWNLOAD tab). The available data include raw predictions pre-calculated using the well-established tools, as well as the findings based on our deep statistical analysis. Finally, the DOCS tab provides detailed documentation related to the description of all the components of BIAPSS.

SingleSEQ web applications for robust analysis of an individual LLPS sequence

The SingleSEQ component of BIAPSS (https://biapss.chem.iastate.edu/single_seq.html) is designed for a high throughput biophysical and evolutionary analysis, visualization, and interactive exploration of sequence features (Figure 2). A comprehensive featurization of sequences was done, which provides users with informative metrics with known impact on protein phase-separation and conformational propensities. Below we summarize the capabilities of each web application followed by their performance in case studies on two proteins, FUS and hnRNPA1. For more details on the tools and approaches used, see the Methods section.

1.1 Entry Summary and Annotation web application

The first interactive application of SingleSEQ analysis provides a compact one-page summary for a single LLPS protein sequence. A brief overview of all key parameters helps to get an insight into the essence of individual sequence characteristics. The extended cross-references section shows coverage in external resources and integrates available metadata for a particular LLPS protein. Over there, the user can (i) easily follow the corresponding entries in the primary LLPS-related databases, such as PhaSepDB ¹⁴, DrLLPS ¹⁷, LLPSTDB ¹⁶, and PhaSePro ¹⁵, (ii) explore the principal annotation data and functional information in UniProt resource ¹⁹, (iii) identify manually curated intrinsically disordered regions in DisProt database ²⁰, (iv) check subcellular location through Compartments ²¹ or display it on the enlarged image, (v) see experimental mapping in The Human Protein Atlas ²². If an experimental structure is known, (vi) the Structure button links to the PDBe-KB repository ²³, otherwise the computationally modeled structure is accessible via the MODBASE database ²⁴ and/or SWISS-MODEL server ²⁵, and the newly provided AlphaFold database ²⁶. There, you can also find PSPredictor score estimating the propensity to phase behavior ²⁷, empirically determined LLPS-related region(s), and PubMed identifiers referring to the literature.

1.2 Sequence Complexity & Composition web application

The application allows (i) exploring the sequence composition (amino acid content and patterning along the sequence), and (ii) detecting the low-complexity regions (LCRs), accompanied with overall sequence information content (Shannon entropy). Since LCRs are known to compose imperfect sequence repeats,

and polar or hydrophobic blocks of residues, the detection and systematic analysis of these regions can facilitate the identification of the nature of multiple weak interaction motifs in phase separating proteins.²⁸ The AA content section provides the detailed contents and decoration of 20 biogenic amino acids (by default, those into which the sequence is enriched) in a given LLPS protein referenced to the user-selected benchmark dataset of various groups of proteins. Finally, the low-complexity section provides along the sequence the merged and unified regions of low-complexity and Shannon information content calculated on the fly based on the user-selected length of the frame.

1.3 Patterns of Chemical Decoration web application

The Chemical Patterns application allows combining several key biophysical properties for an individual LLPS-driver protein sequence, all within a single multi-row chart. This includes chemical properties, such as polarity, hydrophobicity, aromaticity, and charge. Then, the robust consensus predictions of secondary structure, solvent accessibility, structural disorder, and low-complexity regions are provided. The position-based vertical alignment of all features gives a deeper glimpse into the correlation of various chemical decorations and structural properties. Furthermore, the all-in-one chart facilitates the detection of regions sensitive for key interactions. For example, parallel comparison of specific patterns of charged or aromatic residues and unique insertion in a structurally disordered region helps to identify hot spot residues or short motifs, which may play an important role in the LLPS behavior.

1.4 Domains, Motifs, Repeats web application

This multi-row application is designed to detect Pfam functional domains and offers a broad screening of short motifs, both sequential and structural, in the individual sequences of phase-separating proteins. The findings are displayed on the background of regions experimentally confirmed as LLPS-related and/or predicted to be structurally disordered or low-complexity. In the subsequent rows the additional custom feature patterns can be displayed, e.g., physicochemical decoration to facilitate the identification of the nature of interactions or evolutionary conservation of short motifs within intrinsically disordered fragments that clearly points to their significance for the regulatory function, hence may also be essential for the collective formation of MLOs. Additionally, short fragments frequently repeated along an individual sequence are detectable, including random and tandem repeats. The detailed information on the location, motif class, and instance are displayed on the interactive labels.

1.5 Sequence Conservation web application

The conservation strength of a sequence region may be a consequence of its functional or structural importance. Therefore, the unique insertions, deletions, or substitutions, identified within the LLPS protein sequence may prove to influence the chemical properties of the polypeptide chain towards multivalent interactions. The Sequence Conservation application delivers (i) the multiple sequence alignment (MSA) of the LLPS sequence against the reference databases of protein sequences, (ii) the consensus sequence profile, (iii) and evolutionary conservation (of LOGO type) at the level of individual amino acids. Also, the identification of functional domains along the query sequence is provided. For each individual domain, the Pfam seed-MSA can be loaded, which significantly increases the reliability of evolutionary

conservation in this region of the LLPS protein. The MSA section contains HMMER-aligned ²⁹ protein sequences colored by the chemical character of amino acids and sorted by the E-value and score. Each position in the sequence is annotated interactively by corresponding positions in the query sequence, consensus profile, and full-MSA. The labels also include amino acid, confidence level, and secondary structure assignment taken from known PDBs.

1.6 The Secondary Structure web application

Conformational preferences of local sequence fragments are conditioned by the local composition of amino acids through the formation of regular patterns of hydrogen bonds along the protein backbone. Recent findings imply that some intrinsically disordered fragments of the sequence tend to get structurally organized into steric zippers or aromatic-rich kinked β -sheets ^{30–33}. The most advanced sequence-based secondary structure predictors provide not only the binary assignment of ordered regions but also the probability fractions for the helical (H), extended (E), or coiled (C) structures of the individual amino acids along the sequence ^{34,35}. This opens up the possibility to determine factors supporting multivalent interactions, as well as detecting regions showing some tendency to form ordered structure, which can be acting as a switch depending on the environment conditions and molecular crowding. The Secondary Structure application presents the sequence-based predictions and agreed consensus of secondary structure elements assigned along with the selected LLPS sequences. In the last section, the detailed probabilities to form each type of HEC element are attributed along the sequence with amino acid resolution.

1.7 The Solvent Accessibility web application

The Solvent Accessibility application presents the sequence-based prediction of structurally buried or exposed regions along the protein sequence. Since the structural properties of many phase-separating proteins are hardly graspable, the analysis of surface solvent accessibility may support, to a certain extent, the identification of regulatory regions prone to form temporarily stabilized interactions at the contact interface ^{36,37}. For compatibility and user-friendliness, the overall layout of the interactive graph is similar to that of the Secondary Structure app and contains three main sections: (i) on the top, the agreed consensus of solvent accessibility predictions provided at residue-resolution level, (ii) followed in the next few rows by the original results derived and unified from the individual methods, (ii) below which there is a section of detailed probabilities for exposure or burial of individual residues along the sequence. The interactive annotations include residue index, amino acid type, and solvent accessibility in three-state notation or particular probabilities.

1.8 The Structural Disorder web application

Identification of low-complexity and unstructured fragments within proteins has recently gained prominence, following the discovery that the liquid-liquid phase separation emerges as a common functioning mechanism for partially or fully disordered proteins ^{1,2,4,5}. From these findings, we know that protein sequences not only encode for well-ordered tertiary structures but can also explicitly encode intrinsic disorder. The Structural Disorder application provides predictions of disordered regions resulting

from the benchmark of numerous widely used methods. Within the app, we provide both the detailed probability distributions of disorder along the sequence and the agreed binary (order/disorder) outcome of all tools assigned per amino acid position.

1.9 Contact Map web application

The Contact Map application provides sequence-based predictions of spatial contacts between residues distant in a protein sequence. The most powerful contact predictors, such as those selected for this work, employ the convolutional neural networks, machine learning, and evolutionary coupling techniques to prepare the list of residue pairs occurring close enough in space with a certain probability. An in-depth analysis of the contacts predicted for the LLPS proteins, despite their largely disordered nature, can help identify regions essential for maintaining structure, dynamics, and function, and may indicate a lead relevant to phase behavior.

3. Discussion

Case Study I: FUS, LLPS regulated in the context-dependent tuning of preferred forces

Fused in sarcoma (FUS) is one of the early discovered biological systems undergoing self-organization by liquid-liquid phase separation (LLPS) ³⁸. Since then, the protein has been the subject of extensive experimental and computational research to understand the molecular mechanisms and interactions driving this phenomenon. FUS can be found in the BIAPSS service by the UniProt identifier (P35637), gene (FUS), or using the “*RNA-binding*” search key. The summary page contains a high-quality image of the experimentally confirmed cellular location (left panel in Figure 3A). Due to its multifunctionality in RNA processing, FUS is mostly observed in the nucleus ³⁹. In physiological conditions, the low levels of the protein are distributed in the cytoplasm ⁴⁰, where FUS transports and manages RNA through the dynamic liquid-like subcellular compartments, such as ribonucleoprotein or stress granules ⁴¹. However, the cytoplasmic concentration of FUS significantly increases when noxious mutations lead to aggregation ⁴². This progressively aberrant process is manifested by neurodegenerative diseases in humans ⁴². Although plenty of accumulated evidence points to the influence of distinct factors on the cellular behavior of FUS, its primary sequence still holds many cues. To frame the physicochemical properties of full-length FUS, we used the analytical approach offered by the SingleSEQ module of BIAPSS.

The average metrics indicate that the 526 residues long sequence of FUS contains over 80% disorder and only 8% order. The solvent accessibility predictions show the same aspect ratio between exposure and burial. The contents of aromatic, hydrophobic, polar, and charged residues are 10%, 42%, 40%, and 17% respectively, with a slight excess of positive charge. Such a rough overview described by a set of averages gives some general insight about the protein properties but conceals some local distributions that are important for the identification of the preferential interactions. Therefore, we conduct a detailed

analysis of the composition and complexity of the FUS sequence and present the resulting patterns in Figure 3B. Compared to any reference set of proteins, this one is extremely enriched in glycine making up nearly 1/3 of the full sequence. Another 20% of the amino acid content consists primarily of serine and glutamine. Although the dominant content of these three amino acids suggests generally low-complexity of the sequence, their distribution along the sequence is strongly heterogeneous. Indeed, the calculated low information content of the sequence is mainly localized around protein terminals and clearly corresponds to three fragments with high glycine concentration (LCR2: 164-267, LCR3: 370-420, LCR4: 454-507). These regions also exclusively accumulate total arginines, which together with glycine form a series of RGG repeating motifs known to bind RNA specifically ⁴⁴. Both serine and glutamine are mostly localized at the N-terminus, being more clearly clustered within LCR1 (1-163). LCR1 additionally gathers 24/35 available tyrosines, and thus, it has visibly distinct enrichment (SQYG) known to occur in prion-like domains (PLD) ⁴⁵. Using *Domains, Motifs, Repeats* application we also found that the remaining compositionally more complex regions of C-terminus (I287-L365 and R422-D453), match the PF00076 and PF00641 Pfam domains, i.e., the RNA recognition motif (RRM) and RNA binding zinc finger (ZnF), respectively. The robust predictions (for details see Methods) unanimously show that RRM is a well-folded FUS domain, while the other fragments remain disordered. The seed MSAs prepared for FUS within *Sequence Conservation* application further confirms that both domains are evolutionarily conserved members of Pfam families: RRM_1 and zf-RanBP, respectively (see bottom rows in Figure 3B).

The visual inspection of the amino acid content and distribution of FUS allows us to identify and isolate specific regions in the protein (Figure 3C). Furthermore, we have performed a physicochemical featurization of these segments, which reveal preferred interactions when coupled with biomolecular conditionals known from experiments. The recent experimental reports showed that the isolated prion-like domain (PLD, 1-214 or even 1-163) can undergo self-organization forming liquid droplets when kept in high protein levels or high salt concentrations ^{46,47}. This N-terminal fragment is enriched in amino acids whose side chains are multivalent, as shown in Figure 3C. Thus, the dense pattern of polarity comes from enrichment in S, Q, Y, where Y, Q, and G also provide π -electron centers for π - π -stacking. Most of them are also able to be both donors and acceptors of side-chain protons for hydrogen bonding (HB). In line with this, the intermolecular interaction profiles derived from simulations of the 120-163 region indicated the most frequent contacts between QQ > QY > YY > SY and other pairs of enriched amino acids ⁴⁸. All of these observations suggest that homotypic phase separation of wild-type PLD monomers is driven by balanced contributions from hydrogen bonding and π -stacking. Indeed, several mutagenesis studies showed that Y $\rightarrow$ A substitution disrupts phase separation by removal both components of the interaction, while Y $\rightarrow$ F mutants are significantly more aggregation-prone, due to strengthening of binding via tighter hydrophobic F- π -stacking at the cost of losing HB contributions of polar tyrosine ^{46,48}. It is also worth noting that the PLD region is completely deficient of positive charge with a minor net charge per residue of -0.01 (M1-S165: -0.012 and M1-G212: -0.024), which place it within the weak polyelectrolytes region on the CIDER diagram (right panel in Figure 3A) ⁴³. However, an excess of serine and threonine in this region

provides an ability to introduce a strongly negative charge through multiple phosphorylations. After phosphorylation, the dominant force becomes electrostatic repulsion, which is known to disrupt both phase separation and aggregation³⁸. The central region of PLD (39-95) was proposed as the core of aberrant fibrils, which in solid-state form structured cross- β -sheets³⁸. The same structural properties have not been unambiguously confirmed in the condensed phase of liquid-liquid mixing. Undoubtedly, however, our algorithms detected along this region structural motifs known as low-complexity, amyloid-like, reversible, folded segments (LARKS)³¹. In our analysis, the most effective predictors of structural properties showed for these motifs some tendency toward extended secondary structure and a slightly increased probability of burial (bottom panel in Figure 3C). Interestingly, the prediction in 8-letter notation detected a turn or bend within each of the structural motifs, which explains their flexible nature. These findings together with ambiguous experimental results may suggest some variations of structural state in the PLD core and specifically the disorder-to-order transition driven by biomolecular conditionals.

The remaining part of the FUS sequence, referred to as the C-terminus, contains two well-known domains (RRM and ZnF) and three glycine-arginine-rich regions (GARs). All components are significant players in binding RNA. Zinc finger supports only the recognition of specific GGU motif, while RRM domain and RGG repeats are universal towards a variety of RNAs⁴⁹. Both folded domains of FUS are much less polar than PLD, as seen from the BIAPSS-based physicochemical featurization in Figure 3C. They also have a lower content of side chains that are able to engage in π -stacking or hydrogen bonding. However, the charged residues are pretty abundant in the composition of RRM and ZnF, which explains the functional role of electrostatic interactions towards the binding of nucleic acids or stabilizing folds via salt bridges^{50,51}.

All three GARs are the least polar regions of the protein (see Figure 3C). The dense patterning of hydrophobicity arises from glycine excess. The rich π -electron-containing systems, other than aromatic side chains, originate mainly from the abundance of the arginine's guanidino group. Arginine is also a source of excess positive charge at the C-terminus. The experimental studies consistently confirm that the isolated C-terminus does not undergo phase separation⁴⁶. However, liquid-liquid droplets rapidly occur when mixed with N-terminal monomers⁴⁶. Moreover, the LLPS of full-length wild-type FUS is more robust than heterotypic mixing of N- and C-terminals and homotypic self-assembling of N-terminal monomers⁴⁶. This suggests the higher priority of cation- π (R-Y) stacking over π - π (Y-Y) stacking, while both are reinforced by hydrogen bonds. Another experimental study showed that R $\rightarrow$ K mutants, who no longer have the ability of π - π -stacking but retain charge, can still undergo phase separation. In turn, R $\rightarrow$ A substitutions prevent phase separation because they lose the π -system, cation, and ability of side-chain hydrogen bonding. Interestingly, the recent report indicates that stacking interactions, including cation- π (e.g., RY, KF) and especially π - π (e.g., YY and RY, and even RQ), are most robust over a wide range of salt concentrations⁴⁷. The hydrophobic contribution from π -electron-containing systems becomes the main force that strengthens the contact in high salt. In these conditions, the screening of usually dominant electrostatic contributions is significant. Surprisingly, changing the partitioning of the different forces

makes the interaction of the two positively charged arginines attractive under these conditions⁴⁷. The set of diverse chemical groups in arginine is a unique feature among the other amino acid side chains. With its high reactivity, the need for precise regulation comes, and so arginine can be tuned to a preferred state by posttranslational methylation. Thus, under physiological conditions, FUS is highly methylated⁴⁶. This limits self-assembly via interactions with tyrosine and promotes a functional role of intermolecular interactions with other proteins and nucleic acids. Therefore, phase separation and gelation of FUS can increase by hypomethylation of arginines within RGG-rich regions or insertion of additional ones in the C-terminus⁴⁶. All of these findings come together to demonstrate the significant role of the arginine side chain in phase separation. Tyrosine and glutamine are similarly relevant. They also contain multifunctional chemical groups that make them reactive and multivalent. These features aid in context-dependent tuning between preferred forces of interactions. They can work synergistically or alternatively. And their regulation depends on environmental conditions, the state of posttranslational modifications, and the presence of binding partners.

Case Study II: hnRNPD - cellular behavior regulated by the priority of interactions

Heterogeneous nuclear ribonucleoproteins are a large and diverse family of proteins that play active roles at every stage of RNA regulation⁵². An important member of this family, the heterogeneous ribonucleoprotein D-like sequence of humans (hnRNPD), occurs in the nucleus in three isoforms generated by alternative splicing. These isoforms differ by the presence or absence of structurally flexible terminal regions and, depending on this, phase separate (DL1, full-length), aggregate (DL2, missing N-terminus), or remain soluble (DL3, missing both N- and C-terminus)⁵³. It is therefore thought that a unique combination of these regions is used for regulating the function and cellular behavior (phase state, mobility, location) of the isoforms. It is known that various short sequential or structural motifs, especially located in the disordered regions, enhance multivalent interactions and tune the ability of proteins to form higher-order assemblies^{31,54}. It has been shown that even a single mutation in a key region of hnRNPD can significantly promote an aberrant aggregation⁵³. To understand the differences in self-assembly properties of different isoforms we carry out deep biophysical and bioinformatic analysis of full-length hnRNPD by using BIAPSS SingleSEQ web applications.

The protein can be found in the BIAPSS online repository by the UniProt identifier (O14979), gene (hnRNPD), or common name. The brief summary shows that the entry is covered in most primary LLPS databases except PhaSePro. Interestingly we find that PSPredictor score = 0.27 which indicates a lack of phase separation contrary to the experimental reports. The simplified biophysical characteristic of the sequence shows 49% disorder and 31% order consisting mostly of the helical structure. Solvent accessible surface area analysis shows that 64% of the structure is likely exposed to the solvent. The contents of polar, hydrophobic, and aromatic residues are 30%, 42%, 12%, respectively.

Using the *Composition and Complexity app* allows us to carry out a more in-depth analysis of the patterning of LLPS promoting residues in the full-length sequence of hnRNPD. We find the protein is enriched with polar (Y, Q), positively charged (R) residues, and π -electron-containing systems (Y, R, Q, G) at the expense of structure promoting hydrophobic ones (I, L, V) with respect to the all, globular proteins, SwissProt sequence set, and even the disordered proteins collected in DisProt or LLPS-driving sequences collected in BIAPSS (Figure 4A). In particular, the agreed consensus of low-complexity (LCR) predictors shows that the G > Y > Q clusters are mostly localized in the C-terminal region, with enrichment typical for prion-like domains. The N-terminus on the other hand is enriched with P and R residues. The K residues are mostly localized in the linker of two folded domains located in the central part of the sequence. In line with these observations, the Shannon entropy reveals that the C-terminal region is much lower in complexity than N-terminus, however, their amino acid composition indicates that both are still classified as LCRs. Additionally, robust predictions of structural disorder (taken from the *Structural Disorder app*) further confirm that both LCRs contain intrinsically disordered regions (IDRs). Specifically, the IDR1:M1-Q145 is located on the N-terminus, while IDR2:V313-F350 and IDR3:S397-Y420 are located on the C-terminus. Thus we can already see that there is significant asymmetry between the terminals of the protein which is in harmony with experiments showing distinct physicochemical and functional implications for these two LCRs ⁵³.

For a more in-depth analysis of the sequence-specific nature of the two terminal LCRs and their physicochemical properties we next turn to the *Chemical Properties Patterns app* (Figure 4B). For the C-terminal region, we find a compact cluster of mostly polar and π -stacking (including many aromatic) residues which are capable of being a donor and/or acceptor of side-chain hydrogen bonds thereby supporting attractive interactions and explaining the observed burial tendency in this region. More interestingly, when combining in Figure 4B the detailed outcome of structural properties and detected short motifs, we can easily correlate the burial bias with a predicted propensity for extended structure. This region is located in between two IDRs (N347 to Y396) and aligns with the enrichment by Y that contains two LARKS-like structural motifs (**GYDYTG**: 376-381 and **GYADYS**: 392-397). Furthermore, the unstructured and exposed flanking IDRs (IDR2 and IDR3) correlate well with the G-rich region (G323-G411), where several glycine-arginine motifs (GARs) have been detected. The LARKS motifs are known to shape cross- β -sheets through weak attractive interactions, including π -stacking, van der Waals and hydrogen bonding, which are all possible according to the amino acid composition of the C-terminus ³¹. Notably, despite its apparent similarity to the steric zipper, the LARKS motif has a smaller buried interface and looser packing of the side chains between the mating sheets, which means weaker binding force and higher flexibility. In this case, the enrichment with flexible glycine, aromatic tyrosine, and polar glutamine effectively support multivalent interactions since all of them can engage in π -stacking and form hydrogen bonding (Y and Q also through their side chains being both donor and acceptor of a proton). Generally the resultant of binding forces and environmental conditions tend to decide whether such low-complexity regions, decorated with some functional motifs, remain disordered or undergo phase separation or even

aggregate. Our findings are also very much in line with ATR-FTIR experiments of Taylor-Ventura et al., showing that a sequence isoform DL2, possessing only C-terminal of two LCRs, leads to aggregates containing intermolecular β -sheet⁵³

For the N-terminal region, we find that the content of residues able to π -stacking and side-chain hydrogen bonding is much smaller, which together with the enrichment in positively charged residues (mostly R) and proline (known to be a secondary structure breaker) leads to a reduction in the number of possible short-range interactions. As a result, this region seems unstructured (with a slight tendency towards short α -helices) and almost completely solvent-exposed consistent with the predictions of local structure and solvent accessibility (Figure 4B). When analyzing the predicted contact map on the left panel in Figure 4C (taken from the *Contact Map app*), again we can easily point out the differences between the LCRs at the two ends of the protein. In particular, despite the enhanced intra-regions contacts for domains, the algorithms detected Q320-G410 inter-region with the slight probability of internal contacts (score ~ 0.35 while contact threshold is ≥ 0.5), which covers almost the entire C-terminal LCR, while the N-terminus appears without any internal contacts. Interestingly, using the outcome of the *Structural Disorder app* (row "BINDING" in Figure 4B), we see that the N-terminal region despite being structurally featureless is detected by the ANCHOR method as a protein binding region (A20-I95) with three picks above 0.8 score. This region coincides with a region enriched by arginine, which is known to be abundant in the LLPS-driving motifs and RNA binding regions⁴⁴. Therefore, we suspect that the experimentally observed dissolution of DL1 droplets in presence of highly concentrated RNA⁵³, is most likely due to strong electrostatic attraction between negatively charged nucleic acid and positively charged arginines. When the environment lacks ionic interaction partners, the DL1 isoform undergoes phase separation driven by weaker but still favorable π -cation interactions between C-terminal tyrosines and N-terminal arginines. And finally, the absence of the R-rich N-terminus in the isoform DL2 leads to fibril formation, driven by π - π -stacking of C-terminal tyrosines in two monomers.

Thus, using a programming analogy, we can say that the final behavior of hnRNPD isoforms appears as the result of a biomolecular algorithm composed of a series of if-then-else conditionals, depending on the presence of N-terminal and environmental conditions. Furthermore, the experiments confirmed that substitution of cation (R $\rightarrow$ K) or/and aromatic system (Y $\rightarrow$ F) significantly weakens the phase separation of DL1⁵³, which we think, points to the stronger matching of the R-Y pair. This complementarity results from the high cooperativity of multivalent interactions, in addition to the π -cation component also includes π - π -stacking (CZ=NH1 π -electrons donated by arginine compared to lysine) and side-chain hydrogen bonding (polar hydroxyl donated by tyrosine compared to nonpolar phenylalanine and multiple protons bonded to nitrogens donated by arginine).

Finally, while understandably the focus tends to be on the low-complexity regions decorated with some binding and/or switching structural properties motifs, it is nevertheless also important to consider the well-folded domains, especially in cases like hnRNPD where they play a mediator between the two IDRs.

The *Domains, Motifs, Repeats app* reveals the two HMMER-detected Pfam domains (both with ID: PF00076) located in M151-D219 and V235-I305 regions (Figure 4A, 4B). The presence of the domains is also well captured by the contact map prediction which reveals details about likely binding activities going on between distinct regions of the sequence (left panel in Figure 4C). Both, predictors of the secondary structure and known structures (PDBs) assigned to the sequences used in the Pfam seed-MSA, indicate that each domain has well-structured globular $\alpha+\beta$ topology (composed of 2 α -helices and single β -sheet), which perfectly explains the solubility of the third isoform (DL3) lacking both flexible ends. The list of corresponding PDBs can be found on the interactive labels assigned to each sequence in the Pfam seed-MSA in the *Sequence Conservation & MSA app* (e.g., 1P1T, residues: 18-88), and/or by following the link of the “STRUCTURE” button on the main page of the SingleSEQ tab. Specifically, for human hnRNPD (see the right panel in Figure 4C), the BIAPSS provides a direct link to the entry in the novel AlphaFold database collecting the high-quality predictions of structures that proved to be competitive with the experimental ones ²⁶.

Discussion

Case Study I: FUS, LLPS regulated in the context-dependent tuning of preferred forces

Fused in sarcoma (FUS) is one of the early discovered biological systems undergoing self-organization by liquid-liquid phase separation (LLPS) ³⁸. Since then, the protein has been the subject of extensive experimental and computational research to understand the molecular mechanisms and interactions driving this phenomenon. FUS can be found in the BIAPSS service by the UniProt identifier (P35637), gene (FUS), or using the “*RNA-binding*” search key. The summary page contains a high-quality image of the experimentally confirmed cellular location (left panel in **Figure 3A**). Due to its multifunctionality in RNA processing, FUS is mostly observed in the nucleus ³⁹. In physiological conditions, the low levels of the protein are distributed in the cytoplasm ⁴⁰, where FUS transports and manages RNA through the dynamic liquid-like subcellular compartments, such as ribonucleoprotein or stress granules ⁴¹. However, the cytoplasmic concentration of FUS significantly increases when noxious mutations lead to aggregation ⁴². This progressively aberrant process is manifested by neurodegenerative diseases in humans ⁴². Although plenty of accumulated evidence points to the influence of distinct factors on the cellular behavior of FUS, its primary sequence still holds many cues. To frame the physicochemical properties of full-length FUS, we used the analytical approach offered by the SingleSEQ module of BIAPSS.

The average metrics indicate that the 526 residues long sequence of FUS contains over 80% disorder and only 8% order. The solvent accessibility predictions show the same aspect ratio between exposure and burial. The contents of aromatic, hydrophobic, polar, and charged residues are 10%, 42%, 40%, and 17%

respectively, with a slight excess of positive charge. Such a rough overview described by a set of averages gives some general insight about the protein properties but conceals some local distributions that are important for the identification of the preferential interactions. Therefore, we conduct a detailed analysis of the composition and complexity of the FUS sequence and present the resulting patterns in **Figure 3B**. Compared to any reference set of proteins, this one is extremely enriched in glycine making up nearly 1/3 of the full sequence. Another 20% of the amino acid content consists primarily of serine and glutamine. Although the dominant content of these three amino acids suggests generally low-complexity of the sequence, their distribution along the sequence is strongly heterogeneous. Indeed, the calculated low information content of the sequence is mainly localized around protein terminals and clearly corresponds to three fragments with high glycine concentration (LCR2: 164-267, LCR3: 370-420, LCR4: 454-507). These regions also exclusively accumulate total arginines, which together with glycine form a series of RGG repeating motifs known to bind RNA specifically ⁴⁴. Both serine and glutamine are mostly localized at the N-terminus, being more clearly clustered within LCR1 (1-163). LCR1 additionally gathers 24/35 available tyrosines, and thus, it has visibly distinct enrichment (SQYG) known to occur in prion-like domains (PLD) ⁴⁵. Using *Domains, Motifs, Repeats* application we also found that the remaining compositionally more complex regions of C-terminus (I287-L365 and R422-D453), match the PF00076 and PF00641 Pfam domains, i.e., the RNA recognition motif (RRM) and RNA binding zinc finger (ZnF), respectively. The robust predictions (for details see Methods) unanimously show that RRM is a well-folded FUS domain, while the other fragments remain disordered. The seed MSAs prepared for FUS within *Sequence Conservation* application further confirms that both domains are evolutionarily conserved members of Pfam families: RRM_1 and zf-RanBP, respectively (see bottom rows in **Figure 3B**).

The visual inspection of the amino acid content and distribution of FUS allows us to identify and isolate specific regions in the protein (**Figure 3C**). Furthermore, we have performed a physicochemical featurization of these segments, which reveal preferred interactions when coupled with biomolecular conditionals known from experiments. The recent experimental reports showed that the isolated prion-like domain (PLD, 1-214 or even 1-163) can undergo self-organization forming liquid droplets when kept in high protein levels or high salt concentrations ^{46,47}. This N-terminal fragment is enriched in amino acids whose side chains are multivalent, as shown in **Figure 3C**. Thus, the dense pattern of polarity comes from enrichment in S, Q, Y, where Y, Q, and G also provide π -electron centers for π - π -stacking. Most of them are also able to be both donors and acceptors of side-chain protons for hydrogen bonding (HB). In line with this, the intermolecular interaction profiles derived from simulations of the 120-163 region indicated the most frequent contacts between QQ > QY > YY > SY and other pairs of enriched amino acids ⁴⁸. All of these observations suggest that homotypic phase separation of wild-type PLD monomers is driven by balanced contributions from hydrogen bonding and π -stacking. Indeed, several mutagenesis studies showed that Y $\rightarrow$ A substitution disrupts phase separation by removal both components of the interaction, while Y $\rightarrow$ F mutants are significantly more aggregation-prone, due to strengthening of binding via tighter hydrophobic F- π -stacking at the cost of losing HB contributions of polar tyrosine ^{46,48}. It is also worth noting that the PLD region is completely deficient of positive charge with a minor net charge per residue of -0.01 (M1-S165: -0.012 and M1-G212: -0.024), which place it within the weak polyelectrolytes region on

the CIDER diagram (right panel in **Figure 3A**)⁴³. However, an excess of serine and threonine in this region provides an ability to introduce a strongly negative charge through multiple phosphorylations. After phosphorylation, the dominant force becomes electrostatic repulsion, which is known to disrupt both phase separation and aggregation³⁸. The central region of PLD (39-95) was proposed as the core of aberrant fibrils, which in solid-state form structured cross- β -sheets³⁸. The same structural properties have not been unambiguously confirmed in the condensed phase of liquid-liquid mixing. Undoubtedly, however, our algorithms detected along this region structural motifs known as low-complexity, amyloid-like, reversible, folded segments (LARKS)³¹. In our analysis, the most effective predictors of structural properties showed for these motifs some tendency toward extended secondary structure and a slightly increased probability of burial (bottom panel in **Figure 3C**). Interestingly, the prediction in 8-letter notation detected a turn or bend within each of the structural motifs, which explains their flexible nature. These findings together with ambiguous experimental results may suggest some variations of structural state in the PLD core and specifically the disorder-to-order transition driven by biomolecular conditionals.

The remaining part of the FUS sequence, referred to as the C-terminus, contains two well-known domains (RRM and ZnF) and three glycine-arginine-rich regions (GARs). All components are significant players in binding RNA. Zinc finger supports only the recognition of specific GGU motif, while RRM domain and RGG repeats are universal towards a variety of RNAs⁴⁹. Both folded domains of FUS are much less polar than PLD, as seen from the BIAPSS-based physicochemical featurization in **Figure 3C**. They also have a lower content of side chains that are able to engage in π -stacking or hydrogen bonding. However, the charged residues are pretty abundant in the composition of RRM and ZnF, which explains the functional role of electrostatic interactions towards the binding of nucleic acids or stabilizing folds via salt bridges^{50,51}.

All three GARs are the least polar regions of the protein (see **Figure 3C**). The dense patterning of hydrophobicity arises from glycine excess. The rich π -electron-containing systems, other than aromatic side chains, originate mainly from the abundance of the arginine's guanidino group. Arginine is also a source of excess positive charge at the C-terminus. The experimental studies consistently confirm that the isolated C-terminus does not undergo phase separation⁴⁶. However, liquid-liquid droplets rapidly occur when mixed with N-terminal monomers⁴⁶. Moreover, the LLPS of full-length wild-type FUS is more robust than heterotypic mixing of N- and C-terminals and homotypic self-assembling of N-terminal monomers⁴⁶. This suggests the higher priority of cation- π (R-Y) stacking over π - π (Y-Y) stacking, while both are reinforced by hydrogen bonds. Another experimental study showed that R $\rightarrow$ K mutants, who no longer have the ability of π - π -stacking but retain charge, can still undergo phase separation. In turn, R $\rightarrow$ A substitutions prevent phase separation because they lose the π -system, cation, and ability of side-chain hydrogen bonding. Interestingly, the recent report indicates that stacking interactions, including cation- π (e.g., RY, KF) and especially π - π (e.g., YY and RY, and even RQ), are most robust over a wide range of salt concentrations⁴⁷. The hydrophobic contribution from π -electron-containing systems becomes the main force that strengthens the contact in high salt. In these conditions, the screening of usually dominant electrostatic contributions is significant. Surprisingly, changing the partitioning of the different forces makes the interaction of the two positively charged arginines attractive under these conditions⁴⁷. The set

of diverse chemical groups in arginine is a unique feature among the other amino acid side chains. With its high reactivity, the need for precise regulation comes, and so arginine can be tuned to a preferred state by posttranslational methylation. Thus, under physiological conditions, FUS is highly methylated⁴⁶. This limits self-assembly via interactions with tyrosine and promotes a functional role of intermolecular interactions with other proteins and nucleic acids. Therefore, phase separation and gelation of FUS can increase by hypomethylation of arginines within RGG-rich regions or insertion of additional ones in the C-terminus⁴⁶. All of these findings come together to demonstrate the significant role of the arginine side chain in phase separation. Tyrosine and glutamine are similarly relevant. They also contain multifunctional chemical groups that make them reactive and multivalent. These features aid in context-dependent tuning between preferred forces of interactions. They can work synergistically or alternatively. And their regulation depends on environmental conditions, the state of posttranslational modifications, and the presence of binding partners.

Case Study II: hnRNPD - cellular behavior regulated by the priority of interactions

Heterogeneous nuclear ribonucleoproteins are a large and diverse family of proteins that play active roles at every stage of RNA regulation⁵². An important member of this family, the heterogeneous ribonucleoprotein D-like sequence of humans (hnRNPD), occurs in the nucleus in three isoforms generated by alternative splicing. These isoforms differ by the presence or absence of structurally flexible terminal regions and, depending on this, phase separate (DL1, full-length), aggregate (DL2, missing N-terminus), or remain soluble (DL3, missing both N- and C-terminus)⁵³. It is therefore thought that a unique combination of these regions is used for regulating the function and cellular behavior (phase state, mobility, location) of the isoforms. It is known that various short sequential or structural motifs, especially located in the disordered regions, enhance multivalent interactions and tune the ability of proteins to form higher-order assemblies^{31,54}. It has been shown that even a single mutation in a key region of hnRNPD can significantly promote an aberrant aggregation⁵³. To understand the differences in self-assembly properties of different isoforms we carry out deep biophysical and bioinformatic analysis of full-length hnRNPD by using BIAPSS SingleSEQ web applications.

The protein can be found in the BIAPSS online repository by the UniProt identifier (O14979), gene (hnRNPD), or common name. The brief summary shows that the entry is covered in most primary LLPS databases except PhaSePro. Interestingly we find that PSPredictor score = 0.27 which indicates a lack of phase separation contrary to the experimental reports. The simplified biophysical characteristic of the

sequence shows 49% disorder and 31% order consisting mostly of the helical structure. Solvent accessible surface area analysis shows that 64% of the structure is likely exposed to the solvent. The contents of polar, hydrophobic, and aromatic residues are 30%, 42%, 12%, respectively.

Using the *Composition and Complexity app* allows us to carry out a more in-depth analysis of the patterning of LLPS promoting residues in the full-length sequence of hnRNPD. We find the protein is enriched with polar (Y, Q), positively charged (R) residues, and π -electron-containing systems (Y, R, Q, G^[1]) at the expense of structure promoting hydrophobic ones (I, L, V) with respect to the all, globular proteins, SwissProt sequence set, and even the disordered proteins collected in DisProt or LLPS-driving sequences collected in BIAPSS (**Figure 4A**). In particular, the agreed consensus of low-complexity (LCR) predictors shows that the G > Y > Q clusters are mostly localized in the C-terminal region, with enrichment typical for prion-like domains. The N-terminus on the other hand is enriched with P and R residues. The K residues are mostly localized in the linker of two folded domains located in the central part of the sequence. In line with these observations, the Shannon entropy reveals that the C-terminal region is much lower in complexity than N-terminus, however, their amino acid composition indicates that both are still classified as LCRs. Additionally, robust predictions of structural disorder (taken from the *Structural Disorder app*) further confirm that both LCRs contain intrinsically disordered regions (IDRs). Specifically, the IDR1:M1-Q145 is located on the N-terminus, while IDR2:V313-F350 and IDR3:S397-Y420 are located on the C-terminus. Thus we can already see that there is significant asymmetry between the terminals of the protein which is in harmony with experiments showing distinct physicochemical and functional implications for these two LCRs⁵³.

^[1] It is proved that glycine can stack (due to the lack of the side chain) via π -electrons from a peptide bond and hydrogen bonding via backbone carbonyl or amide.⁵⁵

For a more in-depth analysis of the sequence-specific nature of the two terminal LCRs and their physicochemical properties we next turn to the *Chemical Properties Patterns app* (**Figure 4B**). For the C-terminal region, we find a compact cluster of mostly polar and π -stacking (including many aromatic) residues which are capable of being a donor and/or acceptor of side-chain hydrogen bonds thereby supporting attractive interactions and explaining the observed burial tendency in this region. More interestingly, when combining in **Figure 4B** the detailed outcome of structural properties and detected short motifs^[1], we can easily correlate the burial bias with a predicted propensity for extended structure. This region is located in between two IDRs (N347 to Y396) and aligns with the enrichment by Y that contains two LARKS-like structural motifs (**GYDYTG**: 376-381 and **GYADYS**: 392-397). Furthermore, the unstructured and exposed flanking IDRs (IDR2 and IDR3) correlate well with the G-rich region (G323-G411), where several glycine-arginine motifs (GARs) have been detected. The LARKS motifs are known to shape cross- β -sheets through weak attractive interactions, including π -stacking, van der Waals and hydrogen bonding, which are all possible according to the amino acid composition of the C-terminus³¹. Notably, despite its apparent similarity to the steric zipper, the LARKS motif has a smaller buried interface and looser packing of the side chains between the mating sheets, which means weaker binding force and

higher flexibility. In this case, the enrichment with flexible glycine, aromatic tyrosine, and polar glutamine effectively support multivalent interactions since all of them can engage in π -stacking and form hydrogen bonding (Y and Q also through their side chains being both donor and acceptor of a proton). Generally the resultant of binding forces and environmental conditions tend to decide whether such low-complexity regions, decorated with some functional motifs, remain disordered or undergo phase separation or even aggregate. Our findings are also very much in line with ATR-FTIR experiments of Taylor-Ventura et al., showing that a sequence isoform DL2, possessing only C-terminal of two LCRs, leads to aggregates containing intermolecular β -sheet⁵³

For the N-terminal region, we find that the content of residues able to π -stacking and side-chain hydrogen bonding is much smaller, which together with the enrichment in positively charged residues (mostly R) and proline (known to be a secondary structure breaker) leads to a reduction in the number of possible short-range interactions. As a result, this region seems unstructured (with a slight tendency towards short α -helices) and almost completely solvent-exposed consistent with the predictions of local structure and solvent accessibility (**Figure 4B**). When analyzing the predicted contact map on the left panel in **Figure 4C** (taken from the *Contact Map app*), again we can easily point out the differences between the LCRs at the two ends of the protein. In particular, despite the enhanced intra-regions contacts for domains, the algorithms detected Q320-G410 inter-region with the slight probability of internal contacts (score ~ 0.35 while contact threshold is ≥ 0.5), which covers almost the entire C-terminal LCR, while the N-terminus appears without any internal contacts. Interestingly, using the outcome of the *Structural Disorder app* (row "BINDING" in **Figure 4B**), we see that the N-terminal region despite being structurally featureless is detected by the ANCHOR method as a protein binding region (A20-I95) with three picks above 0.8 score. This region coincides with a region enriched by arginine, which is known to be abundant in the LLPS-driving motifs and RNA binding regions⁴⁴. Therefore, we suspect that the experimentally observed dissolution of DL1 droplets in presence of highly concentrated RNA⁵³, is most likely due to strong electrostatic attraction between negatively charged nucleic acid and positively charged arginines. When the environment lacks ionic interaction partners, the DL1 isoform undergoes phase separation driven by weaker but still favorable π -cation interactions between C-terminal tyrosines and N-terminal arginines. And finally, the absence of the R-rich N-terminus in the isoform DL2 leads to fibril formation, driven by π - π -stacking of C-terminal tyrosines in two monomers.

Thus, using a programming analogy, we can say that the final behavior of hnRNPD isoforms appears as the result of a biomolecular algorithm composed of a series of if-then-else conditionals, depending on the presence of N-terminal and environmental conditions. Furthermore, the experiments confirmed that substitution of cation (R $\rightarrow$ K) or/and aromatic system (Y $\rightarrow$ F) significantly weakens the phase separation of DL1⁵³, which we think, points to the stronger matching of the R-Y pair. This complementarity results from the high cooperativity of multivalent interactions, in addition to the π -cation component also includes π - π -stacking (CZ=NH1 π -electrons donated by arginine compared to lysine) and side-chain hydrogen bonding (polar hydroxyl donated by tyrosine compared to nonpolar phenylalanine and multiple protons bonded to nitrogens donated by arginine).

^[1] The “II-STRUCT.”, “SA”, and “MOTIFS” rows in Figure 3B are taken from the outcome of the *Secondary structure*, *Solvent Accessibility*, and *Domains, Motifs, and Repeats* apps, respectively.

Finally, while understandably the focus tends to be on the low-complexity regions decorated with some binding and/or switching structural properties motifs, it is nevertheless also important to consider the well-folded domains, especially in cases like hnRNPD where they play a mediator between the two IDRs. The *Domains, Motifs, Repeats app* reveals the two HMMER-detected Pfam domains (both with ID: PF00076) located in M151-D219 and V235-I305 regions (**Figure 4A, 4B**). The presence of the domains is also well captured by the contact map prediction which reveals details about likely binding activities going on between distinct regions of the sequence (left panel in **Figure 4C**). Both, predictors of the secondary structure and known structures (PDBs) assigned to the sequences used in the Pfam seed-MSA, indicate that each domain has well-structured globular $\alpha+\beta$ topology (composed of 2 α -helices and single β -sheet), which perfectly explains the solubility of the third isoform (DL3) lacking both flexible ends. The list of corresponding PDBs can be found on the interactive labels assigned to each sequence in the Pfam seed-MSA in the *Sequence Conservation & MSA app* (e.g., 1P1T, residues: 18-88), and/or by following the link of the “STRUCTURE” button on the main page of the SingleSEQ tab. Specifically, for human hnRNPD (see the right panel in **Figure 4C**), the BIAPSS provides a direct link to the entry in the novel AlphaFold database collecting the high-quality predictions of structures that proved to be competitive with the experimental ones ²⁶.

Methods

1. Sequence complexity and physicochemical decoration

1.1 Sequence Complexity

Low complexity regions in proteins (LCRs) are compositionally biased fragments of sequences that often have low amino acid diversity, and repeats of short motifs of sequential or structural kinds. Many reports point to their functional ⁵⁶ or regulatory roles ⁵⁷, frequently also associated with subcellular phase separation ²⁸. The LCRs of LLPS proteins have been detected by using several state-of-the-art tools, such as SIMPLE ⁵⁸, CAST ⁵⁹, fLPS ⁶⁰, and SEG ⁶¹. The original hits were parsed by in-house algorithms to merge overlapping regions enriched in different amino acids and only the integrated and unified results have been kept.

The Shannon Entropy describes the information content held in data and is a frequently used measure of protein sequence complexity ⁶². We implemented a module for on-the-fly calculation of it within BIAPSS services. The typical window length for compositional effects is between 5 and 20. The results can be displayed in:

$$S_i = \frac{1}{N} \sum_{j=1}^N S_{(j,N)}$$

- residue resolution mode (residue option; smoother output):

where the Shannon entropy, $S_{(i)}$, at sequence position i is a sum of entropies at all windows containing this position normalized by the window length N .

$$S_{(j,N)} = - \sum_{aa=1}^{AA=20} f_{aa} \log_2(f_{aa})$$

- window resolution mode (block option):

where the Shannon entropy, $S_{(j,N)}$, at j -th sequence window of length N is summed over the fractions, f_{aa} , of 20 biogenic amino acids. The value is assigned to the center position within the window. $S_{(j,N)}$ ranges from 0, where only one residue is present within the sequence window) to $\log_2(N)$ (all positions are different). Therefore, the lower the Shannon entropy, the less complex sequence is.

1.2 Physicochemical decoration

To examine the physicochemical properties of LLPS-driving proteins, we identified along each sequence the patterns of polarity (Ser, Thr, Tyr, Gln, Asn, Cys, Met), hydrophobicity (Gly, Ala, Val, Ile, Leu, Pro, Phe), and detected π -stacking centers (Arg, Asn, Asp, Gln, Glu, Gly*) including those within aromatic rings (Phe, Tyr, Trp, His). We also provided the charge distributions split between positively (His, Lys, Arg) and negatively (Glu, Asp) charged residues. For each feature, both the arrangement along the sequence and the fraction of residues are provided.

1.3 Electrostatics

It is well established that the electrostatic interactions often affect the solubility and stabilize the binding interface in liquid-liquid demixing of biomolecules. The recently proposed charge decoration parameters emerged as a measure of charge distribution along the protein sequence. In addition to the overall charge content, these descriptors are seen as important factors shaping the protein conformations, especially within low-complexity regions¹⁸. Following these discoveries, we calculated and compared charge decoration parameters, namely:

- SCD, sequence charge decoration, is implemented following the formulation by Sawle & Ghosha⁶³
- OCS, overall charge symmetry, is implemented following the formulation by Das & Pappu⁶⁴
- FCR, a fraction of charged residues, is defined as a sum of the fractions of positive and negative charges.

2. HMMER-based sequence conservation and functional domains detection

The multiple sequence alignment (MSA) and consensus profile were prepared using an efficient HMMER method (*phmmer* + *hmmalign* and *hmmbuild*, respectively) that employs a probabilistic hidden Markov model (HMM) ²⁹ and is significantly more accurate compared to BLAST-based searches. Because some of the LLPS sequences are highly unique (detection of the remote homologs is needed) and because the MSA is reliable if at least several dozen of homologous sequences are available, we used sequences selected from various UniProt subsets. Specifically, SwissProt, UniRef50, and UniRef90 differ in the size and increasing sequence identity of entries ¹⁹. To identify sequence regions with significant evolutionary conservation, we have derived three additional MSA-based parameters: strength, diversity, and character. The MSA strength of the sequence conservation informs how much the specific position is held by evolution. This measure normalizes results from the *hmmlogo* tool to a discrete range from 0 (poorly conserved) to 5 (highly conserved). The *hmmlogo* computes letter heights along the sequence depending on the information content of the position. The MSA diversity defines the number of different amino acids detected at a given position in the MSA and is provided in discrete scale from 0 - highly conserved to 5 - poorly conserved (0 - one, 1 - two, 2 - three, 3 - four, 4 - five or six, 5 - 7 and more amino acids at the aligned position). The MSA character describes the chemical nature of the most common amino acids at a given position in the multiple sequence alignment. We distinguished the following attributes: polar, charge, aromatic, another π -system, hydrophobic, other (G or P).

Some LLPS proteins are composed of one or more well-known domains. The identification of these functional regions alongside regions of low complexity or disorder can provide additional insights into the regulatory role of phase separation. Therefore, we have performed a Pfam search for all LLPS proteins, reporting the detected domains and incorporating the original Pfam seed-MSAs for corresponding regions of LLPS sequence (instead of full-length ones) to derive more reliable evolutionary conservation descriptors.

3. Short sequential and structural motifs specific for LLPS sequences

Short Linear Motifs (SLiMs) are short fragments along the sequence, often situated in the intrinsically disordered regions generally showing high structural flexibility and evolutionary conservation. We systematically detected various short sequential and structural motifs. The implemented algorithms used the list of grouped motifs' instances defined by regular expressions as the keys to search protein

sequences prone to phase separation. Among motifs known from the literature as relevant for phase behavior, our analysis includes short structural stretches of protein sequence such as LARKS³¹ and steric zippers³², glycine-arginine-rich regions (GARs)⁴⁴, S-D/S-E motifs specific for phosphosites⁶⁵, and new sequential repetitive n-mers. Accordingly, the eukaryotic linear motifs (ELMs)⁶⁶, are indicated along the sequence with a discrimination of the main ELM classes.

4. Structural properties derived from sequence-based predictions

Bearing in mind the predictive nature of sequence-based methods and, hence, their limited accuracy, comparing several of them and choosing the final consensus has proven to be successful in many approaches⁶⁷. In our study, we comprised predictions from at least three to six widely used tools for each biomolecular characteristic. While almost every method is available as a web server, due to the size and complexity of our analyses, we employed standalone versions. The raw data derived from these standalone tools during the high-performance computing was initially parsed, filtered, and simplified to a uniform CSV format and deposited in our online repository at <https://biapss.chem.iastate.edu/download.html>.

4.1 Secondary structure

Protein secondary structure is a regular three-dimensional organization of local fragments along a polypeptide chain. The two most common secondary structural elements are alpha helices and beta sheets. Most of the predictors provide the secondary structure assignment in 3-letter notation (ss3): H - helix, E - strand, C - coil, while the advanced ones (RaptorX and PORTER5) deliver also more detailed 8-letter notation (ss8): H - α -helix, G - 310-helix, I - π -helix, E - β -strand, B - β -bridge, T - HB-turn, S - bend, C - loop. In our benchmark study, we employed five well-established secondary structure predictors: PSIPRED⁶⁸, RaptorX⁶⁹, PORTER-5⁶⁹, SPIDER-3⁷⁰, FESS⁷¹.

4.2 Solvent accessibility

Solvent accessibility gives some insight into protein structural flexibility, indicating the exposed patches on the protein surface available for interactions with the solvent molecules. Some surface sites have high

evolutionary conservation, which is suggestive of functional or structural importance. Since not many structures of phase-separating IDPs are known, the robust prediction of solvent accessibility can help to identify flexible regions prone to conformational changes upon binding. The assignment of solvent accessibility is usually provided in the 3-letter code: B - buried, E - exposed, M - medium. In our benchmark study, we employed three well-established solvent accessibility predictors: RaptorX⁷², PaleAle5⁶⁸, SPOT-1D⁷³.

4.3 Structural disorder

The sequence-based predictions indicate regions of increased structural flexibility, usually estimating the disorder probability at a given position in the sequence. Detecting highly flexible regions may support the identification of short sequence stretches of multivalent interactions which can be relevant to phase separation. In our benchmark study, we employed seven well-established predictors of structural disorder: RaptorX⁷², IUPred2⁷⁴, SPOT-Disorder⁷⁵, DISOPRED2⁷⁶, DISOPRED3⁷⁷, VSL2⁷⁸, PONDR-VLXT⁷⁹, PONDR-FIT⁸⁰. Most of these methods return the probability of disorder for each position in the sequence. Usually, the residue is considered as ordered when the score is below 0.5. The protein binding regions in disordered fragments were estimated by using the ANCHOR method⁸¹.

4.4 Contact map

Contact map application provides a more reduced representation of a protein structure using a binary two-dimensional matrix of distances between all possible amino acid residue pairs. The commonly used definition assumes the threshold 6-10Å as the distance between the pair of two C α or C β atoms being in contact. The contact number of protein residues limits the number of possible protein conformations and helps encode a three-dimensional structure. In our benchmark study, we employed three state-of-the-art predictors of intramolecular contacts: RaptorX-Contact⁸², ResPRE⁷⁴, SPOT-Contact⁷³.

5. Data availability

BIAPSS is a free online resource accessible at <https://biapss.chem.iastate.edu/>. The web interface of BIAPSS is developed within the Python web framework and Plotly-Dash graphing libraries, with all the major browsers supported including the mobile device accessibility. The SingleSEQ repository is freely available for download <https://biapss.chem.iastate.edu/download.html>. The shared data includes raw predictions which are pre-calculated using a set of well-established tools as well as findings of our deep statistical analysis of overall characteristics of LLPS sequences. The comprehensive documentation for

the BIAPSS services is available online at <https://biapss.chem.iastate.edu/documentation.html>. For the new users, BIAPSS provides many simple tutorials demonstrating the usage of available web applications highlighting their options and features. Specifically, the orange Quick Guide button is located in the top-left corner of the particular web application and context-rich tooltips are attached to most selection forms.

Declarations

Author Contribution

Conceptualization, A.E.B-D.; Software development, A.E.B-D.; Data curation, A.E.B-D.; Supervision & review, V.N.U., and D.A.P.; Writing an original draft, A.E.B-D., V.N.U., and D.A.P.

Acknowledgments

A.E.B-D. acknowledges financial support by Roy J. Carver Charitable Trust through Iowa State University Bioscience Innovation Postdoctoral Fellowship. This work was supported by the National Institute Of General Medical Sciences of the National Institutes of Health [R35GM138243 to D.A.P.]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: none declared.

References

1. Brangwynne, C. P. *et al.* Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **324**, 1729–1732 (2009).
2. Brangwynne, C. P., Mitchison, T. J. & Hyman, A. A. Active liquid-like behavior of nucleoli determines their size and shape in *Xenopus laevis* oocytes. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4334–4339 (2011).
3. Sawyer, I. A., Bartek, J. & Dundr, M. Phase separated microenvironments inside the cell nucleus are linked to disease and regulate epigenetic state, transcription and RNA processing. *Semin. Cell Dev. Biol.* **90**, 94–103 (2018).
4. Banjade, S. *et al.* Conserved interdomain linker promotes phase separation of the multivalent adaptor protein Nck. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E6426–35 (2015).

5. Li, P. *et al.* Phase transitions in the assembly of multivalent signalling proteins. *Nature* **483**, 336–340 (2012).
6. Choi, J.-M., Holehouse, A. S. & Pappu, R. V. Physical Principles Underlying the Complex Biology of Intracellular Phase Transitions. *Annu. Rev. Biophys.* **49**, 107–133 (2020).
7. Wang, J. *et al.* A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* **174**, 688–699.e16 (2018).
8. Dignon, G. L., Best, R. B. & Mittal, J. Biomolecular Phase Separation: From Molecular Driving Forces to Macroscopic Properties. *Annu. Rev. Phys. Chem.* **71**, 53–75 (2020).
9. Savojardo, C., Martelli, P. L. & Casadio, R. Protein–Protein Interaction Methods and Protein Phase Separation. *Annu. Rev. Biomed. Data Sci.* **3**, 89–112 (2020).
10. Borchers, W., Bremer, A., Borgia, M. B. & Mittag, T. How do intrinsically disordered protein regions encode a driving force for liquid-liquid phase separation? *Curr. Opin. Struct. Biol.* **67**, 41–50 (2020).
11. Zaslavsky, B. Y., Ferreira, L. A. & Uversky, V. N. Driving Forces of Liquid-Liquid Phase Separation in Biological Systems. *Biomolecules* **9**, (2019).
12. Tsang, B., Pritišanac, I., Scherer, S. W., Moses, A. M. & Forman-Kay, J. D. Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations. *Cell* **183**, 1742–1756 (2020).
13. Saar, K. L. *et al.* Machine learning models for predicting protein condensate formation from sequence determinants and embeddings. *Cold Spring Harbor Laboratory* 2020.10.26.354753 (2020) doi:10.1101/2020.10.26.354753.
14. You, K. *et al.* PhaSepDB: a database of liquid-liquid phase separation related proteins. *Nucleic Acids Res.* **48**, D354–D359 (2020).
15. Mészáros, B. *et al.* PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res.* **48**, D360–D367 (2020).
16. Li, Q. *et al.* LLPSDB: a database of proteins undergoing liquid–liquid phase separation in vitro. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz778.
17. Ning, W. *et al.* DrLLPS: a data resource of liquid–liquid phase separation in eukaryotes. *Nucleic Acids Res.* **48**, D288–D295 (2019).
18. Bianchi, G., Longhi, S., Grandori, R. & Brocca, S. Relevance of Electrostatic Charges in Compactness, Aggregation, and Phase Separation of Intrinsically Disordered Proteins. *Int. J. Mol. Sci.* **21**, (2020).
19. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
20. Piovesan, D. *et al.* DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.* **45**, D219–D227 (2017).
21. Binder, J. X. *et al.* COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database* 2014, bau012 (2014).
22. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

23. Varadi, M. *et al.* PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Res.* **48**, D344–D353 (2020).
24. Pieper, U. *et al.* ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* **42**, D336–46 (2014).
25. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
26. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
27. Radivojac, P. *et al.* Protein flexibility and intrinsic disorder. *Protein Sci.* **13**, 71–80 (2004).
28. Martin, E. W. & Mittag, T. Relationship of Sequence and Phase Separation in Protein Low-Complexity Regions. *Biochemistry* **57**, 2478–2487 (2018).
29. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
30. Nelson, R. *et al.* Structure of the cross-beta spine of amyloid-like fibrils. *Nature* **435**, 773–778 (2005).
31. Hughes, M. P. *et al.* Atomic structures of low-complexity protein segments reveal kinked β sheets that assemble networks. *Science* **359**, 698–701 (2018).
32. Riek, R. The Three-Dimensional Structures of Amyloids. *Cold Spring Harb. Perspect. Biol.* **9**, (2017).
33. Sawaya, M. R. *et al.* Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature* **447**, 453–457 (2007).
34. Wardah, W., Khan, M. G. M., Sharma, A. & Rashid, M. A. Protein secondary structure prediction using neural networks and deep learning: A review. *Comput. Biol. Chem.* **81**, 1–8 (2019).
35. Yang, Y. *et al.* Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Brief. Bioinform.* **19**, 482–494 (2016).
36. Aydin, Z., Azginoglu, N., Bilgin, H. I. & Celik, M. Developing structural profile matrices for protein secondary structure and solvent accessibility prediction. *Bioinformatics* **35**, 4004–4010 (2019).
37. Mukherjee, S. & Bahadur, R. P. An account of solvent accessibility in protein-RNA recognition. *Sci. Rep.* **8**, 10546 (2018).
38. Murray, D. T. *et al.* Structure of FUS Protein Fibrils and Its Relevance to Self-Assembly and Phase Separation of Low-Complexity Domains. *Cell* **171**, 615–627.e16 (2017).
39. Birsa, N. *et al.* FUS-ALS mutants alter FMRP phase separation equilibrium and impair protein translation. *Sci Adv* **7**, (2021).
40. Birsa, N., Bentham, M. P. & Fratta, P. Cytoplasmic functions of TDP-43 and FUS and their role in ALS. *Semin. Cell Dev. Biol.* **99**, 193–201 (2020).
41. Protter, D. S. W. & Parker, R. Principles and Properties of Stress Granules. *Trends Cell Biol.* **26**, 668–679 (2016).
42. Emmanouilidis, L. *et al.* NMR and EPR reveal a compaction of the RNA-binding protein FUS upon droplet formation. *Nat. Chem. Biol.* **17**, 608–614 (2021).

43. Holehouse, A. S., Ahad, J., Das, R. K. & Pappu, R. V. CIDER: classification of intrinsically disordered ensemble regions. *Biophys. J.* **108**, 228a (2015).
44. Chong, P. A., Vernon, R. M. & Forman-Kay, J. D. RGG/RG Motif Regions in RNA Binding and Phase Separation. *J. Mol. Biol.* **430**, 4650–4665 (2018).
45. King, O. D., Gitler, A. D. & Shorter, J. The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease. *Brain Res.* **1462**, 61–80 (2012).
46. Qamar, S. *et al.* FUS Phase Separation Is Modulated by a Molecular Chaperone and Methylation of Arginine Cation- π Interactions. *Cell* **173**, 720–734.e15 (2018).
47. Krainer, G. *et al.* Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions. *Nat. Commun.* **12**, 1085 (2021).
48. Murthy, A. C. *et al.* Molecular interactions underlying liquid–liquid phase separation of the FUS low-complexity domain. *Nature Structural & Molecular Biology* vol. 26 637–648 (2019).
49. Loughlin, F. E. *et al.* The Solution Structure of FUS Bound to RNA Reveals a Bipartite Mode of RNA Recognition with Both Sequence and Shape Specificity. *Mol. Cell* **73**, 490–504.e6 (2019).
50. Maris, C., Dominguez, C. & Allain, F. H.-T. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS Journal* vol. 272 2118–2131 (2005).
51. Shazman, S. & Mandel-Gutfreund, Y. Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput. Biol.* **4**, e1000146 (2008).
52. Lin, Y., Protter, D. S. W., Rosen, M. K., Correspondence, R. P. & Parker, R. Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. *Mol. Cell* **60**, 208–219 (2015).
53. Batlle, C. *et al.* hnRNPD Phase Separation Is Regulated by Alternative Splicing and Disease-Causing Mutations Accelerate Its Aggregation. *Cell Rep.* **30**, 1117–1128.e5 (2020).
54. Gabryelczyk, B. *et al.* Hydrogen bond guidance and aromatic stacking drive liquid-liquid phase separation of intrinsically disordered histidine-rich peptides. *Nat. Commun.* **10**, 5465 (2019).
55. Vasilyev, N. *et al.* Crystal structure reveals specific recognition of a G-quadruplex RNA by a β -turn in the RGG motif of FMRP. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5391–400 (2015).
56. Kumari, B., Kumar, R., Chauhan, V. & Kumar, M. Comparative functional analysis of proteins containing low-complexity predicted amyloid regions. *PeerJ* vol. 6 e5823 (2018).
57. Coletta, A. *et al.* Low-complexity regions within protein sequences have position-dependent roles. *BMC Syst. Biol.* **4**, 43 (2010).
58. Albà, M. M., Laskowski, R. A. & Hancock, J. M. Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics* **18**, 672–678 (2002).
59. Promponas, V. J. *et al.* CAST: an iterative algorithm for the complexity analysis of sequence tracts. *Bioinformatics* **16**, 915–922 (2000).
60. Harrison, P. M. fLPS: Fast discovery of compositional biases for the protein universe. *BMC Bioinformatics* **18**, 476 (2017).

61. Wootton, J. C. & Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17**, 149–163 (1993).
62. Li, X. & Kahveci, T. A Novel algorithm for identifying low-complexity regions in a protein sequence. *Bioinformatics* **22**, 2980–2987 (2006).
63. Sawle, L. & Ghosh, K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J. Chem. Phys.* **143**, 085101 (2015).
64. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 13392–13397 (2013).
65. Owen, I. & Shewmaker, F. The Role of Post-Translational Modifications in the Phase Transitions of Intrinsically Disordered Proteins. *Int. J. Mol. Sci.* **20**, (2019).
66. Kumar, M. *et al.* ELM—the eukaryotic linear motif resource in 2020. *Nucleic Acids Res.* **48**, D296–D306 (2019).
67. Pollastri, G., Martin, A. J. M., Mooney, C. & Vullo, A. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics* **8**, 201 (2007).
68. Buchan, D. W. A. & Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407 (2019).
69. Wang, Z., Zhao, F., Peng, J. & Xu, J. Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics* **11**, 3786–3792 (2011).
70. Heffernan, R. *et al.* Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *J. Comput. Chem.* **39**, 2210–2216 (2018).
71. Piovesan, D., Walsh, I., Minervini, G. & Tosatto, S. C. E. FIELDS: fast estimator of latent local structure. *Bioinformatics* **33**, 1889–1891 (2017).
72. Wang, S., Li, W., Liu, S. & Xu, J. RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res.* **44**, W430–5 (2016).
73. Hanson, J., Paliwal, K., Litfin, T., Yang, Y. & Zhou, Y. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics* **35**, 2403–2410 (2019).
74. Li, Y., Hu, J., Zhang, C., Yu, D.-J. & Zhang, Y. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics* **35**, 4647–4655 (2019).
75. Hanson, J., Paliwal, K. K., Litfin, T. & Zhou, Y. SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning. *Genomics Proteomics Bioinformatics* **17**, 645–656 (2019).
76. Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F. & Jones, D. T. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 2138–2139 (2004).

77. Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857–863 (2015).
78. Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K. & Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**, 208 (2006).
79. Radivojac, P., Obradović, Z., Brown, C. J. & Dunker, A. K. Prediction of boundaries between intrinsically ordered and disordered protein regions. *Pac. Symp. Biocomput.* 216–227 (2003).
80. Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K. & Uversky, V. N. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta* **1804**, 996–1010 (2010).
81. Dosztányi, Z., Mészáros, B. & Simon, I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25**, 2745–2746 (2009).
82. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput. Biol.* **13**, e1005324 (2017).

Figures

Figure 1

The overall layout of the BIAPSS web platform (<https://biapss.chem.iastate.edu>) for comprehensive sequence-based analysis of LLPS proteins. The interactive applications implemented within the SingleSEQ module employ a residue-resolution biophysical analyzer for interrogating individual LLPS proteins.

Figure 2

Graphical overview of the types of analyses employed in BIAPSS SingleSEQ web applications for studying individual protein sequences with confirmed LLPS behavior.

Figure 3

Figure 3A. The left panel shows the cellular location of FUS (image source: BIAPSS). The protein is predominantly located in the nucleus. The physiological low levels of FUS found in the cytoplasm typically self-organize to membraneless compartments such as stress granule or ribonucleoprotein granule. The aberrant disease-related aggregates are mostly localized in the cytoplasm. The methylation (purple stars) of C-terminal arginines (green tail) in the wild-type FUS strongly promotes phase separation and gelation. The phosphorylation (magenta stars) of serine and threonine in the N-terminus (blue tail) dissolve liquid-liquid droplets. The tyrosine to phenylalanine mutants (yellow stars) in the N-terminus and hypomethylation of arginines in the C-terminus increase aggregation. The right panel shows classification of intrinsically disordered ensemble regions (CIDER, ⁴³) for the FUS sequence split into functional segments.

Figure 3B. Sequence composition and complexity of FUS. The upper panel shows the amino acid (AA) content of the query sequence compared to the Eukaryota dataset. The bottom panel shows along the query sequence (i) the patterning of enriched amino acids (S, Q, Y, R, G), (ii) low-complexity measures

provided as regions of particular AA enrichments (LCR row), and the sequence information content (H(S) row, Shannon entropy), (iii) consensus of predicted disorder regions and secondary structure assignment, (iv) detected Pfam domains, and (v) evolutionary conservation derived from the multiple sequence alignment against UniRef50. The amino acid patterning section contains points corresponding to the locations of most relevant serine phosphorylations (30, 42, 54, 61, 84, 87) ³⁸, arginine methylations (216, 259, 407, 472, 473, 476) ⁴⁶, and tyrosine mutations (113, 122, 130, 136, 143, 149, 161) ⁴⁶.

Figure 3C. Physicochemical and structural properties of FUS. The various characteristics are shown along the protein sequence split on the N-terminus (blue) and C-terminus (green). The full-length N-terminus corresponds to a highly polar prion-like domain (PLD). The PLD contains a core region (39-95), in which multiple LARKS motifs were detected and evidenced to form fibrils in a solid state. The C-terminus contains two well-folded domains detected by Pfam search (RRM, ZnF) and three glycine-arginine-rich (GAR) regions. All components of the C-terminus are known to have a functional role in RNA-binding. The following rows show the physicochemical patterning along the full-length FUS sequence, including charge (magenta - positive, blue - negative), polarity, donor (blue) / acceptor (magenta) / both (purple) of side-chain proton for hydrogen bonding, π -electron-containing systems (blue) with separation of aromatic ones (dark-blue), hydrophobicity, predicted solvent accessibility (SA) in 3-letter notation (magenta - exposed, green - buried, blue - medium). The zoom of the PLD core is shown at the bottom panel, where the SS and SA rows contain the predicted probabilities of secondary structure and solvent accessibility. The green arrows indicate the side chains buried in the fibril core, while the black frames highlight segments that form strands of a cross- β motif ³⁸.

Figure 4

Figure 4A. Sequence composition and complexity of hnRNPD. The upper panel shows the amino acid (AA) content of the query sequence compared to globular, disordered, and phase-separating reference datasets. The bottom panel shows along the query sequence (i) the patterning of enriched amino acids (Y, R, Q, P, G), (ii) low-complexity measures provided as regions of particular AA enrichments (*LCR row*), and the sequence information content (*H(S) row*, Shannon entropy), (iii) consensus of predicted disorder regions, and (iv) detected Pfam domains.

Figure 4B. Physicochemical and structural properties of hnRNPD. The various characteristics are shown along the protein sequence split on three main regions: N-terminus (blue), folded globular domains (sand), C-terminus (green). The full-length N-terminus is disordered and contains a PR-rich region detected to show high binding affinity. The C-terminus contains a Y-rich region decorated with two LARKS-like motifs, which is flanked by two disordered regions enriched with glycine-arginine-rich motifs. The low-complexity terminals show different physicochemical patterning (the content of polar, charged, and other amino acids).

π -electron systems) and structural properties (SA - solvent accessibility, II-STRUCT - secondary structure elements).

Figure 4C. The structure and internal contacts of hnRNPD. The left panel shows the RAPTOR-predicted contact map with the highlighted regions corresponding to N-terminus (blue), folded domains (black with two sand squares), and C-terminus (green). The upper rows contain the PORTER-predicted secondary structure and secondary structure assignment taken from the known protein structures (corresponding to the sequences used in the seed-MSA for detected Pfam domains). The right panel shows the structure of hnRNPD predicted using the AlphaFold method and colored by model confidence.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [toc.jpg](#)