Intrinsically disordered regions that drive phase separation form a robustly distinct protein class

Short title: Intrinsic properties of phase-separating IDRs

Ayyam Y. Ibrahim,[1] Nathan P. Khaodeuanepheng,[1] Dhanush L. Amarasekara,[2] John J. Correia,[3] Karen A. Lewis,[1] Nicholas C. Fitzkee,[2] Loren E. Hough,[4,5] and Steven T. Whitten[1*]*

[1] Department of Chemistry and Biochemistry, Texas State University, San Marcos, TX, USA

[2] Department of Chemistry, Mississippi State University, Mississippi State, MS, USA

[3] Department of Cell and Molecular Biology, University of Mississippi Medical Center, Jackson, MS, USA

[4] Department of Physics, University of Colorado Boulder, Boulder, CO, USA

[5] BioFrontiers Institute, University of Colorado Boulder, Boulder, CO, USA

* Co-corresponding authors

E-mail: loren.hough@colorado.edu

E-mail: steve.whitten@txstate.edu

Abstract

Liquid-liquid phase separation (LLPS) of proteins is thought to be a primary driving force for the formation of membraneless organelles, which control a wide range of biological functions from stress response to ribosome biogenesis. LLPS of proteins in cells is primarily, though not exclusively, driven by intrinsically disordered (ID) domains. Accurate identification of ID regions (IDRs) that drive phase separation is important for testing the underlying mechanisms of phase separation, identifying biological processes that rely on phase separation, and designing sequences that modulate phase separation. To identify IDRs that drive phase separation, we first curated datasets of folded, ID, and phase-separating (PS) ID sequences. We then used these sequence sets to examine how broadly existing amino acids scales can be used to distinguish between the three classes of protein regions. We found that there are robust property differences between the classes and, consequently, that numerous combinations of amino acid property scales can be used to make robust predictions of LLPS. This result indicates that multiple, redundant mechanisms contribute to the formation of phase-separated droplets from IDRs. The top-performing scales were used to further optimize our previously developed predictor of PS IDRs, ParSe. We then modified ParSe to account for interactions between amino acids and obtained reasonable predictive power for mutations that have been designed to test the role of amino acid interactions in driving LLPS.

Introduction

Many intracellular reactions occur within membrane-free compartments that form spontaneously from the cellular milieu (1). Examples of such compartments include P-bodies, Cajal bodies, the nucleolus, paraspeckles, and germ granules (2–4). The formation of membraneless organelles is facilitated primarily, though not exclusively (5, 6), by proteins that are intrinsically disordered (ID) or contain large ID regions (IDRs), collectively termed intrinsically disordered proteins (IDPs) (4, 7). Because these protein-rich droplets typically exist in dynamic, liquid-like states rather than as fixed complexes (1, 2), this transition is referred to as liquid-liquid phase separation (LLPS). By forming specific compartments and micro-environments, LLPS exerts control over the spatial organization and biochemical reactivity within cells (8, 9). Indeed, LLPS has been found to modulate chemical and biochemical reactions (10–12) and LLPS dysregulation has been associated with several human diseases (13–15).

Due to the critical role of protein LLPS in cell function and disease, significant efforts have been made to determine the physical mechanisms responsible for driving phase separation. Mutation and sequence analysis have demonstrated the importance of cation-π, π-π, π/sp^2, and hydrophobic interactions (16–21). Groups of amino acids driving cohesive interactions are often characterized as "stickers" and are frequently interspaced with small polar residues acting as "spacers" (22–25). In addition, charge composition and patterning appear to contribute to the regulation of LLPS by IDRs (20, 26–29). Successfully predicting the relationship between primary sequence and phase separation behavior is key to understanding the molecular mechanisms underlying LLPS and to identifying the cellular processes that rely on LLPS. Effective predictive algorithms might also reveal how mutations affect LLPS-associated disease states.

Several methods have been developed to predict which protein sequences drive LLPS (30, 31). Algorithms including PLAAC, PSPredictor, and PSPer are based on the composition of databases of proteins that are known to drive LLPS (28, 32, 33). Others, including catGRANULE and CRAPome, are associated with cellular structures (34, 35). Uniquely, PScore was developed based on a specific mechanism thought to drive LLPS: the propensity of cation-π and π-π interactions to drive cohesive protein interactions (16, 36). Simulation models of IDRs have also been used to identify which protein domains drive LLPS as well as how mutations will affect LLPS behavior of those proteins (37–41). The diversity of successful approaches for predicting LLPS indicates that multiple complementary mechanisms may be responsible for this phenomenon.

We previously developed a predictive model of LLPS behavior, ParSe ("**Par**tition **Se**quence"), that identifies phase-separating (PS) IDRs starting from predictions of hydrodynamic size, which is indicative of the relative strength of intramolecular as compared to solvent interactions (42). The ParSe algorithm uses a sequence-calculated polymer scaling exponent, ν_{model}, to quantify hydrodynamic size (43, 44). When ν_{model} is combined with a second sequence-based parameter, the intrinsic propensity for a sequence to form β-turns (45), the algorithm can distinguish between sequences belonging to one of three classes of protein regions: folded, ID, and PS ID (42). We proposed a physical mechanism whereby transient β-turn structures reduce the desolvation penalty of forming a protein-rich phase and increase exposure of atoms involved in π/sp^2 valence electron interactions. In this mechanism, β-turns could act as energetically favored nucleation points,

potentially explaining the observed higher propensity for turns in IDRs that drive phase separation *in vivo* (42, 45).

However, the prior study did not test whether the combination of ν_{model} and β-turn propensity was uniquely able to distinguish folded, ID, and PS ID sequences, as would be required if this putative mechanism is necessary for LLPS. In the current study, we first curated the sequence training sets to expand the folded and ID categories. Our curated list of proteins that are ID but not thought to drive LLPS acts as a key negative control, enabling us to distinguish which features of IDRs in particular drive LLPS (31). Using the expanded sequence sets, we exhaustively tested all amino acid property scales found in the Amino Acid Index Database (46) for their ability to separate folded, ID, and PS ID sequences. We show that the three sequence sets are distinct in their means when quantified by the majority of amino acid scales, revealing that there are robust property differences between ID and PS ID sequences, not unlike the differences between folded and ID sequences.

We applied principal component analysis (PCA) to identify the extent of variability between our sequence sets and the optimal combinations of property scales that maximize the distinction between ID and PS ID sequences. The resulting predictor, ParSe version 2 (v2), uses sequence hydrophobicity to distinguish folded from ID, and, subsequently, ν_{model} and a conformational parameter to distinguish ID from PS ID. In general, PS ID sequences exhibit enriched β-turn and depleted α-helix propensities. ParSe v2 more accurately predicts these regions from the amino acid sequence than the original version. We then compared our predicted propensity for LLPS with experimental results on mutant sequences designed to test the role of π- and charge-based interactions in LLPS behavior. We found that only by including effects representing interactions between amino acids could we accurately predict LLPS behavior of these mutants. Given the high fidelity of ParSe even in the absence of these interaction terms, it appears there are multiple diverse mechanisms that can drive LLPS and that PS ID sequences can be successfully identified through simple combinations of amino acid property scales.

Results

Construction of Protein Sequence Datasets. A limitation of the previous work, including our own (42), has been the relatively small set of sequences used to train predictors. We first sought to alleviate this problem by identifying additional sequences in our two negative control categories, folded proteins and IDRs, which are not thought to phase separate. The importance of well-defined negative control sets has been highlighted recently by Pansca et al (31) and Cai et al (47). For example, some negative control sets like the human proteome are known to contain many false negatives, which can lead to misassignments by the predictor.

We first expanded the set of folded proteins. Previously, we selected only folded regions found within known LLPS proteins. However, this selection may not be justified because it is not known whether folded regions within phase-separating proteins are biased differently in ν_{model} and β-turn propensity compared to folded proteins in general. Subsequently, we expanded the previous folded set (comprised of 82 sequences) to include sequences from 122 human proteins with nonhomologous folded structures (48), 32 proteins with small (N = 36) to large (N = 415) folded

structures (49), 54 folded extremophile proteins (50), 53 folded metamorphic proteins (51), and 90 folded membrane proteins (Table S1). Combined, these folded protein regions represent 421 unique sequences after removing duplicate entries. The folded sets were, overall, similar in both mean ν_{model} (Figure 1A, Table 1) and mean β-turn propensity (Table 2) to the previous folded set obtained from known LLPS proteins (Tables S2, S3), indicating that folded regions within LLPS proteins are indeed similar to folded regions more generally.

Similarly, we expanded the set of IDR sequences not enriched for LLPS potential, called the "ID" set, by adding ID sequences found in the Biological Magnetic Resonance Bank (BMRB) (52) and DisProt (53, 54) databases. NMR experiments are typically performed at relatively high concentrations (≥100 μM), and so BMRB entries that do not explicitly address LLPS likely have a low propensity to phase separate. In addition, proteins known to drive LLPS are now annotated in DisProt; therefore, DisProt entries lacking such annotation are at least nominally depleted in LLPS drivers. Moreover, we only selected IDRs from DisProt that were both predicted to be disordered by MetaPredict (55) and were not highly homologous to proteins with folded structures in the Protein Data Bank (PDB) (56) using seqatoms (57). The combined ID set contains 121 unique protein domains (Table S4).

While these expanded datasets show slight differences in mean predicted ν_{model} or β-turn propensity from the datasets used in our previous work (Tables 1, 2, S2, S3), the expanded sets reinforce our previous findings that there exist significant differences in ν_{model} or β-turn propensity between the three classes of protein regions: folded, ID, and PS ID (Figure 1A). These results, as such, confirm that the two sequence-calculated metrics, ν_{model} and β-turn propensity, can be used in combination, as done previously, to predict phase separating regions within proteins (42). Additionally, when ν_{model} and β-turn propensity are calculated for homopolymers of the common amino acids and then presented in a β-turn propensity versus ν_{model} plot (Figure 1B), the values obtained are consistent with previous characterizations of "order promoting" as compared to "disorder promoting" amino acids (Figure 1B). In particular, we find that homopolymers of Trp, Cys, Phe, Ile, Tyr, Val, Leu, Ala, His, Met, and Thr fall within the "folded" region of the β-turn versus ν_{model} plot, and so are predicted to act as "order promoting" amino acids, while by similar analysis Arg, Gln, Pro, Glu, Lys, and Asp are "disorder promoting", and Asn, Ser, and Gly are "phase separation promoting". This result is similar to conclusions from analyses of protein structures (58, 59), where Trp, Cys, Phe, Ile, Tyr, Val, Leu, and Asn are enriched in folded proteins ("order promoting"), while Ala, Arg, Gln Pro, Glu, Lys, Gly, and Ser are enriched in IDPs ("disorder promoting"), and His, Met, Thr, and Asp are "ambiguous".

The clear segregation of some amino acids into the PS ID sector of the β-turn propensity versus ν_{model} plot motivated us to consider whether an approach as simple as color coding of the amino acids would enable identification of PS regions in proteins known to phase separate. Indeed, the PS driving regions of many proteins are visually apparent by our simple visualization tool based on the location of homopolymers in the β-turn propensity versus ν_{model} plot (Figure 1B). The magnitude is related to the propensity and the color indicates the quadrant of the plot; therefore, a shaded bar chart predicts the propensity for a sequence to promote order, disorder, or phase separation. The rapid identification of PS regions in proteins (Figure 1C) such as Ddx4, FUS, and Sup35 (3, 17, 22, 60) led us to conclude that PS regions in proteins are distinctly different than

other ID regions. We therefore sought to determine whether these classes of proteins were distinguishable by other amino acid property scales.

Most amino acids property scales find significant differences between folded, ID, and phase-separating protein regions. We sought to determine if additional sequence-based intrinsic properties were significantly different between protein regions that are folded, ID, or ID with high potential for driving LLPS. To explore this idea, 566 scales of amino acid properties were obtained from the Amino Acid Index Database (46), which is a curated set of numerical indices representing various physicochemical and biochemical properties of the amino acids. This approach is similar to work done to improve coarse-grained models by testing multiple hydrophobicity scales (40). We added to these scales a newly developed hydrophobicity scale designed to predict sequences that drive protein LLPS (19), as well as ν_{model}. For each scale and for each sequence, we summed the amino acid scale for amino acids in the sequence, and divided by the length, N. Welch's unequal variances t-test (61), given as a one-tail p-value, was used to find scales that show a statistical difference in the means of the sequence sets. Using the nonparametric Mann-Whitney U-test (62) gave overall similar results (Figure S1).

Figures 2A-C show that the different sequence sets have significantly different mean values for most scales when compared. For example, 81% of scales give p-values <0.05 (indicating means that are statistically different), when comparing ID and PS ID sequences (Figure 2A). Moreover, 13% and 22% of scales yield p-values smaller (thus showing a more significant difference) than the p-values obtained from ν_{model} and β-turn propensity, respectively, used in ParSe (42). Each scale type (e.g., α-helix propensity, β-turn propensity, hydrophobicity, etc.) had some scales with very low p-values and some with p-values ≥0.05, suggesting that, overall, most, but not all, conformational and physicochemical based scales could substitute for ν_{model} or β-turn propensity in ParSe and likely exhibit some ability for identifying PS IDRs from sequence. This analysis reveals that the physical differences between PS and conventional IDRs are robust across many different scales of amino acid properties (Figure 2A). We conclude that PS regions likely contain a variety of complementary, redundant sequence features that drive LLPS.

The differences between folded and ID (both ID and PS ID) datasets are also robust to different scales of amino acid properties (Figures 2B-C). 95% and 93% of scales produced p-values <0.05 when means were compared between the folded and PS ID, and folded and ID sets, respectively. Almost all amino acid property scales yield statistically different means when comparing ID and folded sequences; the best performing scales were based on hydrophobicity. Those hydrophobicity scales with the lowest p-values when comparing means in the folded and ID sets had among the highest p-values when comparing means in the ID and PS ID sets (and *vice versa*), consistent with our previous findings that a single metric was insufficient to separate the three datasets.

PCA identifies two principal modes of variation between proteins. We next sought to determine the degree to which amino acid scales could be combined without significant redundancy when comparing protein sequences. To do so, we used principal component analysis (PCA), which characterizes the variability in a dataset (63), in this case variability arising from different scales being applied to our sequences. Our primary focus is on distinguishing PS IDRs from conventional IDRs because many disordered predictors already exist to separate folded from disordered domains (55, 64, 65). We first selected the scale in each scale type (listed in Figure 2A)

with the smallest *p*-value when comparing the ID and PS ID sets; that is representative scales from each type that are best able to separate ID and PS ID sequences. We additionally included ν_{model}, which we found previously to give complementary information to β-turn propensity. Each scale was then used to calculate sequence properties via a sliding 25-residue window applied to protein domains in a combined set including both the ID and PS ID datasets or the human proteome. We used a sliding window to avoid averaging properties between regions of proteins with different characteristics (42).

The results of the PCA indicate that most of the variability measured by high-performing scales within these datasets can be captured by 2-3 parameters (Figures 2D-E, S2). For both the combined ID dataset including ID and PS ID sequences and the human proteome, approximately 70% of the variability is captured by the first two principal components. Moreover, 58% of the variability in the combined ID set is captured by a single component. The variance arising from conformational propensity scales tend to cluster, as do those with physicochemical metrics like charge, hydrophobicity, and other compositional details. These results are robust to both the number of top-performing scales chosen and to the choice of reference set; we saw similar clustering when we extended this analysis to include the top three performing scales in each type and to the entire human proteome (Figure S2).

Within these two categories (conformational propensity and physicochemical metrics), high-performing scales function very similarly. As such, the predictive capabilities of amino acid scale combinations within each category are limited. In particular, turn and coil scales applied to protein sequences yield strongly correlated modes of variation that also are mostly anti-correlated with the variance produced from α-helix propensity scales (Figures 2D, S2). In our previous work, we proposed that β-turns could serve as a site for cohesive interactions between protein chains, driving LLPS (42). Our current results, while consistent with this hypothesis, show that this hypothesis cannot easily be distinguished from other structural hypotheses, e.g., that coils drive LLPS or helix inhibits LLPS, because the variation between these scales when applied to our datasets are all highly correlated. In contrast, the variances arising from hydrophobicity, charge, or ν_{model} in our datasets have patterns that, in general, are different from the variances arising from turn, coil, and α-helix conformational propensities.

To illustrate the separation obtained when using complementary top-performing scales, we selected three scales to best separate our three datasets: 1) the top performing hydrophobicity scale for separating folded from either ID set (from Vendruscolo and coworkers (66)), 2) the top performing conformational scale in separating ID from PS ID sets, in this case one predicting α-helical propensity (from Tanaka and Scheraga (67)), and 3) ν_{model} because it was most orthogonal to the latter helix scale in the PCA of our combined ID datasets. As can be seen in Figure 2F, significant separation is observed between our different datasets using these three intrinsic sequence properties. In general, the folded domains occupy a region with $\varphi > 0.08$, and the greatest separation between the two disordered sets is observed in the α-helix/ν_{model} plane.

When this approach is used to assess homopolymers of the common amino acids by their placement into a plot of hydrophobicity, α-helix propensity, and ν_{model}, the homopolymer results predict that Trp, Cys, Phe, Ile, Tyr, Val, Leu, His, and Met are “order promoting” amino acids,

while Ala, Arg, Gln, Pro, Glu, Lys, and Asp are "disorder promoting", and Asn, Ser, Thr, and Gly are "phase separation promoting" (Figure S3), similar to what we found previously (Figure 1B).

Predicting folded, ID, and phase-separating protein regions from sequence. Next, we used the separation obtained from this method to identify protein sequences belonging to folded, ID or PS ID categories, analogous to what we did for ParSe. Our aim was to see if using these top-performing scales would provide better predictions of PS ID domains. We modified the algorithm making a second-generation version, ParSe version 2 (v2). In this version, as with the original (42), we apply a 25-residue window and then slide this window across a whole sequence in 1-residue steps (Figure 3A) to label individual amino acids as either P (for PS ID), D (for ID) or F (for folded), and then to regions that are at least 90% of any one of these labels (see Methods, Figure 3C). Both ParSe v1 and v2 accurately delineate regions of Sup35 that have been experimentally determined (60) to behave as ID, PS ID, or folded regions (Figure 3C), and good accuracy is similarly found for other well-studied proteins (3, 17, 22, 68–72) utilizing diverse reported mechanisms driving LLPS (Figure S4).

One advantage of our algorithm is that it is very fast, and so can easily be applied to large datasets, e.g., the human proteome. We measured the prevalence of protein regions predicted by ParSe v2 to have LLPS potential in the human proteome (Figure 4) by two methods. First, as previously, we measured the longest predicted region with high LLPS potential (contiguous regions that are at least 90% labeled P). The results from ParSe v2 are mostly identical to results obtained previously using ParSe (42), whereby only ~5% of proteins in the human proteome have a predicted P-labeled region that is at least 50 residues in length. Disordered regions taken from DisProt (minus the LLPS annotated IDPs) (53, 54) and folded regions taken from SCOPe (Structural Classification of Proteins extended, version 2.07) (73, 74) gave results mirroring the human proteome result in the sense that these sequences are mostly devoid of long regions predicted to have high LLPS potential. In contrast, the 43 proteins assembled by Vernon *et al* (16) that have been verified *in vitro* to exhibit homotypic phase separation behavior tend to contain long stretches labeled P by ParSe v2, with ~90% of this set having predicted PS regions ≥50 residues in length. Only ~63% of the 98 parent proteins from which the PS ID set was derived have predicted PS regions ≥50 residues, wherein not all in this set have been shown to phase separate as purified components.

Second, we developed a numerical score to give a quantitative measure of the confidence of our assignment of P, F and D labels, and to give a single metric to define the LLPS potential of every protein. Our justification for using a single numerical score is, in part, the dominance of a single principal component in the PCA of the combined ID set (Figure 2E), although we generalized this approach to F-labeled positions as well. In the combined ID datasets, most of the variability was in a single direction nearly orthogonal to the line separating P and D sectors in our plot. As such, we used the linear distance of a 25 amino-acid window into its classifier sector (i.e., F, D, or P sector), relative to the cutoff boundary and normalized by the distance to the boundary of the training set mean (Figure 3B). Values greater than 1 in this *classifier distance* indicate a window located at a distance further from the sector boundary than the distance of the training set mean, whereas values less than 1 indicate a window closer to the cutoff boundary than the training set mean and, as such, possibly with some uncertainty for its classifier label. Classifier distances calculated from the Sup35 sequence are shown in Figure 3D, wherein window values have been assigned to the central residue of the window, as we did with the window label.

We used the summed classifier distance for every window labeled P to obtain an overall score for each protein. This method is more robust to situations where multiple smaller regions drive phase separation (Figure S4), as compared to, e.g., Sup35, where a single domain drives phase separation. Windows labeled either F or D do not contribute to this sum. The assumption we are making is that single regions that promote phase separation are sufficient to drive phase separation of larger proteins. This is consistent with the observation that many PS proteins still undergo LLPS when they contain other protein regions or GFP tags (2, 75). As before, we found that only a small fraction of the human proteome consists of proteins with IDRs driving phase separation (42). Indeed, using a cutoff for the summed classifier distance of 100 retains 100% and 76% of the proteins in the Vernon et al *in vitro* sufficient set and the parent proteins of our PS ID set, respectively. In contrast, only 10% of human proteins are predicted to drive phase separation through their IDRs by this cutoff (Figure 4B). Because we are focused solely on IDRs which drive phase separation, excluding multivalent interactions that involve ordered domains, nucleic acids, or other drivers of LLPS, the total number of LLPS drivers is somewhat larger than this.

We used this whole protein metric (the summed classifier distance of P-labeled windows) to create a recall plot, used to assess prediction performance, for multiple datasets (Figures 4C, S5). The success in recall plots is typically quantified using the area under the curve (AUC), when comparing a test dataset to a comparison dataset (47, 76, 77). Here, in all cases, we used the human proteome as the comparison dataset. The SCOPe database and DisProt (excluding LLPS annotated entries) both have AUC values < 0.5 (Figure 4C), indicating that the human proteome does contain more proteins predicted to drive LLPS than these negative control groups. As a result, this approach likely gives a lower bound on the success of a predictor. As expected, our calculated AUC using ParSe v2 is highest on the *in vitro* sufficient LLPS drivers from Vernon *et al* (AUC = 0.99, Figure 4C), which constitute a significant fraction of our positive control dataset (i.e., the parent proteins of the PS ID set). This is likely both because this is the dataset we used for training and because it is also the most highly curated dataset. To further test its efficacy, we measured AUC values for ParSe v2 on datasets of LLPS drivers curated by other groups (16, 47, 76–78), and found it to perform quite well, with AUC values >0.8 (Figure S5).

Figures 4A and S6 suggest ParSe v2 is an improvement (i.e., slightly higher recall), albeit marginally, compared to the original ParSe. The strong performance of ParSe v1 is, in part, because even in the original version, we used scales that gave strong separation between datasets. Utilizing scales with weaker predictive value leads to a less efficient predictor, as expected (Figure S7).

We then sought to compare ParSe to other published predictors. Although their data are not as highly curated as others, recent published work by Chen et al included predictions from multiple predictors on a publicly available dataset, facilitating comparison to other LLPS predictors (76). Of note, the negative control set in Chen et al contains, by our prediction, a higher fraction of IDRs driving phase separation than the human proteome (Figure S8D), although whether this is a problem with the database or with our prediction method is unclear. On their datasets, ParSe performs similarly as measured by AUC scores, to PScore (16), CatGranule (34), and PLAAC (32) in identifying proteins that drive LLPS (Figure S8A-C). The quality of the test one can make of these predictors depends significantly on the quality of the datasets, and so a true test of these

predictors will require significantly more experimental data from both positive and negative controls (31, 47).

Predicting the effects from mutation on phase separation behavior. Despite its simplicity, ParSe can predict the IDR(s) driving phase separation for a wide range of known phase-separating proteins, including FUS, Ddx4, LAF1, and A1. Several of these proteins have been the targets of mutagenesis studies implicating specific interactions between amino acids (i.e., cation-π or cation-anion) in the formation of phase-separated droplets. Cation-π interactions are thought to occur between different amino acids in the chain, and the balance of residues, e.g., Arg and Tyr, is thought to be important for LLPS (16, 22, 36). Similarly, net charge per residue, as opposed to simply the number of negative or positive charges (39), as well as the specific charge pattern (27), are also thought the be key determinants of LLPS.

Because ParSe is based only on the amino acid composition, and so does not include these higher-order effects involving combinations of amino acid types, we hypothesize that ParSe will have little predictive value for mutations that specifically alter the ratio of these pairwise interactions. More generally, we sought to determine if ParSe v2 could model the effects on phase separation behavior arising from mutations in the protein sequence. We hypothesize that sequence changes targeting P-labeled positions would have the greatest ability to modulate phase separation behavior. To assess this idea, we used the classifier distance whereby a phase separation "potential" was modeled as the summed classifier distance of P-labeled windows in the protein, as we did above in the recall plots. We compared the summed classifier distance with quantitative measures of LLPS behavior from four mutational studies involving three IDRs that individually exhibit LLPS behavior *in vitro* as purified components (3, 18, 27, 39) with sets of published mutations modulating either charge patterning or π-based interactions (Figures 5, S9).

As the different studies used different metrics to quantitatively assess phase separation, we first began by simply asking whether the summed classifier distance could accurately reproduce the rank ordering of variants. In Figure 5A, we ordered, from left-to-right, in decreasing phase separation "potential" as reported within each individual study the mutant and wild type sequences. Shown is the summed classifier distance of P-labeled windows. In the LAF-1 RGG study (27), mutants forming phase separated droplets at elevated temperatures indicated increased phase separation potential, whereas changes in the saturation concentration, c_{sat}, at a given temperature was used in studies with A1-LCD (18, 39). However, the mutant rank order in c_{sat} can change with the temperature; caused by differences in the standard molar enthalpy associated with phase separation, $\Delta h°$, which reflects the temperature dependence to c_{sat}. To manage this issue, mutant data were separated into two sets. One set corresponding to those mutants with experimental c_{sat} at 4 °C (Table S6), and a second corresponding to those mutants with experimental $\Delta h°$, $\Delta s°$, and $\Delta g°$ (Table S5). Figure 5A shows rank order in $\Delta h°$ for the A1-LCD mutants. Figure S9 ranks the A1-LCD mutants according to c_{sat} at 4 °C. The summed classifier distance (i.e., ParSe v2 predicted PS potential) of each mutant trended somewhat with the experimental rank order, correctly predicting an increase or decrease relative to the wild type in ~60% of the mutants as presented in Figure 5 (i.e., with A1-LCD mutants ranked by $\Delta h°$) and ~65% in Figure S9 (i.e., with A1-LCD mutants ranked by c_{sat}). Thus, ParSe is only moderately able to predict the effects of mutations designed to disrupt pairwise interactions between amino acids such as those arising from aromatic,

cation-π, and charge-based interactions. This performance is similar to the performance of PScore, PLAAC, and catGranule (Figure S10).

To test the importance of pairwise interactions, we explicitly included different types of interactions in our model to try to account for these contributions and possibly improve the trend of calculated potential *versus* observed phase separation behavior. We expanded our calculation of LLPS potential to include both the summed P classifier distance and terms quantifying the effects of interactions between amino acids, termed U_π for π-π and cation-π interactions, and U_q for charge-based effects. The contribution of these terms toward predicting the effects of mutations can give information on the relative importance of the individual terms. We used c_{sat}, $\Delta h°$, $\Delta s°$, and $\Delta g°$ separately to train this calculation; via 31 A1-LCD variants with c_{sat} and 27 A1-LCD and Ddx4 variants with Δh, $\Delta s°$, and $\Delta g°$ (Figures 5, S9). As c_{sat} is highly sensitive to the temperature (39), we expected the thermodynamic properties to be the more reliable metrics of LLPS. Indeed, we were best able to predict the effects of sequence changes on the measured $\Delta h°$ (Figure 5E). The predicted PS potential combining summed classifier distance with U_π and U_q correctly predicts the directional change relative to wild type in ~90% of the mutants when U_π and U_q were trained against $\Delta h°$, and the correlation between experimentally measured $\Delta h°$ and ParSe-calculated PS potential was reasonably high (R=0.76; Figure 5C). Thus, explicit consideration of interactions between amino acid types is important for determining PS potential in these mutational studies. It remains to be seen whether ParSe is able to accurately predict PS potential of mutants designed to test other aspects of LLPS, such as its dependence on the presence of partner molecules or on a specific set of solution conditions (e.g., pH, ionic strength, temperature).

Finally, we sought to determine what effect including these corrections to ParSe had on the identification of proteins driving phase separation. Overall, including U_π and U_q into ParSe increases the number of proteins identified that drive phase separation in both the PS sets and the human proteome (Figure S11). As a result, the AUC when comparing either our PS ID set or the Vernon highly curated set to the human proteome is slightly reduced. However, whether this is a result of correctly classifying more human proteins as driving LLPS, or whether we have simply increased the false negative rate remains to be seen.

Discussion

In this work, we focused on identifying IDRs that drive phase separation, with a particular focus on separating PS IDRs from conventional IDRs that do not drive phase separation. Using carefully curated datasets of ID, PS ID and folded domains (Figures 1, 2), we developed a sequence-based predictor of phase separation (ParSe; Figure 3) which is fast enough to scan the entire human proteome in minutes on a single computer, and as or more accurate than other published predictors in identifying both proteins and regions within proteins that drive phase separation (Figures 2, 3, S5, S8, S10). We recognized that a wide variety of amino acid scales show significant differences between the ID and PS ID datasets, indicating that PS IDRs are a robustly different class of protein region than non-phase separating IDRs (Figure 2). We conclude that a redundant combination of molecular mechanisms driving cohesive interactions between amino acids is likely at play. This helps to explain why our general predictor of IDR hydrodynamic size (ν_{model}) is a strong indicator of LLPS potential, as we found previously (42). Moreover, by including interactions between

amino acids thought to drive phase separation, we were able to match existing data on mutant sequences (Figure 5). This extension highlights the importance of pair-wise interactions in modulating phase separation.

While our approach has proved very successful, it, like other approaches to this problem, has significant limitations, including limitations in predicting responses to changes in solvent, limitations of the datasets, and limitations of the constraints of the approach chosen. The formation of phase separated droplets by polymer chains is a result, very generally, of interactions between chains that are stronger than the interactions of the chain with the solvent. As a result, LLPS is strongly dependent on the solution environment. Within cells, there are many proteins which assemble into membraneless organelles only within specific cellular conditions, e.g., upon lowering of pH (60). To accurately predict which solution conditions drive phase separation of any individual protein domain would require a detailed understanding of which mechanisms proteins use to drive phase separation, how those mechanisms are modulated by solutions conditions, and how cells modulate solution conditions in different cellular states. As a first step in this process, our aim is to simply improve identification of which IDRs and which potential mechanisms are used by IDRs to drive phase separation in a variety of cellular and solution conditions. Thus, although our predictor has high success in identifying proteins that have been seen experimentally to drive phase separation, we do not yet distinguish between responses to different cellular conditions, or, e.g., upper- versus lower-critical temperature. The temperature dependence of hydrophobicity scales as used by Dignon et al (79) could be a potential future approach to do this.

A primary limitation of our work, as well as others, is that even our well curated datasets have misidentified regions. For example, because the IDR in a protein that is responsible for phase separation has not always been identified, we simply used all IDRs from known phase-separating proteins. As a result, our PS ID set likely includes some IDRs which are not involved in phase separation. Similarly, our ID set was curated from proteins that have not yet been identified to phase separate, including those with experiments done at high protein concentration. However, the lack of observation of phase separation at any one experimental condition does not preclude its formation. Indeed, a long history of solution screening for crystallography would indicate that protein behavior can vary dramatically based on solution conditions (80). However, it appears that our PS ID and ID datasets are sufficiently enriched or depleted for PS IDRs for us to identify key properties of IDRs that drive phase separation. For example, the performance of our predictor is improved as the rigor with which the dataset was curated improves. ParSe gives the highest AUC on the dataset from Vernon et al containing only those proteins shown to drive homotypic LLPS *in vitro*, compared to datasets containing PS drivers more generally, and weaker still on datasets including both LLPS drivers and proteins recruited to existing droplets (Figure S5) (16, 76, 77).

Our approach is based primarily on sequence composition and not on sequence patterning or combinations of amino acids. It is surprising how effective this strategy is and how many different scales can be used to distinguish PS IDRs successfully. Nevertheless, our approach, while fast and effective, is unable to identify pairwise protein interactions that contribute to LLPS. In our analysis of mutants, we introduced a simple potential whereby amino acid pairs are counted, and this clearly improves the ability to predict the effects of mutation on phase separation (Figure 5). Pairwise interaction patterns are probably better identified by machine learning algorithms or simulation (27, 28, 40, 41, 47, 77). However, the efficacy of our approach appears to indicate that the primary

determinant of whether any one sequence will phase separate depends on the overall amino composition, whereas rearrangements, mutations, or post-translational modifications of that base sequence will modulate that propensity for phase separation. Thus, it appears that the identification of sequences that have the potential to phase separate is an easier problem than identifying how mutation of a few residues will impact that phase separation potential. This result is not specific to our predictor, as none of the predictors tested here showed significantly better correlation with changes in phase separation potential upon mutation (Figure S10). We additionally note that different experimental measurements of LLPS potential give different ordering of mutants (Figures 5, S9), further compounding the issue.

Finally, our approach differs from several others in that we are focused solely on the problem of separating PS IDRs from IDRs that do not phase separate (47, 76). We are thus not able to identify proteins that utilize multivalent interactions between folded domains and other folded, ID, or nucleic acid binding domains as a primary mechanism for driving phase separation (5, 6, 24, 25). Moreover, we are primarily focused on IDRs that drive phase separation, as opposed to those that are recruited to existing phase separated droplets, a case which has been recently considered by Chen et al (76). Our motivation for this narrow focus is that a broader focus might obscure mechanisms used only by PS IDRs, and that interactions between folded domains are, in general, better understood than those between disordered domains.

The strong performance of ParSe on existing datasets, the robust nature of differences between PS IDRs and conventional IDRs, and the high correlation between ParSe and other predictors on databases of phase separating proteins all give confidence that ParSe is able to identify PS IDRs with significant accuracy. Because of its speed, ParSe can easily be applied to datasets of arbitrarily large size. As an example, we measured the summed classifier distance for the human proteome and found that only a small fraction of the human proteome is likely to drive phase separation (Figure 4B). Moreover, we identified the 500 proteins with the highest summed classifier distance in the human proteome, as well as their longest predicted PS ID region (Table S7). Many proteins involved in stress granule formation, RNA processing, and other functions known to be associated with membraneless organelles are identified in this process. However, many proteins are also identified that are not yet associated with a biological process driven by phase separation. This suggests that, while the fraction of human proteins driving phase separation may be small, not all of the biological processes relying on phase separation have yet been identified.

Experimental procedures

Protein databases

A set of 224 IDRs from proteins that exhibit LLPS behavior, used for the PS ID set, was obtained from our prior work (42). For the ID set, we started with 23 IDR sequences used previously (42), and then added all DisProt consensus ID sequences not having the disorder function ontology identifier for LLPS, IDPO:00041 (54). Protein sequences in the BMRB (52) with “disordered” or “IDP” as a keyword or in the entry title were also added to the ID set. BMRB obtained sequences were restricted to those with ≥70% of residue positions classified as disordered by Wishart’s random coil index, using an S^2 cutoff of 0.6 (81). DisProt and BMRB sequences were culled by

Metapredict (55), keeping only those predicted to be ID, and seqatoms (57), excluding those that were highly homologous to folded regions of proteins in the PDB. The folded set started with the 82 folded sequences used previously (42), and then added a set of human proteins with nonhomologous structures (48), proteins with small to large structures (49), extremophile proteins (50), metamorphic proteins (51), and membrane proteins that were found by searching the PDB (56) for the phrase "membrane protein." Using the PISCES Server (82), the human, extremophile, metamorphic, and membrane proteins had a maximum of 50% sequence identity within each folded subset and only X-ray structures with a resolution better than 2.5 Å.

Calculation of β-turn propensity and ν_{model}

The propensity to form β-turn structures was calculated by $\sum scale_i/N$, where $scale_i$ is the value for amino acid type i in the normalized frequencies for β-turn from Levitt (83). The summation is over the protein sequence containing N number of amino acids. ν_{model} was introduced previously (42) as a phenomenological substitute to the polymer scaling exponent (84, 85) and used to normalize protein hydrodynamic size to the chain length,

$$\nu_{model} = log(R_h/R_o)/log(N), \quad [1]$$

where R_o is a constant set to 2.16 Å, and the hydrodynamic radius, R_h, is calculated from sequence using an equation found to be accurate for monomeric IDPs (43, 44, 86–88). The equation to calculate R_h for a disordered sequence is,

$$R_h = 2.16\text{Å} \cdot N^{(0.503-0.11 \cdot ln(f_{PPII}))} + 0.26 \cdot |Q_{net}| - 0.29 \cdot N^{0.5}, \quad [2]$$

where f_{PPII} is the fractional number of residues in the PPII conformation, and Q_{net} is the net charge. f_{PPII} is estimated from $\sum P_{PPII,i}/N$, where $P_{PPII,i}$ is the experimental PPII propensity determined for amino acid type i in unfolded peptides (89) and the summation is over the protein sequence. Q_{net} is determined from the number of lysine and arginine residues minus the number of glutamic acid and aspartic acid.

Principal component analysis

The statistical program R (90) was used to perform PCA on the sequence sets, and the packages ggfortify, ggplot2, factoextra, MetBrewer, and tidyverse were used to render the results. In the PCA, the variables were shifted to be zero centered and scaled to unit variance.

ParSe v2 algorithm

For an input primary sequence, whereby the amino acids are restricted to the 20 common types, ParSe v2 first reads the sequence to determine its length, N. Next, the algorithm uses a sliding window scheme (Figure 3A) to calculate ν_{model}, α-helix propensity, and ϕ for every 25-residue segment of the primary sequence. This window scheme can be applied to proteins with N >25. R_h

is calculated by Equation 2, which in turn is used to determine v_{model} by Equation 1, by the same method used in the original ParSe described previously (42). α-helix propensity is calculated as the sequence sum divided by N using the scale by Tanaka and Scheraga (67). ϕ is calculated as the sequence sum divided by N using the hydrophobicity scale by Vendrusculo and coworkers (66). A window is labeled F if ϕ >0.08 (Figure 3B). If ϕ <0.08, a window is labeled P or D depending on the values of v_{model} and α-helix propensity. Windows with high α-helix propensity and high v_{model} are labeled D, while those with low α-helix propensity and low v_{model} are labeled P. The P/D boundary was determined by the line that bisects the overlapping distributions of v_{model} and α-helix propensity in the PS ID and ID sets, given by v_{model} = -0.244•α-helix propensity + 0.789. The window label is assigned to the central residue in that window. N- and C-terminal residues not belonging to a central window position are assigned the label of the central residue in the first and last window, respectively, of the whole sequence. Protein regions predicted by ParSe v2 to be PS, ID, or folded are determined by finding contiguous residue positions of length ≥20 that are ≥90% of only one label P, D, or F, respectively. When overlap occurs between adjacent predicted regions, owing to the up to 10% label mixing allowed, this overlap is split evenly between the two adjacent regions.

Classifier distance calculation

The classifier distance is the normalized distance of a ParSe v2 generated window into its classifier sector (i.e., F, D, or P sector) and relative to the cutoff boundary (Figure 3B). For F labeled windows, the classifier distance is ϕ (of the window) minus the cutoff value of 0.08, and then normalized to distance of the folded set mean ϕ (0.1164) to the cutoff. Specifically, this is (ϕ – 0.08)/(0.1164 – 0.08). For P or D labeled windows, first we find the point on the P/D boundary (v_{model} = -0.244 · α-helix propensity + 0.789) that makes a perpendicular bisector when paired with the window values of v_{model} and α-helix propensity. Then the distance between this point and the point defined by the window values of v_{model} and α-helix propensity is determined. Specifically, this distance is sqrt((α – x) · (α – x) + (v_{model} – y) · (v_{model} – y)) where α is the α-helix propensity, x is (α/0.244 + 0.789 – v_{model})/(0.244 + 1/0.244) and y is (x – α)/0.244 + v_{model}. This distance is normalized by dividing by 0.019 (the distance from the boundary to either of the set means).

PSCORE calculation

PSCORE, which is a phase separation propensity predictor (16), was calculated by computer algorithm using the Python script and associated database files available at https://doi.org/10.7554/eLife.31486.022.

Granule propensity calculation

Granule propensity was calculated by using the catGranule (34) webtool available at http://www.tartaglialab.com.

PLAAC LLR calculation

LLR score, which identifies prion-containing sequences (91), was calculated by using the webtool available at http://plaac.wi.mit.edu.

Metapredict calculation

Metapredict score (55), which predicts the presence of ID in a sequence, was calculated by computer algorithm using the Python script available at http://metapredict.net.

Calculation of U_π

The relative contributions of aromatic and cation-π interactions to LLPS in our calculations followed the observed rank order by Wang *et al*: Tyr-Arg > Tyr-Lys ~ Phe-Arg > Phe-Lys (22). To mimic this ranking, we assumed 3:2:1 weighting and, also, that Phe-Tyr interactions would contribute comparably to Phe-Lys interactions,

$$\begin{aligned} U_\pi = a \cdot (&3\cdot(\#Y \times \#R / (\#Y - \#R)_{\#Y \neq \#R}) \\ &+2\cdot(\#Y \times \#K / (\#Y - \#K)_{\#Y \neq \#K}) + 2\cdot(\#F \times \#R / (\#F - \#R)_{\#F \neq \#R}) \\ &+1\cdot(\#F \times \#K / (\#F - \#K)_{\#F \neq \#K}) + 1\cdot(\#F \times \#Y / (\#F - \#Y)_{\#F \neq \#Y})). \end{aligned} \quad [3]$$

In Equation 3, #Y, #R, #F, and #K represent the number of Tyr, Arg, Phe, and Lys residues, respectively, in a sequence, calculated on a per-window basis, and a is a fitting parameter (see below). Thus, U_π increases with increasing Tyr, Arg, Phe, and Lys content, and more so when interaction partners are present at similar levels. When the divisor is zero (e.g., when #Y = #R), it is changed to 1 to avoid infinite potentials.

Window-specific U_π was added to the classifier distance at windows labeled P. Moreover, U_π was applied to D-labeled windows, allowing for the possibility of labels changing from D to P. This would occur when the value for U_π was larger than the classifier distance at a D-labeled window. Thus, protein regions that otherwise have characteristics more like the ID set, in ν_{model} and α-helix propensity, could be labeled P if U_π was large enough. When this occurs, the given classifier distance was determined by the difference between U_π and the original classifier distance of the window formerly labeled D.

The parameter a in Equation 3 was determined by finding the optimal correlation of ParSe-calculated PS potential to $\Delta h°$ (finding $a = 0.14$; Figure 5B), $\Delta s°$ (finding $a = 0.08$), $\Delta g°$ (finding $a = 0.11$), or c_{sat} (finding $a = 0.28$). In each case, the mutants used to fit a were limited to the subset with identical charge and charge patterns, determined by calculating the net charge per residue, *NCPR*, and sequence charge decoration, *SCD*, of each sequence. *NCPR* is the number of Lys and Arg residues minus the number of Glu and Asp residues, divided by N. *SCD* is calculated by $N^{-1}\sum_i\sum_{j>i}(q_i q_j)|j-i|^{1/2}$, where q is the amino acid specific charge (92).

Calculation of U_q.

To model the contributions of charge-based interactions to LLPS, we build upon the observations by Schuster et al (27) and Bremer et al (39) that changes in *SCD* and *NCPR*, respectively, can affect phase separation potential. Accordingly, a simple charge-based potential was defined,

$$U_q = b \cdot SCD + c \cdot |NCPR|, \quad [4]$$

where b and c are fitting parameters, and U_q is calculated on a per-window basis. U_q is added to the classifier distance at each window labeled P, and is applied to windows labeled D, following the scheme described above for U_π, again allowing for the possibility of labels changing from D to P. As with a, the parameters b and c were fixed by finding the optimal correlation of calculated PS potential and $\Delta h°$ (finding 8.4 and 5.6, respectively; Figure 5C), $\Delta s°$ (finding 4.6 and 7.0, respectively), $\Delta g°$ (finding 5.2 and 5.4, respectively), or c_{sat} (finding -16.0 and 33, respectively).

Calculation of $\Delta h°$, $\Delta s°$, and $\Delta g°$ from temperature dependence to c_{sat}.

For some Ddx4 and A1-LCD sequences, $\Delta h°$ and $\Delta s°$ (and thus $\Delta g°$) were not available, but c_{sat} measured at different temperatures has been reported (3, 18). For these proteins, the standard molar chemical potential, $\mu°$, was used to relate c_{sat} in the dilute and dense phases, c_{dilute} and c_{dense}, respectively, to the standard molar enthalpy and entropy associated with phase separation (39),

$$\begin{aligned}\Delta\mu° &= \Delta g° = \Delta h° - T \cdot \Delta s° \\ &= \mu°_{dense} - \mu°_{dilute} \\ &= \mu_{dense} - R \cdot T \cdot \ln(c_{dense}/c_{ref}) - (\mu_{dilute} - R \cdot T \cdot \ln(c_{dilute}/c_{ref})) \\ &= R \cdot T \cdot \ln(c_{dilute}/c_{dense}),\end{aligned}$$

where $\mu_{dense} - \mu_{dilute}$ is zero at equilibrium, R is the universal gas constant, and T is temperature. By plotting the natural logarithm of c_{sat} at different temperatures, a linear fit versus $1/T$ yields $\Delta h°$ and $\Delta s°$. For A1-LCD mutants, 0.03 M was used for c_{dense} (39). For Ddx4 mutants, 0.01 M was used for c_{dense} (3). $\Delta g°$ was calculated from $\Delta h° - T \cdot \Delta s°$ and the standard temperature (273.15 K).

Data availability

The Parse v2 algorithm written in Fortran, Parse_v2.f, can be downloaded at https://github.com/stevewhitten/ParSe_v2. A webtool version can be used at https://stevewhitten.github.io/Parse_v2_web.

Author contributions—L. E. H. and S. T. W. conceptualization; A. Y. I., N. P. K., D. L. A., J. J. C., K. A. L., N. C. F., L. E. H., and S. T. W., formal analysis; K. A. L., N. C. F., L. E. H., and S. T. W., investigation; L. E. H and S. T. W., methodology; K.A.L., N. C. F., L. E. H, and S. T. W., writing–original draft; J. J. C., writing–review and editing.

Funding and additional information—This work was supported by the National Institutes of Health under grants R25GM102783 (South Texas Doctoral Bridge Program; B. O. Oyajobi and S. T. W.), R35GM119755 (L. E. H.), T32GM065103 (J. H. A.), and R01AI139479 (N. C. F.), as well as the National Science Foundation under grants 1818090 (N. C. F.) and 1943488 (L. E. H.). No nongovernmental sources were used to fund this project. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or NIH.

Conflict of interest—The authors declare that they have no conflicts of interest with the contents of this article.

References

1. Gomes, E., and Shorter, J. (2019) The molecular language of membraneless organelles. *J. Biol. Chem.* **294**, 7115–7127
2. Brangwynne, C. P., Eckmann, C. R., Courson, D. S., Rybarska, A., Hoege, C., Gharakhani, J., Jülicher, F., and Hyman, A. A. (2009) Germline P Granules Are Liquid Droplets That Localize by Controlled Dissolution/Condensation. *Science*. **324**, 1729–1732
3. Brady, J. P., Farber, P. J., Sekhar, A., Lin, Y.-H., Huang, R., Bah, A., Nott, T. J., Chan, H. S., Baldwin, A. J., Forman-Kay, J. D., and Kay, L. E. (2017) Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E8194–E8203
4. Mitrea, D. M., and Kriwacki, R. W. (2016) Phase separation in biology; functional organization of a higher order. *Cell Communication and Signaling*. **14**, 1
5. Su, X., Ditlev, J. A., Hui, E., Xing, W., Banjade, S., Okrut, J., King, D. S., Taunton, J., Rosen, M. K., and Vale, R. D. (2016) Phase separation of signaling molecules promotes T cell receptor signal transduction. *Science*. **352**, 595–599
6. Bouchard, J. J., Otero, J. H., Scott, D. C., Szulc, E., Martin, E. W., Sabri, N., Granata, D., Marzahn, M. R., Lindorff-Larsen, K., Salvatella, X., Schulman, B. A., and Mittag, T. (2018) Cancer Mutations of the Tumor Suppressor SPOP Disrupt the Formation of Active, Phase-Separated Compartments. *Mol Cell*. **72**, 19-36.e8
7. Uversky, V. N., Kuznetsova, I. M., Turoverov, K. K., and Zaslavsky, B. (2015) Intrinsically disordered proteins as crucial constituents of cellular aqueous two phase systems and coacervates. *FEBS Letters*. **589**, 15–22
8. Shin, Y., and Brangwynne, C. P. (2017) Liquid phase condensation in cell physiology and disease. *Science*. **357**, eaaf4382
9. Li, X.-H., Chavali, P. L., Pancsa, R., Chavali, S., and Babu, M. M. (2018) Function and Regulation of Phase-Separated Biological Condensates. *Biochemistry*. **57**, 2452–2461
10. Prouteau, M., and Loewith, R. (2018) Regulation of Cellular Metabolism through Phase Separation of Enzymes. *Biomolecules*. **8**, 160
11. Jacobs, M. I., Jira, E. R., and Schroeder, C. M. (2021) Understanding How Coacervates Drive Reversible Small Molecule Reactions to Promote Molecular Complexity. *Langmuir*. **37**, 14323–14335

12. Zhang, Y., Narlikar, G. J., and Kutateladze, T. G. (2021) Enzymatic Reactions inside Biological Condensates. *J Mol Biol*. **433**, 166624
13. Alberti, S., and Dormann, D. (2019) Liquid–Liquid Phase Separation in Disease. *Annual Review of Genetics*. **53**, 171–194
14. Tsang, B., Pritišanac, I., Scherer, S. W., Moses, A. M., and Forman-Kay, J. D. (2020) Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations. *Cell*. **183**, 1742–1756
15. Prasad, A., Bharathi, V., Sivalingam, V., Girdhar, A., and Patel, B. K. (2019) Molecular Mechanisms of TDP-43 Misfolding and Pathology in Amyotrophic Lateral Sclerosis. *Front Mol Neurosci*. **12**, 25
16. Vernon, R. M., Chong, P. A., Tsang, B., Kim, T. H., Bah, A., Farber, P., Lin, H., and Forman-Kay, J. D. (2018) Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife*. 10.7554/eLife.31486
17. Das, S., Lin, Y.-H., Vernon, R. M., Forman-Kay, J. D., and Chan, H. S. (2020) Comparative roles of charge, π, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *Proc Natl Acad Sci U S A*. **117**, 28795–28805
18. Martin, E. W., Holehouse, A. S., Peran, I., Farag, M., Incicco, J. J., Bremer, A., Grace, C. R., Soranno, A., Pappu, R. V., and Mittag, T. (2020) Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*. **367**, 694–699
19. Dannenhoffer-Lafage, T., and Best, R. B. (2021) A Data-Driven Hydrophobicity Scale for Predicting Liquid-Liquid Phase Separation of Proteins. *J Phys Chem B*. **125**, 4046–4056
20. Murthy, A. C., Dignon, G. L., Kan, Y., Zerze, G. H., Parekh, S. H., Mittal, J., and Fawzi, N. L. (2019) Molecular interactions underlying liquid-liquid phase separation of the FUS low-complexity domain. *Nat Struct Mol Biol*. **26**, 637–648
21. Murthy, A. C., Tang, W. S., Jovic, N., Janke, A. M., Seo, D. H., Perdikari, T. M., Mittal, J., and Fawzi, N. L. (2021) Molecular interactions contributing to FUS SYGQ LC-RGG phase separation and co-partitioning with RNA polymerase II heptads. *Nat Struct Mol Biol*. **28**, 923–935
22. Wang, J., Choi, J.-M., Holehouse, A. S., Lee, H. O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D., Poser, I., Pappu, R. V., Alberti, S., and Hyman, A. A. (2018) A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell*. **174**, 688-699.e16
23. Choi, J.-M., Dar, F., and Pappu, R. V. (2019) LASSI: A lattice model for simulating phase transitions of multivalent proteins. *PLOS Computational Biology*. **15**, e1007028
24. Dao, T. P., Kolaitis, R.-M., Kim, H. J., O'Donovan, K., Martyniak, B., Colicino, E., Hehnly, H., Taylor, J. P., and Castañeda, C. A. (2018) Ubiquitin Modulates Liquid-Liquid Phase Separation of UBQLN2 via Disruption of Multivalent Interactions. *Mol Cell*. **69**, 965-978.e6
25. Yang, Y., Jones, H. B., Dao, T. P., and Castañeda, C. A. (2019) Single Amino Acid Substitutions in Stickers, but Not Spacers, Substantially Alter UBQLN2 Phase Transitions and Dense Phase Material Properties. *J. Phys. Chem. B*. **123**, 3618–3629
26. Pak, C. W., Kosno, M., Holehouse, A. S., Padrick, S. B., Mittal, A., Ali, R., Yunus, A. A., Liu, D. R., Pappu, R. V., and Rosen, M. K. (2016) Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein. *Mol Cell*. **63**, 72–85

27. Schuster, B. S., Dignon, G. L., Tang, W. S., Kelley, F. M., Ranganath, A. K., Jahnke, C. N., Simpkins, A. G., Regy, R. M., Hammer, D. A., Good, M. C., and Mittal, J. (2020) Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. *Proc Natl Acad Sci U S A*. **117**, 11421–11431
28. Chu, X., Sun, T., Li, Q., Xu, Y., Zhang, Z., Lai, L., and Pei, J. (2022) Prediction of liquid-liquid phase separating proteins using machine learning. *BMC Bioinformatics*. **23**, 72
29. Crabtree, M. D., Holland, J., Kompella, P., Babl, L., Turner, N., Baldwin, A. J., and Nott, T. J. (2021) Repulsive electrostatic interactions modulate dense and dilute phase properties of biomolecular condensates. bioRxiv doi: 10.1101/2020.10.29.357863
30. Vernon, R. M., and Forman-Kay, J. D. (2019) First-generation predictors of biological protein phase separation. *Current Opinion in Structural Biology*. **58**, 88–96
31. Pancsa, R., Vranken, W., and Mészáros, B. (2021) Computational resources for identifying and describing proteins driving liquid-liquid phase separation. *Brief Bioinform*. 10.1093/bib/bbaa408
32. Lancaster, A. K., Nutter-Upham, A., Lindquist, S., and King, O. D. (2014) PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics*. **30**, 2501–2502
33. Orlando, G., Raimondi, D., Tabaro, F., Codicè, F., Moreau, Y., and Vranken, W. F. (2019) Computational identification of prion-like RNA-binding proteins that form liquid phase-separated condensates. *Bioinformatics*. **35**, 4617–4623
34. Klus, P., Bolognesi, B., Agostini, F., Marchese, D., Zanzoni, A., and Tartaglia, G. G. (2014) The cleverSuite approach for protein characterization: predictions of structural properties, solubility, chaperone requirements and RNA-binding abilities. *Bioinformatics*. **30**, 1601–1608
35. Mellacheruvu, D., Wright, Z., Couzens, A. L., Lambert, J.-P., St-Denis, N. A., Li, T., Miteva, Y. V., Hauri, S., Sardiu, M. E., Low, T. Y., Halim, V. A., Bagshaw, R. D., Hubner, N. C., Al-Hakim, A., Bouchard, A., Faubert, D., Fermin, D., Dunham, W. H., Goudreault, M., Lin, Z.-Y., Badillo, B. G., Pawson, T., Durocher, D., Coulombe, B., Aebersold, R., Superti-Furga, G., Colinge, J., Heck, A. J. R., Choi, H., Gstaiger, M., Mohammed, S., Cristea, I. M., Bennett, K. L., Washburn, M. P., Raught, B., Ewing, R. M., Gingras, A.-C., and Nesvizhskii, A. I. (2013) The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods*. **10**, 730–736
36. Forman-Kay, J. D., Ditlev, J. A., Nosella, M. L., and Lee, H. O. (2022) What are the distinguishing features and size requirements of biomolecular condensates and their implications for RNA-containing condensates? *RNA*. **28**, 36–47
37. Dignon, G. L., Zheng, W., Kim, Y. C., Best, R. B., and Mittal, J. (2018) Sequence determinants of protein phase behavior from a coarse-grained model. *PLOS Computational Biology*. **14**, e1005941
38. Zheng, W., Dignon, G. L., Jovic, N., Xu, X., Regy, R. M., Fawzi, N. L., Kim, Y. C., Best, R. B., and Mittal, J. (2020) Molecular Details of Protein Condensates Probed by Microsecond Long Atomistic Simulations. *J. Phys. Chem. B*. **124**, 11671–11679
39. Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., and Mittag, T. (2022) Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat. Chem.* **14**, 196–207
40. Tesei, G., Schulze, T. K., Crehuet, R., and Lindorff-Larsen, K. (2021) Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of

single-chain properties. *Proceedings of the National Academy of Sciences*. **118**, e2111696118
41. Joseph, J. A., Reinhardt, A., Aguirre, A., Chew, P. Y., Russell, K. O., Espinosa, J. R., Garaizar, A., and Collepardo-Guevara, R. (2021) Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nat Comput Sci*. **1**, 732–743
42. Paiz, E. A., Allen, J. H., Correia, J. J., Fitzkee, N. C., Hough, L. E., and Whitten, S. T. (2021) Beta turn propensity and a model polymer scaling exponent identify intrinsically disordered phase-separating proteins. *Journal of Biological Chemistry*. **297**, 101343
43. Tomasso, M. E., Tarver, M. J., Devarajan, D., and Whitten, S. T. (2016) Hydrodynamic Radii of Intrinsically Disordered Proteins Determined from Experimental Polyproline II Propensities. *PLoS Comput. Biol.* **12**, e1004686
44. English, L. R., Voss, S. M., Tilton, E. C., Paiz, E. A., So, S., Parra, G. L., and Whitten, S. T. (2019) Impact of Heat on Coil Hydrodynamic Size Yields the Energetics of Denatured State Conformational Bias. *J. Phys. Chem. B*. **123**, 10014–10024
45. Zhang, Y., Zai-Rose, V., Price, C. J., Ezzell, N. A., Bidwell, G. L., Correia, J. J., and Fitzkee, N. C. (2018) Modeling the Early Stages of Phase Separation in Disordered Elastin-like Proteins. *Biophys. J.* **114**, 1563–1578
46. Kawashima, S., and Kanehisa, M. (2000) AAindex: amino acid index database. *Nucleic Acids Res*. **28**, 374
47. Cai, H., Vernon, R. M., and Forman-Kay, J. D. (2022) An interpretable machine learning algorithm to predict disordered protein phase separation based on biophysical interactions. bioRxiv doi: 10.1101/2022.07.06.499043
48. Wang, S., Gu, J., Larson, S. A., Whitten, S. T., and Hilser, V. J. (2008) Denatured-State Energy Landscapes of a Protein Structural Database Reveal the Energetic Determinants of a Framework Model for Folding. *Journal of Molecular Biology*. **381**, 1184–1201
49. Fitzkee, N. C., and Rose, G. D. (2004) Reassessing random-coil statistics in unfolded proteins. *PNAS*. **101**, 12497–12502
50. Panja, A. S., Maiti, S., and Bandyopadhyay, B. (2020) Protein stability governed by its structural plasticity is inferred by physicochemical factors and salt bridges. *Sci Rep*. **10**, 1822
51. Chen, N., Das, M., LiWang, A., and Wang, L.-P. (2020) Sequence-Based Prediction of Metamorphic Behavior in Proteins. *Biophysical Journal*. **119**, 1380–1390
52. Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Kent Wenger, R., Yao, H., and Markley, J. L. (2008) BioMagResBank. *Nucleic Acids Research*. **36**, D402–D408
53. Piovesan, D., Tabaro, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C. J., Aspromonte, M. C., Davey, N. E., Davidović, R., Dosztányi, Z., Elofsson, A., Gasparini, A., Hatos, A., Kajava, A. V., Kalmar, L., Leonardi, E., Lazar, T., Macedo-Ribeiro, S., Macossay-Castillo, M., Meszaros, A., Minervini, G., Murvai, N., Pujols, J., Roche, D. B., Salladini, E., Schad, E., Schramm, A., Szabo, B., Tantos, A., Tonello, F., Tsirigos, K. D., Veljković, N., Ventura, S., Vranken, W., Warholm, P., Uversky, V. N., Dunker, A. K., Longhi, S., Tompa, P., and Tosatto, S. C. E. (2017) DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res*. **45**, D219–D227
54. Hatos, A., Hajdu-Soltész, B., Monzon, A. M., Palopoli, N., Álvarez, L., Aykac-Fas, B., Bassot, C., Benítez, G. I., Bevilacqua, M., Chasapi, A., Chemes, L., Davey, N. E.,

Davidović, R., Dunker, A. K., Elofsson, A., Gobeill, J., Foutel, N. S. G., Sudha, G., Guharoy, M., Horvath, T., Iglesias, V., Kajava, A. V., Kovacs, O. P., Lamb, J., Lambrughi, M., Lazar, T., Leclercq, J. Y., Leonardi, E., Macedo-Ribeiro, S., Macossay-Castillo, M., Maiani, E., Manso, J. A., Marino-Buslje, C., Martínez-Pérez, E., Mészáros, B., Mičetić, I., Minervini, G., Murvai, N., Necci, M., Ouzounis, C. A., Pajkos, M., Paladin, L., Pancsa, R., Papaleo, E., Parisi, G., Pasche, E., Barbosa Pereira, P. J., Promponas, V. J., Pujols, J., Quaglia, F., Ruch, P., Salvatore, M., Schad, E., Szabo, B., Szaniszló, T., Tamana, S., Tantos, A., Veljkovic, N., Ventura, S., Vranken, W., Dosztányi, Z., Tompa, P., Tosatto, S. C. E., and Piovesan, D. (2020) DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* **48**, D269–D276

55. Emenecker, R. J., Griffith, D., and Holehouse, A. S. (2021) Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys J.* **120**, 4312–4319
56. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Research*. **28**, 235–242
57. Brandt, B. W., Heringa, J., and Leunissen, J. A. M. (2008) SEQATOMS: a web tool for identifying missing regions in PDB in sequence context. *Nucleic Acids Res*. **36**, W255-259
58. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. (2001) Intrinsically disordered protein. *J Mol Graph Model*. **19**, 26–59
59. Campen, A., Williams, R. M., Brown, C. J., Meng, J., Uversky, V. N., and Dunker, A. K. (2008) TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept Lett*. **15**, 956–963
60. Franzmann, T. M., Jahnel, M., Pozniakovsky, A., Mahamid, J., Holehouse, A. S., Nüske, E., Richter, D., Baumeister, W., Grill, S. W., Pappu, R. V., Hyman, A. A., and Alberti, S. (2018) Phase separation of a yeast prion protein promotes cellular fitness. *Science*. 10.1126/science.aao5654
61. Welch, B. L. (1947) The Generalization of 'Student's' Problem when Several Different Population Varlances are Involved. *Biometrika*. **34**, 28–35
62. Mann, H. B., and Whitney, D. R. (1947) On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*. **18**, 50–60
63. Pearson, K. (1901) LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. **2**, 559–572
64. Kozlowski, L. P., and Bujnicki, J. M. (2012) MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics*. **13**, 111
65. Erdős, G., and Dosztányi, Z. (2020) Analyzing Protein Disorder with IUPred2A. *Current Protocols in Bioinformatics*. **70**, e99
66. Bastolla, U., Porto, M., Roman, H. E., and Vendruscolo, M. (2005) Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. *Proteins*. **58**, 22–30
67. Tanaka, S., and Scheraga, H. A. (1977) Statistical Mechanical Treatment of Protein Conformation. 5. Multistate Model for Specific-Sequence Copolymers of Amino Acids. *Macromolecules*. **10**, 9–20

68. Tremblay, M.-L., Xu, L., Lefèvre, T., Sarker, M., Orrell, K. E., Leclerc, J., Meng, Q., Pézolet, M., Auger, M., Liu, X.-Q., and Rainey, J. K. (2015) Spider wrapping silk fibre architecture arising from its modular soluble protein precursor. *Scientific Reports*. **5**, 11502
69. Muiznieks, L. D., and Keeley, F. W. (2016) Phase separation and mechanical properties of an elastomeric biomaterial from spider wrapping silk and elastin block copolymers. *Biopolymers*. **105**, 693–703
70. Elbaum-Garfinkle, S., Kim, Y., Szczepaniak, K., Chen, C. C.-H., Eckmann, C. R., Myong, S., and Brangwynne, C. P. (2015) The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *PNAS*. **112**, 7189–7194
71. Harami, G. M., Kovács, Z. J., Pancsa, R., Pálinkás, J., Baráth, V., Tárnok, K., Málnási-Csizmadia, A., and Kovács, M. (2020) Phase separation by ssDNA binding protein controlled via protein−protein and protein−DNA interactions. *PNAS*. **117**, 26206–26217
72. Lin, Y., Protter, D. S. W., Rosen, M. K., and Parker, R. (2015) Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. *Mol Cell*. **60**, 208–219
73. Fox, N. K., Brenner, S. E., and Chandonia, J.-M. (2014) SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res*. **42**, D304-309
74. Chandonia, J.-M., Fox, N. K., and Brenner, S. E. (2019) SCOPe: classification of large macromolecular structures in the structural classification of proteins-extended database. *Nucleic Acids Res*. **47**, D475–D481
75. Li, P., Banjade, S., Cheng, H.-C., Kim, S., Chen, B., Guo, L., Llaguno, M., Hollingsworth, J. V., King, D. S., Banani, S. F., Russo, P. S., Jiang, Q.-X., Nixon, B. T., and Rosen, M. K. (2012) Phase transitions in the assembly of multivalent signalling proteins. *Nature*. **483**, 336–340
76. Chen, Z., Hou, C., Wang, L., Yu, C., Chen, T., Shen, B., Hou, Y., Li, P., and Li, T. (2022) Screening membraneless organelle participants with machine-learning models that integrate multimodal features. *Proc Natl Acad Sci U S A*. **119**, e2115369119
77. Saar, K. L., Morgunov, A. S., Qi, R., Arter, W. E., Krainer, G., Lee, A. A., and Knowles, T. P. J. (2021) Learning the molecular grammar of protein condensates from sequence determinants and embeddings. *Proc Natl Acad Sci U S A*. **118**, e2019053118
78. Farahi, N., Lazar, T., Wodak, S. J., Tompa, P., and Pancsa, R. (2021) Integration of Data from Liquid-Liquid Phase Separation Databases Highlights Concentration and Dosage Sensitivity of LLPS Drivers. *Int J Mol Sci*. **22**, 3017
79. Dignon, G. L., Zheng, W., Kim, Y. C., and Mittal, J. (2019) Temperature-Controlled Liquid–Liquid Phase Separation of Disordered Proteins. *ACS Cent. Sci.* **5**, 821–830
80. McPherson, A. (2017) Protein Crystallization. *Methods Mol Biol*. **1607**, 17–50
81. Berjanskii, M. V., and Wishart, D. S. (2005) A Simple Method To Predict Protein Flexibility Using Secondary Chemical Shifts. *J. Am. Chem. Soc.* **127**, 14970–14971
82. Wang, G., and Dunbrack, R. L. (2005) PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* **33**, W94-98
83. Levitt, M. (1978) Conformational preferences of amino acids in globular proteins. *Biochemistry*. **17**, 4277–4285
84. Flory, P. J. (1949) The Configuration of Real Polymer Chains. *J. Chem. Phys.* **17**, 303–310
85. Flory, P. J. (1969) *Statistical mechanics of chain molecules*, Interscience Publishers, New York

86. Perez, R. B., Tischer, A., Auton, M., and Whitten, S. T. (2014) Alanine and proline content modulate global sensitivity to discrete perturbations in disordered proteins. *Proteins*. **82**, 3373–3384
87. English, L. R., Tilton, E. C., Ricard, B. J., and Whitten, S. T. (2017) Intrinsic α helix propensities compact hydrodynamic radii in intrinsically disordered proteins. *Proteins*. **85**, 296–311
88. Langridge, T. D., Tarver, M. J., and Whitten, S. T. (2014) Temperature effects on the hydrodynamic radius of the intrinsically disordered N-terminal region of the p53 protein. *Proteins*. **82**, 668–678
89. Elam, W. A., Schrank, T. P., Campagnolo, A. J., and Hilser, V. J. (2013) Evolutionary conservation of the polyproline II conformation surrounding intrinsically disordered phosphorylation sites. *Protein Science*. **22**, 405–417
90. R Core Team (2022) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria
91. Alberti, S., Halfmann, R., King, O., Kapila, A., and Lindquist, S. (2009) A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. *Cell*. **137**, 146–158
92. Sawle, L., and Ghosh, K. (2015) A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J Chem Phys*. **143**, 085101

Table

Table 1. Summary of mean v_{model} in protein sequence sets.

Set	Number	v_{model} [a]	*t*-test [b]	*U*-test [b]
PS ID	224	0.542 ± 0.020	-	-
ID	121	0.558 ± 0.022	$2.5e^{-10}$	$1.6e^{-11}$
Folded	421	0.537 ± 0.008	$1.2e^{-3}$	$1.5e^{-3}$

[a] Mean ± standard deviation.
[b] One-tail *p*-value, where *p*-value <0.05 indicates the compared sets are statistically different in their means. Comparisons are to the PS ID sequence set.

Table 2: Summary of mean β-turn propensity in the protein sequence sets.

Set	Number	β-turn propensity [a]	*t*-test [b]	*U*-test [b]
PS ID	224	1.152 ± 0.087	-	-
ID	121	1.101 ± 0.075	$4.6e^{-8}$	$4.9e^{-9}$
Folded	421	0.971 ± 0.040	$2.0e^{-33}$	$1.1e^{-89}$

[a] Mean ± standard deviation.
[b] One-tail *p*-value, where *p*-value <0.05 indicates the compared sets are statistically different in their means. Comparisons are to the PS ID sequence set.

Figures

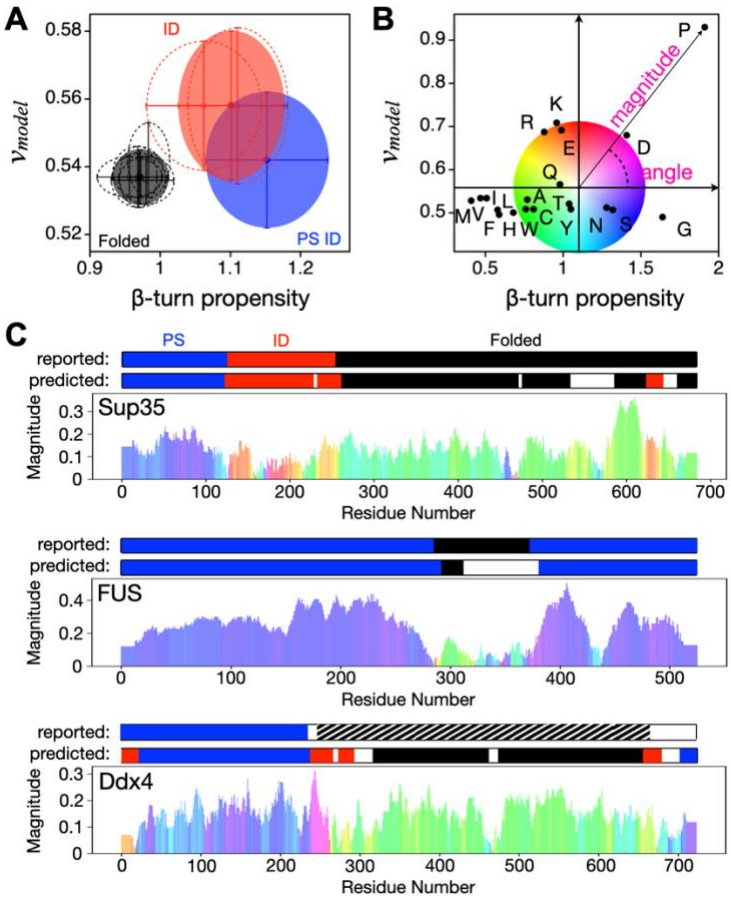

Figure 1. Sequence-calculated ν_{model} and β-turn propensity separate protein regions by class. *A,* comparing ν_{model} and β-turn propensity in each sequence set. Filled circles show the mean and standard deviation in ν_{model} and β-turn propensity in the PS ID (blue), ID (red), and folded (black) sets. Open and dashed circles show the mean and standard deviation in individual subsets: previous ID and BMRB & DisProt (red); previous folded, human, small-to-large, extremophile, membrane, and metamorphic (black). *B,* comparing ν_{model} and β-turn propensity in homopolymers (N = 100) where amino acid type is identified by its one-letter code. A centralized origin was mapped into this plot at the β-turn propensity and ν_{model} values of 1.101 and 0.558, respectively, which are the means in the ID set. From this origin, every amino acid type can be represented by a distance magnitude and angular displacement; as shown for proline. A color wheel is used to convey angular displacement. *C*, magnitude/color plots are compared to the ParSe (original version) predictions for Sup35 (UniProt ID P05453), FUS (UniProt ID P35637), and Ddx4 (UniProt ID Q9NQI0), and to regions reported (i.e., identified) by experiment. Each figure shows the magnitude (y-axis) and color (angular displacement) by residue number (x-axis), as determined by amino acid type and its magnitude/color from panel B. ParSe predictions use blue (PS), red (ID), and black (folded). Striped represents ≥50% identity to a known folded protein.

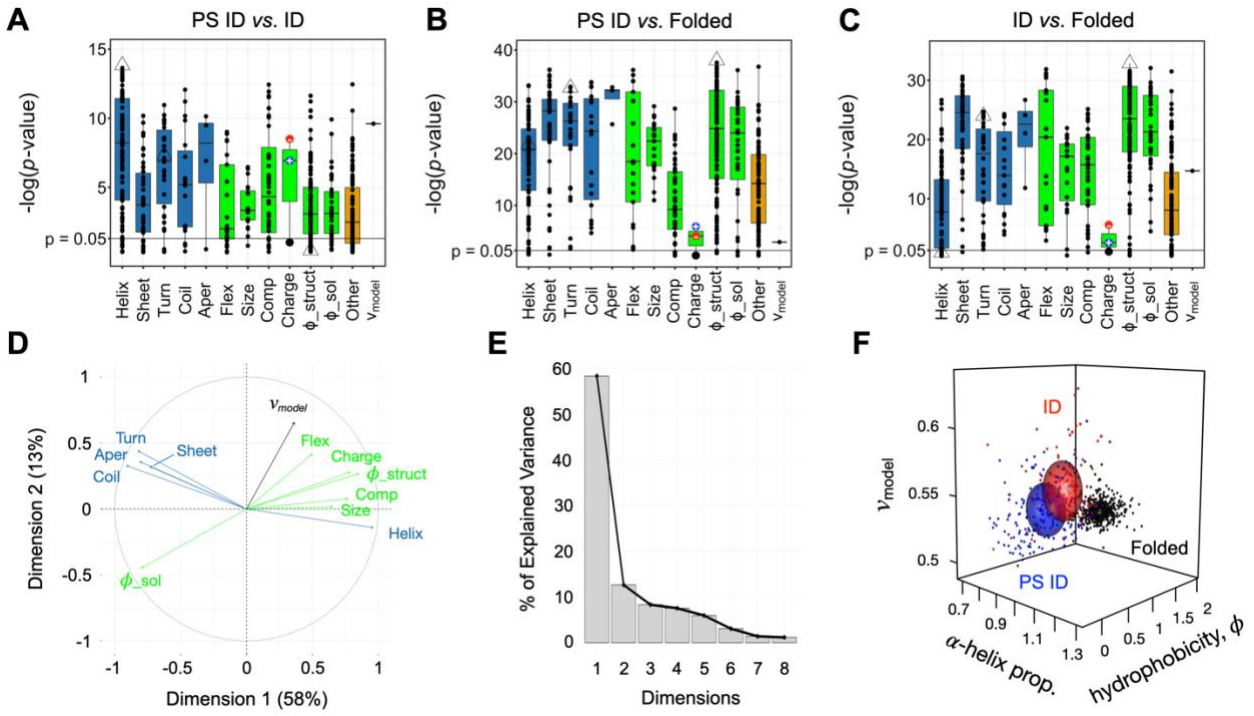

Figure 2. Robust differences in intrinsic sequence-calculated properties are found when comparing means by protein region class. *A-C*, *p*-values calculated by Welch's unequal variances *t*-test, shown as -log(*p*-value), compares set means in 567 amino acid scales and ν_{model}. Conformation-based scales, highlighted by blue boxplots, are grouped by type according to α-helix (Helix), sheet or strand (Sheet), β-turn, tight turn, or reverse turn (Turn), coil or loop (Coil), and aperiodic (Aper) propensities. Physicochemical-based scales, highlighted by green box plots, are grouped by type according to flexibility (Flex), size (Size), composition (Comp), negative charge, positive charge, or net charge (Charge), and hydrophobicity (ϕ). Hydrophobicity scales were separated into two types: structure-based (ϕ_struct), where the scale is derived from a structural metric like burial or contact frequency in surveys of high-resolution protein structures, and solution-based (ϕ_sol), where the scale is obtained from solution studies like measuring the transfer free energy of the amino acids from water to an organic solvent. Scales (e.g., refractivity, crystal melting point) that did not easily map into a conformation- or physicochemical-based group were combined separately (Other). Boxplots show the dataset median (50th percentile) with the central bar, and the vertical width spans the 25th to 75th percentiles. Open triangles highlight the smallest *p*-value when comparing means in the PS ID and ID sets (from an α-helix propensity scale), the smallest *p*-value when comparing means in either the PS ID or ID sets with the folded set (from a structure-based hydrophobicity scale), and the β-turn propensity scale used in ParSe. *D*, bidimensional plot from PCA showing the modes of variance in the combined ID set (PS ID and ID) arising from conformation- (blue arrows) and physicochemical-based (green arrows) scales relative to the two principal components of variance, given as Dimension 1 and Dimension 2. *E*, scree plot showing the percent of the total variance in the combined set of ID sequences that is captured by each principal component (i.e., dimension). *F*, sequence calculated ν_{model}, α-helix propensity, and hydrophobicity for the sequences in the PS ID (blue), ID (red), and folded (black) sets; spheres show the set mean ± σ.

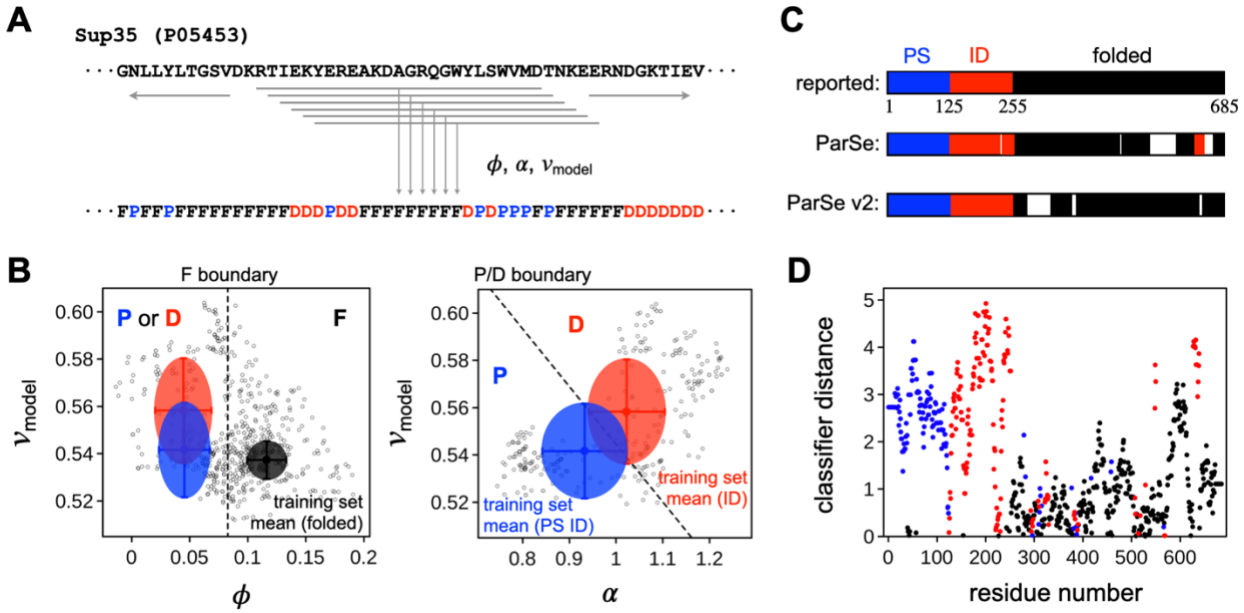

Figure 3. Predicting protein regions from sequence using the ParSe v2 algorithm. *A*, a sliding window algorithm is used to identify from sequence regions within a protein that match the PS ID, ID, and folded classes. Hydrophobicity (ϕ), α-helix propensity (α), and ν_{model} are calculated for each contiguous stretch of 25-residues, or "window", in the primary sequence. *B*, each window is assigned a label, F, P, or D, depending on the values of ϕ, α, and ν_{model}. In the left figure, open circles are ϕ and ν_{model} calculated for each 25-residue window in the Sup35 sequence (UniProt ID P05453); filled circles are the mean ± σ in ϕ and ν_{model} in the ID (red), PS ID (blue), and folded (black) sequence sets. Windows with $\phi \geq$ the folded set mean - 2σ (dashed line) are labeled F. For windows with $\phi <$ the folded set mean - 2σ, the label is determined by α and ν_{model}; P for low α with low ν_{model}, or D for high α with high ν_{model}, as shown in the right figure. Filled circles show the mean ± σ in α and ν_{model} in the ID (red) and PS ID (blue) sets. *C*, contiguous regions ($N \geq 20$) in the Sup35 primary sequence that were 90% of only one label P, D, or F are colored blue, red, or black, respectively, to represent predicted PS, ID, or folded regions. Predictions from the original ParSe and ParSe v2 are compared to the reported regions identified by experiment. *D*, classifier distance of each window, assigned to the central residue of the window and then colored according to its label P (blue), D (red), or F (black).

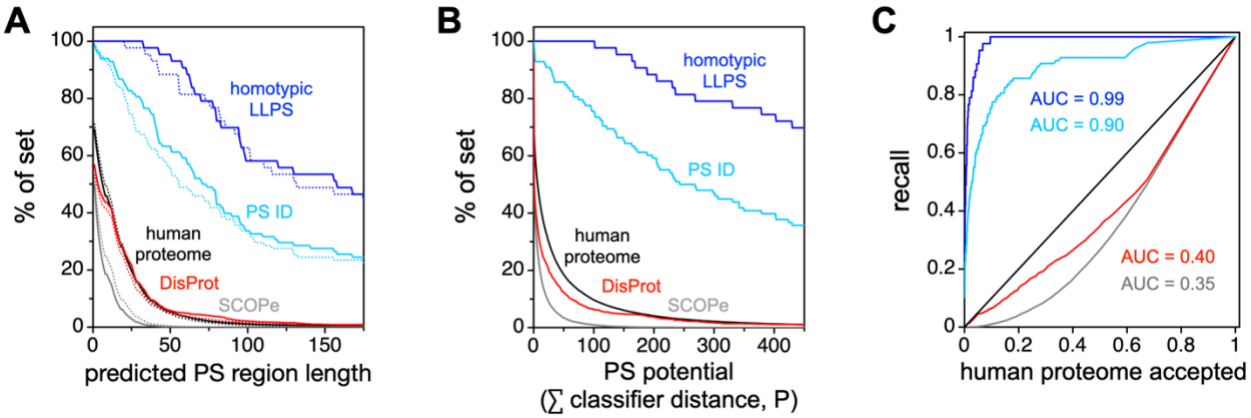

Figure 4. ParSe predicted PS regions are rarely found in the human proteome. *A*, ParSe (stippled lines) and ParSe v2 (solid lines) were used to identify regions in proteins that were ≥90% labeled P, which are referred to as phase-separating, PS, regions. Shown by the y-axis is the percent of proteins in a set with PS regions at least as long as the length indicated by the x-axis. The human proteome (UniProt reference proteome UP000005640) is given by black lines; DisProt (minus LLPS annotated entries) by red lines; SCOPe (version 2.07) by grey lines; a set of *in vitro* sufficient homotypic LLPS proteins by blue lines; and the full sequences of the proteins in the PS ID set by light blue lines. *B*, the summed P classifier distance was calculated by ParSe v2 for the protein sets in panel A. Shown by the y-axis is the percent of proteins in a set with a summed P classifier distance at least as much as the value indicated by the x-axis. Lines were colored using the same coloring scheme as in panel A. *C*, reproduction of the results in panel B wherein each set was directly compared to the human proteome result. Here, lines show the % of a set (using the same coloring scheme) plotted against the human proteome % of set for values of the summed P classifier distance.

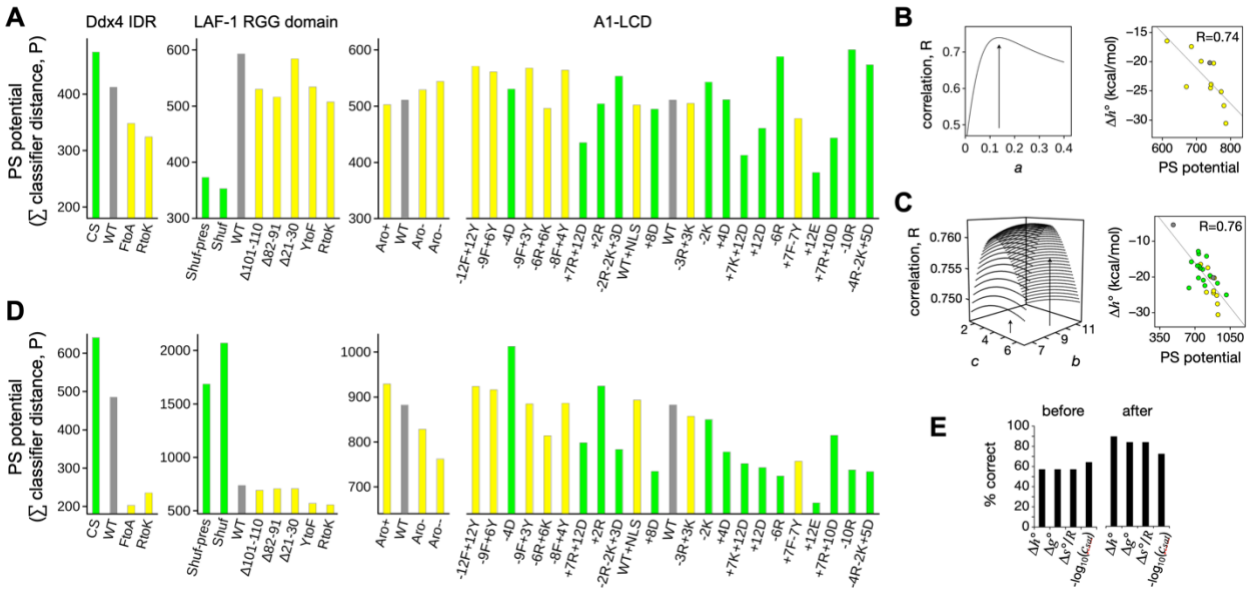

Figure 5. Predicting mutation effects on phase separation behavior. *A*, the summed classifier distance of P-labeled windows was used to calculate a phase-separating (PS) potential from sequence. Mutants were grouped by experimental study and colored grey for wildtype (WT), yellow for mutants with both *NCPR* and *SCD* identical to the WT values, and green otherwise (non-WT *NCPR* and *SCD*). Placement left-to-right within a study follows the reported PS potential in rank, from high-to-low, for comparison to the predicted PS potential. A1-LCD mutants used $\Delta h°$ and not c_{sat} to establish rank. *B*, A1-LCD mutants with *NCPR* and *SCD* matching the WT values were used to fix a in Equation 3 by optimizing the correlation of Parse-calculated PS potential (including U_π) to $\Delta h°$; the right figure shows the optimal correlation. *C*, similarly, all A1-LCD and Ddx4 mutants with experimental $\Delta h°$ were then used to fix b and c in Equation 4 by optimizing the correlation of ParSe-calculated PS potential (including U_π and U_q) to $\Delta h°$; the right figure shows the optimal correlation. *D*, ParSe-calculated PS potentials (including U_π and U_q optimized to $\Delta h°$) for the mutant and WT sequences. *E*, percent of mutants correctly predicting an increase or decrease in PS potential relative to the WT before and after including U_π and U_q in the calculations. Results are binned according to experimental value that was used to fix a, b, and c in U_π and U_q.

Supporting Information for:

Intrinsically disordered regions that drive phase separation form a robustly distinct protein class

Short title: Intrinsic properties of phase-separating IDRs

Ayyam Y. Ibrahim, Nathan P. Khaodeuanepheng, Dhanush L. Amarasekara, John J. Correia, Karen A. Lewis, Nicholas C. Fitzkee, Loren E. Hough, and Steven T. Whitten

Contents:

Supporting Tables

Table S1. List of folded protein regions.

Name	Database [a]	UniProt accession number	folded regions (*N*) [b]	PDB entries
PPP5C	Wang *et al*	P53041	19-177 (159)	1a17.pdb
Galectin-3	Wang *et al*	P17931**	114-250 (137)	1a3k.pdb
RB1	Wang *et al*	P06400	378-562 (185)	1ad6.pdb
CD40LG	Wang *et al*	P29965	116-261 (146)	1aly.pdb
FABP5	Wang *et al*	Q01469	3-135 (133)	1b56.pdb
LALBA	Wang *et al*	P00709	20-142 (123)	1b9o.pdb
CDKN2D	Wang *et al*	P55273	7-162 (156)	1bd8.pdb
AMBP	Wang *et al*	P02760	230-339 (110)	1bik.pdb
FKBP1A	Wang *et al*	P62942	2-108 (107)	1bkf.pdb
SPTBN1	Wang *et al*	Q01082	173-280 (108)	1bkr.pdb
TIMP2	Wang *et al*	P16035	27-208 (182)	1br9.pdb
ZBTB16	Wang *et al*	Q05516	6-126 (121)	1buo.pdb
LGALS3BP	Wang *et al*	Q08380	19-127 (111)	1by2.pdb
HSO90AA1	Wang *et al*	P07900	11-223 (213)	1byq.pdb
CRABP2	Wang *et al*	P29373	2-138 (137)	1cbs.pdb
CD4	Wang *et al*	P01730	26-203 (178)	1cdy.pdb
CALM1	Wang *et al*	P0DP23	5-148 (144)	1cll.pdb
HRAS	Wang *et al*	P01112	1-166 (166)	1ctq.pdb
APAF1	Wang *et al*	O14727	1-93 (93)	1cy5.pdb
F5	Wang *et al*	P12259	2066-2224 (159)	1czt.pdb
F8	Wang *et al*	P00451	2190-2348 (159)	1d7p.pdb
ASGR1	Wang *et al*	P07306	154-281 (128)	1dv8.pdb
RNASE1	Wang *et al*	P07998	35-154 (120)	1e21.pdb
CD69	Wang *et al*	Q07108	83-199 (117)	1e87.pdb
PLEKHA1	Wang *et al*	Q9HB21	190-293 (104)	1eaz.pdb

CCL8	Wang *et al*	P80075	25-99 (75)	1esr.pdb
DAPP1	Wang *et al*	Q9UN19	162-261 (100)	1fao.pdb
PFN1	Wang *et al*	P07737	2-140 (139)	1fil.pdb
AIMP1	Wang *et al*	Q12904	150-313 (164)	1fl0.pdb
FN1	Wang *et al*	P02751	1543-1633 (91)	1fna.pdb
FCGR3B	Wang *et al*	O75015	21-193 (173)	1fnl.pdb
IGHE	Wang *et al*	P01854	217-424 (208)	1fp5.pdb
GSTZ1	Wang *et al*	O43708	5-212 (208)	1fw1.pdb
SELE	Wang *et al*	P16581	22-178 (157)	1g1t.pdb
CST3	Wang *et al*	P01034*	36-146 (111)	1g96.pdb
MMP2	Wang *et al*	P08253	461-660 (200)	1gen.pdb
CALML3	Wang *et al*	P27482	5-148 (144)	1ggz.pdb
TXNL1	Wang *et al*	O43396	2-108 (107)	1gh2.pdb
CTSS	Wang *et al*	P25774	115-331 (217)	1glo.pdb
GABARAP	Wang *et al*	O95166*	1-117 (117)	1gnu.pdb
IGF2R	Wang *et al*	P11717	1515-1647 (133)	1gp0.pdb
RNASE2	Wang *et al*	P10153	28-161 (134)	1gqv.pdb
COL10A1	Wang *et al*	Q03692	549-680 (132)	1gr3.pdb
MADCAM1	Wang *et al*	Q13477**	23-227 (206)	1gsm.pdb
NCF4	Wang *et al*	Q15080	2-144 (143)	1h6h.pdb
BLVRB	Wang *et al*	P30043	1-205 (205)	1hdo.pdb
QDPR	Wang *et al*	P09417	9-244 (236)	1hdr.pdb
FABP3	Wang *et al*	P05413	2-132 (131)	1hmt.pdb
GSTM2	Wang *et al*	P28161	2-218 (217)	1hna.pdb
MBL2	Wang *et al*	P11226	108-248 (141)	1hup.pdb
IL4	Wang *et al*	P05112	25-153 (129)	1hzi.pdb
PCMT1	Wang *et al*	P22061	3-226 (224)	1i1n.pdb
GTF2F1	Wang *et al*	P35269	449-517 (73)	1i27.pdb
UBR5	Wang *et al*	O95071	2393-2453 (61)	1i2t.pdb

PRNP	Wang *et al*	P04156**	119-226 (108)	1i4m.pdb
LPA	Wang *et al*	P08519	1274-1355 (82)	1i71.pdb
MMP8	Wang *et al*	P22894	100-262 (163)	1i76.pdb
ICAM1	Wang *et al*	P05362	28-212 (185)	1iam.pdb
ARHGEF1	Wang *et al*	Q92888*	44-233 (190)	1iap.pdb
LMNA	Wang *et al*	P02545	436-544 (113)	1ifr.pdb
FGF9	Wang *et al*	P31371	52-208 (157)	1ihk.pdb
LCK	Wang *et al*	P06239	123-226 (104)	1ijr.pdb
FGF4	Wang *et al*	P08620	79-206 (128)	1ijt.pdb
HSD17B4	Wang *et al*	P51659	622-736 (115)	1ikt.pdb
ABHD14B	Wang *et al*	Q96IU4	2-209 (208)	1imj.pdb
UBE2V2	Wang *et al*	Q15819	7-145 (139)	1j74.pdb
ANAPC10	Wang *et al*	Q9UM13	2-162 (161)	1jhj.pdb
MMP12	Wang *et al*	P39900	106-263 (158)	1jk3.pdb
LYZ	Wang *et al*	P61626	19-148 (130)	1jsf.pdb
TCL1A	Wang *et al*	P56279	4-114 (111)	1jsg.pdb
GGA1	Wang *et al*	Q9UJY5	7-145 (139)	1jwf.pdb
MATK	Wang *et al*	P42679	117-213 (97)	1jwo.pdb
PTK2	Wang *et al*	Q05397	908-1049 (142)	1k04.pdb
BCL3	Wang *et al*	P20749	133-360 (228)	1k1b.pdb
ANG	Wang *et al*	P03950	26-147 (122)	1k59.pdb
RAP2A	Wang *et al*	P10114	1-167 (167)	1kao.pdb
GSN	Wang *et al*	P06396	185-288 (104)	1kcq.pdb
NRP1	Wang *et al*	O14786	273-427 (155)	1kex.pdb
DHFR	Wang *et al*	P00374	2-187 (186)	1kmv.pdb
HINT1	Wang *et al*	P49773	16-126 (111)	1kpf.pdb
COL6A3	Wang *et al*	P12111	3108-3165 (58)	1kth.pdb
PROCR	Wang *et al*	Q9UNN8	25-194 (170)	1l8j.pdb
GNLY	Wang *et al*	P22749	63-136 (74)	1l9l.pdb

CLC	Wang *et al*	Q05315	2-142 (141)	1lcl.pdb
B2M	Wang *et al*	P61769**	21-116 (96)	1lds.pdb
PCTP	Wang *et al*	Q9UKL6	8-210 (203)	1ln1.pdb
RBP7	Wang *et al*	Q96R05	2-134 (133)	1lpj.pdb
THBS1	Wang *et al*	P07996	434-546 (113)	1lsl.pdb
RND3	Wang *et al*	P61587	22-200 (179)	1m7b.pdb
TGFBR2	Wang *et al*	P37173	49-153 (105)	1m9z.pdb
SOD1	Wang *et al*	P00441*	2-154 (153)	1mfm.pdb
RAC1	Wang *et al*	P63000	2-181 (183)	1mh1.pdb
NT5M	Wang *et al*	Q9NPB1	34-227 (194)	1mh9.pdb
SUOX	Wang *et al*	P51687	81-160 (80)	1mj4.pdb
APP	Wang *et al*	P05067**	28-123 (96)	1mwp.pdb
RAB5A	Wang *et al*	P20339	15-181 (167)	1n6h.pdb
KIR2DL1	Wang *et al*	P43626	27-221 (195)	1nkr.pdb
FKBP3	Wang *et al*	Q00688	109-224 (116)	1pbk.pdb
CYTH2	Wang *et al*	Q99418	52-246 (195)	1pbv.pdb
PIK3R1	Wang *et al*	P27986	3-85 (83)	1pht.pdb
PLA2GRA	Wang *et al*	P14555	21-144 (124)	1pod.pdb
CDC25B	Wang *et al*	P30305	388-565 (178)	1qb0.pdb
REG1A	Wang *et al*	P05451	23-166 (144)	1qdd.pdb
ESR1	Wang *et al*	P03372**	304-551 (248)	1qkt.pdb
ACTN2	Wang *et al*	P35609	391-635 (248)	1quu.pdb
RBP4	Wang *et al*	P02753	19-193 (175)	1rbp.pdb
PLA2G4A	Wang *et al*	P47712	17-141 (126)	1rlw.pdb
SPARC	Wang *et al*	P09486	153-303 (151)	1sra.pdb
TNC	Wang *et al*	P24821	802-891 (90)	1ten.pdb
CLEC3B	Wang *et al*	P05452	66-202 (137)	1tn3.pdb
ITGAL	Wang *et al*	P20701	153-333 (181)	1zon.pdb
ICAM2	Wang *et al*	P13598	25-216 (192)	1zxq.pdb

ABL1	Wang *et al*	P00519	57-218 (163)	2abl.pdb
PPIA	Wang *et al*	P62937	2-165 (164)	2cpl.pdb
FCGR2B	Wang *et al*	P31994	46-218 (173)	2fcb.pdb
FTH1	Wang *et al*	P02794	6-177 (172)	2fha.pdb
IL10	Wang *et al*	P22301	24-178 (155)	2ilk.pdb
S100A7	Wang *et al*	P31151	2-97 (96)	2psr.pdb
TGFB2	Wang *et al*	P61812	303-414 (112)	2tgi.pdb
FGG	Wang *et al*	P02679	170-418 (249)	3fib.pdb
CXCL8	Wang *et al*	P10145	32-99 (68)	3il8.pdb
ACP1	Wang *et al*	P24666	2-158 (157)	5pnt.pdb
VIL1	Fitzkee & Rose	P02640	792-826 (36)	1vii.pdb
Prkcd	Fitzkee & Rose	P28867	231-280 (50)	1ptq.pdb
spg	Fitzkee & Rose	P06654	228-282 (56)	2gb1.pdb
FYN	Fitzkee & Rose	P06241	84-142 (59)	1shfA.pdb
cspB	Fitzkee & Rose	P32081	1-67 (67)	1csp.pdb
UBC	Fitzkee & Rose	P0CG48	609-684 (76)	1ubq.pdb
cI	Fitzkee & Rose	P03034	7-93 (87)	1lmb.pdb
Barstar	Fitzkee & Rose	P11540	2-90 (89)	1a19A.pdb
ACYP1	Fitzkee & Rose	P41500	4-101 (98)	2acy.pdb
PETE	Fitzkee & Rose	P00299	70-168 (99)	2pcy.pdb
CYCS	Fitzkee & Rose	P00004	2-105 (104)	1hrc.pdb
Pik3r1	Fitzkee & Rose	Q63787	321-431 (111)	1fu6A.pdb
Hemerythrin	Fitzkee & Rose	P02246	1-113 (113)	2hmqA.pdb
LALBA	Fitzkee & Rose	P00711	20-141 (122)	1f6sA.pdb
RNASE1	Fitzkee & Rose	P61823	27-150 (124)	1xptA.pdb
cheY	Fitzkee & Rose	P0AE67	2-129 (128)	1ehc.pdb
LYZ	Fitzkee & Rose	P00698	19-147 (129)	1hel.pdb
Fabp2	Fitzkee & Rose	P02693	2-132 (131)	1ifb.pdb
nuc	Fitzkee & Rose	P00644	83-223 (141)	2sns.pdb

CALM	Fitzkee & Rose	P62157	5-147 (143)	1cm1A.pdb
MB	Fitzkee & Rose	P02185	2-154 (153)	1mbo.pdb
rnhA	Fitzkee & Rose	P0A7Y4	1-155 (155)	2rn2.pdb
gag-pol	Fitzkee & Rose	O92956	1331-1487 (162)	1asu.pdb
E (endolysin)	Fitzkee & Rose	P00720	1-164 (164)	2lzm.pdb
DFR1	Fitzkee & Rose	P22906	1-192 (192)	1ai9A.pdb
mutY	Fitzkee & Rose	P17802	1-225 (225)	1mun.pdb
Triosephosphate isomerase	Fitzkee & Rose	P04789	2-250 (249)	5timA.pdb
HAGH	Fitzkee & Rose	Q16775	49-308 (260)	1qh3A.pdb
ecoRIR	Fitzkee & Rose	P00642	17-277 (261)	1eriA.pdb
galE	Fitzkee & Rose	P09147	1-338 (338)	1nah.pdb
CKMT1A	Fitzkee & Rose	P12532	39-417 (379)	1qk1A.pdb
PGK1	Fitzkee & Rose	P00560	2-415 (415)	3pgk.pdb
apr	Panja *et al*	P00782	108-382 (274)	1a2q.pdb
adk	Panja *et al*	P69441	1-214 (214)	1ake.pdb
amy	Panja *et al*	P29957	25-472 (448)	1aqm.pdb
hip	Panja *et al*	P00260	38-122 (85)	1b0y.pdb
amyE	Panja *et al*	P00691	42-466 (425)	1bag.pdb
FGF2	Panja *et al*	P09038	161-285 (125)	1bas.pdb
amyS	Panja *et al*	P06278	32-512 (481)	1bli.pdb
axe-2	Panja *et al*	O59893	28-234 (207)	1bs9.pdb
sodB	Panja *et al*	Q9X6W9	3-213 (211)	1coj.pdb
fer1	Panja *et al*	P00217	2-129 (128)	1doi.pdb
phnA	Panja *et al*	Q51782	2-407 (404)	1ei6.pdb
cyp119	Panja *et al*	Q55080	1-367 (367)	1f4t.pdb
atsA	Panja *et al*	P51691	3-527 (524)	1hdh.pdb
hip2	Panja *et al*	P38524	1-71 (71)	1hpi.pdb
katG2	Panja *et al*	O59651	18-731 (707)	1itk.pdb
aspC	Panja *et al*	Q8RR70	1-388 (388)	1j32.pdb

SSO2706	Panja *et al*	P50389	3-236 (226)	1jds.pdb
mtnN	Panja *et al*	P0AF12	1-230 (226)	1jys.pdb
rpiA	Panja *et al*	O50083	1-229 (229)	1lk5.pdb
VNG_1446H	Panja *et al*	Q9HPW4	11-77 (67)	1mog.pdb
speE	Panja *et al*	Q5SK28	1-312 (309)	1uir.pdb
Endoglucanase	Panja *et al*	P06564	578-761 (181)	1uww.pdb
acyP	Panja *et al*	P84142	2-91 (90)	1v3z.pdb
mdh	Panja *et al*	O59028	2-360 (337)	1v9n.pdb
serC	Panja *et al*	Q9RME2	2-361 (360)	1w23.pdb
amyA	Panja *et al*	Q8GPL8	28-515 (488)	1wza.pdb
APE_2278	Panja *et al*	Q9Y9L0	2-245 (240)	1x0r.pdb
mvaS	Panja *et al*	Q9FD71	1-383 (383)	1x9e.pdb
Rv1264	Panja *et al*	P9WMU9	14-376 (360)	1y10.pdb
adk	Panja *et al*	P27142	1-217 (217)	1zin.pdb
Rv1885c	Panja *et al*	P9WIB9	35-199 (165)	2ao2.pdb
ndk	Panja *et al*	P61136	4-158 (155)	2az1.pdb
gdh	Panja *et al*	Q977U7	1-357 (355)	2b5v.pdb
tdh	Panja *et al*	O58389	3-347 (327)	2d8a.pdb
Lysozyme 1	Panja *et al*	Q7YT16	20-141 (122)	2fbd.pdb
Cat-1	Panja *et al*	Q24940	17-326 (306)	2o6x.pdb
oxc	Panja *et al*	P0AFI0	5-551 (547)	2q27.pdb
Thioredoxin-dependent peroxiredoxin	Panja *et al*	G1K3P1	1-76 (156)	2xhf.pdb
sod	Panja *et al*	Q9Y8H8	1-212 (212)	3ak1.pdb
Alkaline serine protease ver112	Panja *et al*	Q68GV9	104-382 (279)	3f7m.pdb
dapE	Panja *et al*	P44514	1-376 (370)	3ic1.pdb
pepQ	Panja *et al*	Q44238	1-440 (425)	3l24.pdb
sodB	Panja *et al*	P84612	1-192 (192)	3lio.pdb
Enpp2	Panja *et al*	Q9R1E6	51-855 (805)	3nkm.pdb

phoK	Panja *et al*	A1YYW7	31-556 (526)	3q3q.pdb
cheC1	Panja *et al*	Q5V4K4	2-206 (200)	3qta.pdb
FOXG_17421	Panja *et al*	B3A0S5	1-327 (327)	3u7b.pdb
Alkaline phosphatase	Panja *et al*	B5BP20	31-527 (497)	3wbh.pdb
LGMN	Panja *et al*	Q99538	26-288 (267)	4aw9.pdb
LMRG_02624	Panja *et al*	A0A0H3GD84	39-526 (488)	4cdb.pdb
bop	Panja *et al*	Q5UXY6	3-238 (236)	4pxk.pdb
mdh	Panja *et al*	A9W386	2-320 (319)	4ror.pdb
patA	Panja *et al*	P42588	7-459 (453)	4uox.pdb
cysQ	Panja *et al*	P9WKJ1	10-267 (266)	5djf.pdb
F	Chen *et al*	P11209	480-515 (36)	1g2cF.pdb
HA	Chen *et al*	P03437	385-498 (114)	1htmB.pdb
SERPINB14	Chen *et al*	P01012	2-386 (381)	1jtiB.pdb
Plk4	Chen *et al*	Q64702	845-919 (75)	1mbyA.pdb
PVC01_130047600	Chen *et al*	O60989	76-450 (375)	1miqB.pdb
MATALPHA2	Chen *et al*	P0CY08	113-189 (77)	1mnmC.pdb
colG	Chen *et al*	Q9X721	1005-1118 (111)	1nqdA.pdb
SRP102	Chen *et al*	P36057	36-244 (191)	1nrjB.pdb
PDE5A	Chen *et al*	O76074	535-860 (311)	1rkpA.pdb
cobB	Chen *et al*	P75960	40-274 (225)	1s5pA.pdb
F	Chen *et al*	P04849	122-183 (62)	1svfC.pdb
SOD1	Chen *et al*	P00441**	2-154 (153)	1uxmK.pdb
F	Chen *et al*	O89342	143-205 (63)	1wp8C.pdb
S	Chen *et al*	P59594	892-981 (124)	1wyyB.pdb
tll0464	Chen *et al*	Q8DLM0	1-112 (102)	1x0gA.pdb
hlyA	Chen *et al*	P09545	46-741 (663)	1xezA.pdb
SAR-endolysin	Chen *et al*	Q37875	9-185 (170)	1xjtA.pdb
Relb	Chen *et al*	Q04863	276-378 (110)	1zk9A.pdb
ftsH	Chen *et al*	Q9WZ49	150-606 (421)	2ce7C.pdb

suhB	Chen *et al*	O33832	1-254 (254)	2p3vA.pdb
Polyprotein	Chen *et al*	O36607	3-230 (227)	2pbk.pdb
NRP2	Chen *et al*	O60462	276-595 (315)	2qqjA.pdb
MAD2L1	Chen *et al*	Q13257	1-205 (202)	2vfxL.pdb
prgI	Chen *et al*	P41784	19-80 (62)	2x9cA.pdb
FN1	Chen *et al*	P02751	516-606 (91)	3ejhA.pdb
CST3	Chen *et al*	P01034**	38-146 (107)	3gaxA.pdb
R	Chen *et al*	P27359	1-165 (165)	3hdeA.pdb
FBP2	Chen *et al*	O00757	9-337 (326)	3ifaA.pdb
PRIM2	Chen *et al*	P49643	272-457 (167)	3l9qB.pdb
B2M	Chen *et al*	P61769*	21-119 (99)	3lowA.pdb
gag-pol	Chen *et al*	P04585	588-1139 (552)	3meeA.pdb
gp-C	Chen *et al*	Q9ICW1	313-422 (103)	3mkoA.pdb
rsmH	Chen *et al*	P60390	8-313 (283)	3tkaA.pdb
CWC2	Chen *et al*	Q12046	3-227 (225)	3tp2A.pdb
PR	Chen *et al*	Q3L181	1-336 (311)	3uyiA.pdb
macA	Chen *et al*	Q74FY6**	23-346 (320)	4aalA.pdb
Diphtheria toxin	Chen *et al*	P00588	37-567 (499)	4ae0A.pdb
SUN2	Chen *et al*	Q9UH99	522-717 (196)	4dxrA.pdb
bcp	Chen *et al*	Q9YA14	2-160 (160)	4gqcB.pdb
PRNP	Chen *et al*	Q95211	125-221 (97)	4hlsA.pdb
Grem2	Chen *et al*	O88273	50-160 (111)	4jphB.pdb
pimA	Chen *et al*	A0QWG6	1-373 (359)	4n9wA.pdb
plyB	Chen *et al*	Q5W9E8	53-519 (465)	4ov8A.pdb
KWL1	Chen *et al*	P85261	48-213 (158)	4pmkA.pdb
COMT	Chen *et al*	P21964	54-266 (207)	4pyiA.pdb
MJ1213	Chen *et al*	Q58610	1-109 (109)	4qhfA.pdb
gbs1529	Chen *et al*	Q8E473	494-642 (141)	4rmbA.pdb
OAS1	Chen *et al*	Q29599	1-349 (349)	4rwnA.pdb

ply	Chen *et al*	Q7ZAK5	1-471 (471)	5aoeB.pdb
malE	Chen *et al*	P0AEX9	27-393 (402)	5b3zA.pdb
TRAP1	Chen *et al*	Q12931	82-294 (205)	5f3kA.pdb
G	Chen *et al*	P0C2X0	1-409 (409)	5i2mA.pdb
DVL2	Chen *et al*	O14641	416-509 (92)	5suzA.pdb
MADCAM1	membrane protein	Q13477*	23-231 (209)	1bqsA.pdb
MSN	membrane protein	P26038	4-297 (289)	1ef1A.pdb
FCGR2A	membrane protein	P12318	37-207 (171)	1fcgA.pdb
SELP	membrane protein	P16109	42-199 (158)	1g1sA.pdb
EEA1	membrane protein	Q15075	1289-1411 (123)	1jocA.pdb
GGA1	membrane protein	Q9UJY5	494-639 (146)	1na8A.pdb
SDCBP	membrane protein	O00560	197-273 (82)	1r6jA.pdb
CLIC1	membrane protein	O00299	22-234 (213)	1rk4A.pdb
NGF	membrane protein	P01138	132-236 (99)	1sg1A.pdb
ANTXR2	membrane protein	P58335	38-218 (181)	1shuX.pdb
IL1RAPL1	membrane protein	Q9NZN1	403-561 (147)	1t3gA.pdb
PGLYRP3	membrane protein	Q96LB9	177-341 (165)	1twqA.pdb
CD3E	membrane protein	P07766	33-123 (91)	1xiwA.pdb
CFTR	membrane protein	P13569	388-671 (267)	1xmiA.pdb
TRPV2	membrane protein	Q9Y5S1	71-318 (244)	2f37A.pdb
SYNJ2BP	membrane protein	P57105	5-98 (100)	2jikA.pdb
GRIP1	membrane protein	Q9Y3R0	148-239 (94)	2jilA.pdb
SELENOS	membrane protein	Q9BQE4	52-121 (69)	2q2fA.pdb
CD59	membrane protein	P13987	26-102 (78)	2uwrA.pdb
RAMP2	membrane protein	O60895	58-135 (78)	2xvtA.pdb
ARHGEF1	membrane protein	Q92888**	22-233 (165)	3ab3D.pdb
CNKSR2	membrane protein	Q8WXI2	6-80 (74)	3bs5B.pdb
HLA-DRA	membrane protein	P01903	28-205 (178)	3c5jA.pdb
AGER	membrane protein	Q15109	23-240 (219)	3cjjA.pdb

IQGAP1	membrane protein	P46940	962-1339 (369)	3fayA.pdb
GRIK1	membrane protein	P39086	445-820 (256)	3fvoA.pdb
ADAM22	membrane protein	Q9P0K1	233-718 (486)	3g5cA.pdb
AQP4	membrane protein	P55087	32-254 (223)	3gd8A.pdb
RHCG	membrane protein	Q9UBD6	2-443 (403)	3hd6A.pdb
TRIM72	membrane protein	Q6ZMU5	278-470 (193)	3kb5A.pdb
MPP1	membrane protein	Q00013	282-458 (180)	3neyA.pdb
GLIPR1	membrane protein	P48060	22-214 (193)	3q2uA.pdb
PLXNA2	membrane protein	O75051	1490-1600 (102)	3q3jA.pdb
GORASP2	membrane protein	Q9H8Y8	7-208 (200)	3rleA.pdb
MAPKAP1	membrane protein	Q9BPZ7	372-490 (116)	3voqA.pdb
PILRA	membrane protein	Q9UKJ1	32-150 (120)	3wuzA.pdb
macA	membrane protein	Q74FY6*	24-346 (323)	4aanA.pdb
PMP2	membrane protein	P02689	1-132 (132)	4bvmA.pdb
DYSF	membrane protein	O75923	1-124 (127) 943-1051 (109)	4iqhA.pdb 4caiA.pdb
BECN1	membrane protein	Q14457	248-447 (195)	4ddpA.pdb
STING1	membrane protein	Q86WV6	155-337 (173)	4emtA.pdb
DLG1	membrane protein	Q12959	310-406 (97)	4g69A.pdb
FOLR1	membrane protein	P15328	30-233 (206)	4km6A.pdb
SLC4A1	membrane protein	P02730	57-350 (276)	4ky9A.pdb
MR1	membrane protein	Q95460	23-291 (262)	4l4vA.pdb
HLA-B	membrane protein	P01889	25-298 (274)	4lcyA.pdb
PRNP	membrane protein	P04156**	118-224 (107)	4n9oA.pdb
ESYT2	membrane protein	A0FGR8	363-659 (292)	4npjA.pdb
PVDR	membrane protein	P22290	211-508 (282)	4nuuA.pdb
LGR4 - fusion	membrane protein	Q9BXB1	27-399 (443)	4qxeA.pdb
PLK1	membrane protein	P53350	372-599 (223)	4rcpA.pdb
TOR1AIP1	membrane protein	Q5JTV8	360-583 (224)	4tvsA.pdb
VAMP8	membrane protein	Q9BV40	11-74 (64)	4wy4A.pdb

PGRMC1	membrane protein	O00264	72-179 (112)	4x8yA.pdb
GPC1	membrane protein	P35052	25-473 (411)	4ywtA.pdb
GLP1R	membrane protein	P43220	29-128 (100)	5e94H.pdb
SCN2B	membrane protein	O60939	30-148 (122)	5febA.pdb
ADORA2A - fusion	membrane protein	P29274	2-305 (387)	5iu4A.pdb
ADIPOR2	membrane protein	Q86V24	99-380 (282)	5lx9A.pdb
ZMPSTE24	membrane protein	O75844	10-474 (444)	5sytA.pdb
PTGES	membrane protein	O14684	5-152 (147)	5tl9A.pdb
CHRM2 - fusion	membrane protein	P08172	16-458 (384)	5zkcA.pdb
SLMAP	membrane protein	Q14BN4	2-135 (134)	6ar2A.pdb
FZD4 - fusion	membrane protein	Q9ULV1	181-513 (379)	6bd4A.pdb
C5AR1 - fusion	membrane protein	P21730	30-327 (370)	6c1rB.pdb
GRM5 - fusion	membrane protein	P41594	569-836 (409)	6ffiA.pdb
CCDC90B - fusion	membrane protein	Q9GZT6	62-126 (94)	6h9mA.pdb
TACR1 - fusion	membrane protein	P25103	27-327 (483)	6hlpA.pdb
MCOLN2	membrane protein	Q8IZK6	92-282 (173)	6hrrA.pdb
KDELR2	membrane protein	Q5ZKX9	1-207 (207)	6i6hA.pdb
DHODH	membrane protein	Q02127	29-395 (367)	6idjA.pdb
MPLZL1	membrane protein	O95297	38-158 (119)	6igwA.pdb
MFN2	membrane protein	O95140	24-418 (428)	6jfkA.pdb
GPR52 - fusion	membrane protein	Q9Y2T5	21-338 (441)	6li0A.pdb
AQP7	membrane protein	O14520	33-279 (247)	6qziA.pdb
LTC4S	membrane protein	Q16873	2-144 (143)	6r7dA.pdb
PTCH1	membrane protein	Q13635	149-423 (277) 842-935 (94)	6rtwA.pdb 6rvcA.pdb
ERVW-1	membrane protein	Q9UQF0	345-433 (89)	6rx1A.pdb
ERVFRD-1	membrane protein	P60508	380-468 (89)	6rx3A.pdb
CYSLTR2 - fusion	membrane protein	Q9NS75	29-322 (365)	6rz6A.pdb
SLC2A1	membrane protein	P11166	8-455 (448)	6thaA.pdb
HCRTR1	membrane protein	O43613	45-346 (301)	6todA.pdb

KCNMA1	membrane protein	Q12791	408-1121 (594)	6v5aA.pdb
SCN4B	membrane protein	Q8IWT1	37-154 (115)	6vsvA.pdb
JAGN1 - fusion	membrane protein	Q8N5M9	2-183 (397)	6wvdA.pdb
malE - fusion	membrane protein	P0AEX9	26-392 (571)	6zhoA.pdb
DDR2 - fusion	membrane protein	Q16832	561-849 (275)	7aymA.pdb
GABARAP - fusion	membrane protein	O95166**	1-116 (132)	7brqA.pdb

[a] The Protein Data Bank (1) was used to identify folded regions within proteins. Originally, we searched for folded regions within proteins known to exhibit phase separation behavior, finding 82 folded regions (2). The phase-separating proteins were obtained from lists compiled by Vernon et al (3), the PhaSePro database (4), and the DisProt database (5). A complete list of these 82 folded regions has been published elsewhere (2). To that list, we added folded regions from 122 human proteins with nonhomologous structures obtained from Wang et al (6), 32 proteins with small to large structures obtained from Fitzkee and Rose (7), 54 extremophile proteins obtained from Panja et al (8), 53 metamorphic proteins obtained from Chen et al (9), and 90 membrane proteins that were found by searching the Protein Data Bank for the phrase "membrane protein." Duplicate entries were removed from the combined list. For example, human Galectin-3 (UniProt accession number P17931) is found in both the PhaSePro database of phase-separating proteins and the list of human proteins with nonhomologous structures from Wang et al. Duplicate entries in the combined list are identified by an asterisk at the end of the UniProt accession number; two asterisks indicate the duplicate that was removed from the final folded set. Protein names with "- fusion" indicate a protein that is fused to another protein in the crystallographic structure, which is found among a few classified as "membrane protein".

[b] Residue positions with resolved atomic coordinates in a PDB structure (x-ray or NMR) were used to verify regions ($N \geq 20$) that fold. Unresolved residues were not included in folded regions. Protein sequences were extracted from the referenced PDB file and thus may contain substitutions, deletions, and/or insertions (excluding histidine affinity tags) compared to the UniProt sequence. The value of N in parenthesis is the length of the extracted sequence.

Table S2. Summary of mean v_{model} in the ID and folded sequence subsets.

Set	Number	v_{model} [a]	*t*-test [b]	*U*-test [b]
Previous ID	23	0.558 ± 0.019	-	-
BMRB & DisProt	98	0.558 ± 0.023	0.44	0.48
Previous Folded	82	0.536 ± 0.008	-	-
Human	122	0.536 ± 0.007	0.40	0.32
Small-to-large	32	0.537 ± 0.009	0.36	0.41
Extremophile	54	0.542 ± 0.011	$1.2e^{-4}$	$2.4e^{-4}$
Membrane	90	0.537 ± 0.006	0.17	0.21
Metamorphic	53	0.537 ± 0.006	0.15	0.18

[a] Mean ± standard deviation.
[b] One-tail *p*-value, where values <0.05 indicate the compared sets are statistically different in their means. Comparisons are to the previous set; BMRB & DisProt to the Previous ID, and Human, Small-to-large, Extremophile, Membrane, and Metamorphic to the Previous Folded.

Table S3. Summary of mean β-turn propensity in the ID and folded sequence subsets.

Set	Number	β-turn propensity [a]	*t*-test [b]	*U*-test [b]
Previous ID	23	1.062 ± 0.082	-	-
BMRB & DisProt	98	1.110 ± 0.071	$6.5e^{-3}$	$9.3e^{-4}$
Previous Folded Set	82	0.969 ± 0.039	-	-
Human	122	0.980 ± 0.039	0.03	0.07
Small-to-large	32	0.968 ± 0.027	0.42	0.34
Extremophile	54	0.983 ± 0.030	0.01	0.03
Membrane	90	0.956 ± 0.046	0.02	0.02
Metamorphic	53	0.972 ± 0.040	0.30	0.48

[a] Mean ± standard deviation.

[b] One-tail *p*-value, where values <0.05 indicate the compared sets are statistically different in their means. Comparisons are to the previous set; BMRB & DisProt to the Previous ID, and Human, Small-to-large, Extremophile, Membrane, and Metamorphic to the Previous Folded.

Table S4. List of IDRs not known to exhibit phase separation behavior.

Name	Database [a]	Entry number	UniProt accession number	ID region (*N*)
pknG	BMRB	26027	P9WI73	1-75 (75)
HCK	BMRB	27554	P08631	2-79
SIC1	BMRB	16657	P38634	1-90 (90)
SLC9A1	BMRB	26557	P19634	680-815 (136)
ERD14	BMRB	16876	P42763	1-185 (185)
Spp1	DisProt	DP01448	P10923	17-294 (278)
PAGE4	DisProt	DP01435	O60829	1-102 (102)
MAP2K4	DisProt	DP01400	P45985	1-86 (86)
Sufu	DisProt	DP01397	Q9Z0P7	279-359 (81)
HCN1	DisProt	DP01317	O60741	1-93 (93)
SUFU	DisProt	DP01312	Q9UMX1	279-360 (82)
PQBP1	DisProt	DP01308	O60828	82-265 (184)
HIRD11	DisProt	DP01300	Q9SLJ2	1-98 (98)
LEA18	DisProt	DP01299	Q96273	1-97 (97)
PSEN1	DisProt	DP01292	P49768	1-77 (77)
Prothymosin a14	DisProt	DP01228	Q9UMZ1	1-101 (101)
Ppp1r10	DisProt	DP01202	O55000	309-433 (125)
NOLC1	DisProt	DP01178	Q14978	1-699 (699)
Gja4	DisProt	DP01175	P28235	233-333 (101)
DCLRE1C	DisProt	DP01162	Q96SD1	480-575 (96)
ptkA	DisProt	DP01160	P9WPI9	1-81 (81)
H1-0	DisProt	DP01156	P07305	105-194 (90)
CHZ1	DisProt	DP01135	P40019	1-153 (153)
Caskin1	DisProt	DP01127	Q8VHK2	603-1430 (828)
Ttn-1	DisProt	DP01090	A0A2I2LG13	2793-6678 (3886)
PM28	DisProt	DP01088	Q9XES8	1-89 (89)
YRB2	DisProt	DP01079	P40517	1-203 (203)

Ahn-1	DisProt	DP01074	Q7YUB9	1-86 (86)
MSA2	DisProt	DP01067	P19599	21-238 (218)
LMP2A	DisProt	DP01060	A8CDV5	1-118 (118)
Omega gliadin storage protein	DisProt	DP01040	Q9FUW7	1-280 (280)
SLE2	DisProt	DP01036	I1JLC8	1-105 (105)
pscP	DisProt	DP00993	Q9I332	1-253 (253)
Small delta antigen	DisProt	DP00965	P0C6L3	60-195 (136)
SBDS-like protein	DisProt	DP00957	C0J347	264-464 (201)
GAP43	DisProt	DP00955	P06836	1-242 (242)
N	DisProt	DP00948	P59595	182-259 (78)
Ppp1r9b	DisProt	DP00943	O35274	1-154 (154)
BASP1	DisProt	DP00930	P80723	1-227 (227)
NABP2	DisProt	DP00864	Q9BQ15	110-211 (102)
trm10	DisProt	DP00798	O14214	1-83 (83)
CNGB1	DisProt	DP00768	Q28181-4	14-99 (86) 272-590 (319)
Smtnl1	DisProt	DP00742	Q99LM3	1-341 (341)
dre4	DisProt	DP00721	Q8IRG6	889-1044 (156)
Ssrp	DisProt	DP00720	Q05344	437-554 (118) 625-723 (99)
N	DisProt	DP00698	O89339	400-532 (133)
RYBP	DisProt	DP00694	Q8N488	1-228 (228)
L1CAM	DisProt	DP00666	P32004	1144-1257 (114)
GMPM1	DisProt	DP00664	Q01417	1-173 (173)
ALB3	DisProt	DP00662	Q8LBP4	339-462 (124)
MAC-41A	DisProt	DP00659	P16458	233-385 (153)
COR47	DisProt	DP00657	P31168	1-265 (265)
N	DisProt	DP00640	Q89933	400-525 (126)
ERD10	DisProt	DP00606	P42759	1-260 (260)
Genome polyprotein	DisProt	DP00588	P27958	1-82 (82)
stm	DisProt	DP00584	A2VD23	1-613 (613)

SEPTIN4	DisProt	DP00537	O43236	1-119 (119)
DHN1	DisProt	DP00530	P12950	1-168 (168)
MYOM1	DisProt	DP00517	P52179	836-931 (96)
NUPR1	DisProt	DP00510	O60356	1-82 (82)
UBA2	DisProt	DP00486	Q9UBT2	551-640 (90)
HY5	DisProt	DP00469	O24646	1-77 (77)
cna	DisProt	DP00461	P08083	1-90 (90)
Chm	DisProt	DP00458	P37727	108-208 (101)
PPP1R1B	DisProt	DP00421	P07516	1-202 (202)
JAG1	DisProt	DP00418	P78504	1094-1218 (125)
URE1	DisProt	DP00353	P23202	1-90 (90)
DNAJC6	DisProt	DP00351	Q27974	547-813 (267)
col	DisProt	DP00342	P09883	1-83 (83)
Trl	DisProt	DP00328	Q08605	368-444 (77)
PPP1R1A	DisProt	DP00325	P01099	1-166 (166)
ADD2	DisProt	DP00241	P35612	409-726 (318)
ADD1	DisProt	DP00240	P35611	430-737 (308)
SSB	DisProt	DP00229	P05455	326-408 (83)
Nucleoplasmin	DisProt	DP00217	P05221	120-200 (81)
CAST	DisProt	DP00196	P20810	137-277 (141)
HMGN2	DisProt	DP00195	P02313	1-89 (89)
Late embryogenesis abundant protein 1	DisProt	DP00186	Q95V77	1-143 (143)
CTDP1	DisProt	DP00177	Q9Y5B0	879-961 (83)
TCF7L2	DisProt	DP00175	Q9NQB0	1-130 (130)
zipA	DisProt	DP00161	P77173	86-185 (100)
RAD23A	DisProt	DP00156	P54725	79-160 (82)
NEFL	DisProt	DP00151	P02547	444-549 (106)
Slbp	DisProt	DP00144	Q9VAN6	97-175 (79)
PTHLH	DisProt	DP00138	P12272	68-144 (77)

H1-4	DisProt	DP00136	P15865	1-217 (217)
PRB4	DisProt	DP00119	P10163	17-310 (294)
Desiccation-related protein clone PCC6-19	DisProt	DP00112	P22239	1-155 (155)
H1-0	DisProt	DP00097	P10922	96-193 (98)
TOP2	DisProt	DP00076	P06786	1178-1428 (251)
TOP1	DisProt	DP00075	P11387	1-214 (214)
Structural polyprotein	DisProt	DP03350	P03316	1-113 (113)
RPA1	DisProt	DP00061	P27694	105-180 (76)
HMGA1	DisProt	DP00040	P17096	1-107 (107)
HMGN2	DisProt	DP00039	P05204	1-90 (90)
RAP1	DisProt	DP00020	P11938	1-123 (123)

[a] The Biological Magnetic Resonance Data Bank (BMRB) (10) and DisProt (5) databases were used to identify IDRs that are not known to exhibit phase separation behavior. This list of verified IDRs, wherein duplicates have been removed, was combined with a list of 23 IDRs that have been identified and reported elsewhere (2).

Table S5. Enthalpy, entropy, and free energy of phase separation of A1-LCD and Ddx4 mutants.

IDR	Mutant	Primary Sequence	Δh° [a]	$\Delta s^{\circ}/R$ [b]	Δg° [c]
Ddx4	CS	MGDRDWRAEINPHMSSYVPIFEKDRYSGENGRNFNDTP ASSSEMRDGPSERDHFMKSGFASGDNFGNRDAGKCNER DNTSTMGGFGVGKSFGNEGFSNSRFERGDSSGFWRESS NDCRDNPTRNDGFSDRGGYEKGNNSEASGPYERGGRGS FDGCRGGFGLGSPNNRLDPRECMQRTGGLFGSDRPVLS GTGNGDTSQSRSGSGSERGGYKGLNEKVITGSGENSWK SEARGGES	-23.09	38.82	-44.16
Ddx4	WT	MGDEDWEAEINPHMSSYVPIFEKDRYSGENGDNFNRTP ASSSEMDDGPSRRDHFMKSGFASGRNFGNRDAGECNKR DNTSTMGGFGVGKSFGNRGFSNSRFEDGDSSGFWRESS NDCEDNPTRNRGFSKRGGYRDGNNSEASGPYRRGGRGS FRGCRGGFGLGSPNNDLDPDECMQRTGGLFGSRRPVLS GTGNGDTSQSRSGSGSERGGYKGLNEEVITGSGKNSWK SEAEGGES	-5.43	8.47	-10.03
A1-LCD	Aro+	GSMAFASSFQRGRYGSGNFGGGRGGGFGGNDNFGRGGN FSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGS YNDFGNYNNQSSNFGPMKGGNFGGRSSGGSYGGGQYFA KPRNQGGYGGSSFSSSYGSGRRF	-30.58	43.50	-54.18
A1-LCD	Aro-	GSMASASSSQRGRSGSGNSGGGRGGGFGGNDNFGRGGN SSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNSGGGGS SNDFGNYNNQSSNFGPMKGGNFGGRSSGGSGGGGQYSA KPRNQGGYGGSSSSSSSGSGRRF	-17.44	27.00	-32.10
A1-LCD	-12F+12Y	GSMASASSSQRGRSGSGNYGGGRGGGYGGNDNYGRGGN YSGRGGYGGSRGGGGYGGSGDGYNGYGNDGSNYGGGGS YNDYGNYNNQSSNYGPMKGGNYGGRSSGGSGGGGQYYA KPRNQGGYGGSSSSSSYGSGRRY	-27.55	40.89	-49.74
A1-LCD	-9F+6Y	GSMASASSSQRGRSGSGNFGGGRGGGYGGNDNYGRGGN YSGRGGFGGSRGGGGYGGSGDGYNGGGNDGSNYGGGGS YNDSGNYNNQSSNFGPMKGGNYGGRSSGGSGGGGQYGA KPRNQGGYGGSSSSSSYGSGRRY	-25.16	38.78	-46.21
A1-LCD	-4D	GSMASASSSQRGRSGSGNFGGGRGGGFGGNGNFGRGGN FSGRGGFGGSRGGGGYGGSGGGYNGFGNSGSNFGGGGS YNGFGNYNNQSSNFGPMKGGNFGGRSSGPYGGGGQYFA KPRNQGGYGGSSSSSSYGSGRRF	-25.05	39.72	-46.61
A1-LCD	-9F+3Y	GSMASASSSQRGRSGSGNFGGGRGGGYGGNDNGGRGGN YSGRGGFGGSRGGGGYGGSGDGYNGGGNDGSNYGGGGS YNDSGNGNNQSSNFGPMKGGNYGGRSSGGSGGGGQYGA KPRNQGGYGGSSSSSSYGSGRRS	-24.51	38.89	-45.62
A1-LCD	-6R+6K	GSMASASSSQKGKSGSGNFGGGRGGGFGGNDNFGKGGN FSGRGGFGGSKGGGGYGGSGDGYNGFGNDGSNFGGGGS YNDFGNYNNQSSNFGPMKGGNFGGKSSGGSGGGGQYFA KPRNQGGYGGSSSSSSYGSGRKF	-24.29	39.95	-45.98
A1-LCD	-8F+4Y	GSMASASSSQRGRSGSGNFGGGRGGGYGGNDNGGRGGN YSGRGGFGGSRGGGGYGGSGDGYNGGGNDGSNYGGGGS YNDSGNYNNQSSNFGPMKGGNYGGRSSGGSGGGGQYGA KPRNQGGYGGSSSSSSYGSGRRF	-23.91	37.37	-44.19
A1-LCD	+7R+12D	GSMASADSSQRDRDDRGNFGDGRGGGFGGNDNFGRGGN FSDRGGFGGSRGDGRYGGDGDRYNGFGNDGRNFGGGGS YNDFGNYNNQSSNFDPMKGGNFRDRSSGPYDRGGQYFA KPRNQGGYGGSSSSSRSYGSDRRF	-22.45	30.44	-38.97
A1-LCD	+2R	GSMASASSSQRGRSGSGNFGGGRGGGFGGNDNFGRGGN FSGRGGFGGSRGGGGYGGSGDGYNGFRNDGSNFGGGGR YNDFGNYNNQSSNFGPMKGGNFGGRSSGPYGGGGQYFA KPRNQGGYGGSSSSSSYGSGRRF	-21.74	31.96	-39.09
A1-LCD	-2R-2K+3D	GSMASASSSQDGRSGSGNFGGGRGGGFGGNDNFGRGGN FSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGS	-20.92	29.85	-37.12

		YNDFGNYNNQSSNFGPMDGGNFGGRSSGPYGGGGQYFA DPRNQGGYGGSSSSSSYGSGGRF			
A1-LCD	WT+NLS	GSMASASSSQRGRSGSGNFGGGRGGGFGGNDNFGRGGN FSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGS YNDFGNYNNQSSNFGPMKGGNFGGRSSGPYGGGGQYFA KPRNQGGYGGSSSSSSYGSGRRF	-20.27	28.79	-35.90
A1-LCD	+8D	GSMASASSSQRDRSGSGNFGGGRDGGFGGNDNFGRGDN FSGRGDFGGSRDGGGYGGSGDGYNGFGNDGSNFGGGGS YNDFGNYNNQSSNFGPMKGGNFGGRSSDPYGGGGQYFA KPRNQDGYGGSSSSSSYDSGRRF	-20.27	29.50	-36.28
A1-LCD	WT	GSMASASSSQRGRSGSGNFGGGRGGGFGGNDNFGRGGN FSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGS YNDFGNYNNQSSNFGPMKGGNFGGRSSGGSGGGGQYFA KPRNQGGYGGSSSSSSYGSGRRF	-20.22	28.91	-35.91
A1-LCD	-3R+3K	GSMASASSSQRGKSGSGNFGGGRGGGFGGNDNFGRGGN FSGRGGFGGSKGGGGYGGSGDGYNGFGNDGSNFGGGGS YNDFGNYNNQSSNFGPMKGGNFGGRSSGGSGGGGQYFA KPRNQGGYGGSSSSSSYGSGRKF	-19.95	30.44	-36.47
A1-LCD	-2K	GSMASASSSQRGRSGSGNFGGGRGGGFGGNDNFGRGGN FSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGS YNDFGNYNNQSSNFGPMGGGNFGGRSSGPYGGGGQYFA GPRNQGGYGGSSSSSSYGSGRRF	-19.62	26.09	-33.78
A1-LCD	+4D	GSMASASSSQRDRSGSGNFGGGRGGGFGGNDNFGRGGN FSGRGDFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGS YNDFGNYNNQSSNFGPMKGGNFGGRSSDPYGGGGQYFA KPRNQGGYGGSSSSSSYDSGRRF	-17.88	23.62	-30.70
A1-LCD	+7K+12D	GSMASADSSQRDRDDKGNFGDGRGGGFGGNDNFGRGGN FSDRGGFGGSRGDGKYGGDGDKYNGFGNDGKNFGGGGS YNDFGNYNNQSSNFDPMKGGNFKDRSSGPYDKGGQYFA KPRNQGGYGGSSSSKSYGSDRRF	-17.55	25.50	-31.39
A1-LCD	+12D	GSMASADSSQRDRDDSGNFGDGRGGGFGGNDNFGRGGN FSDRGGFGGSRGDGGYGGDGDGYNGFGNDGSNFGGGGS YNDFGNYNNQSSNFDPMKGGNFGDRSSGPYDGGGQYFA KPRNQGGYGGSSSSSSYGSDRRF	-17.01	25.15	-30.66
A1-LCD	-6R	GSMASASSSQGGRSGSGNFGGGRGGGFGGNDNFGGGGN FSGSGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGS YNDFGNYNNQSSNFGPMKGGNFGGSSSGPYGGGGQYFA KPGNQGGYGGSSSSSSYGSGGRF	-16.90	22.33	-29.02
A1-LCD	+7F-7Y	GSMASASSSQRGRSGSGNFGGGRGGGFGGNDNFGRGGN FSGRGGFGGSRGGGGFGGSGDGFNGFGNDGSNFGGGGS FNDFGNFNNQSSNFGPMKGGNFGGRSSGGSGGGGQFFA KPRNQGGFGGSSSSSSFGSGRRF	-16.47	23.39	-29.17
A1-LCD	+12E	GSMASAESSQREREESGNFGEGRGGGFGGNDNFGRGGN FSERGGFGGSRGEGGYGGEGDGYNGFGNDGSNFGGGGS YNDFGNYNNQSSNFEPMKGGNFGERSSGPYEGGGQYFA KPRNQGGYGGSSSSSSYGSERRF	-15.76	23.62	-28.58
A1-LCD	+7R+10D	GSMASADSSQRDRDGRGNFGDGRGGGFGGNDNFGRGGN FSDRGGFGGSRGGGRYGGDGDRYNGFGNDGRNFGGGGS YNDFGNYNNQSSNFDPMKGGNFRDRSSGPYDRGGQYFA KPRNQGGYGGSSSSRSYGSDRRF	-14.18	17.05	-23.43
A1-LCD	-10R	GSMASASSSQGGSSGSGNFGGGGGGGFGGNDNFGGGGN FSGSGGFGGSGGGGGYGGSGDGYNGFGNDGSNFGGGGS YNDFGNYNNQSSNFGPMKGGNFGGSSSGPYGGGGQYFA KPGNQGGYGGSSSSSSYGSGGGF	-13.53	19.16	-23.93
A1-LCD	-4R-2K+5D	GSMASASSSQDGRSGSGNFGGGDGGGFGGNDNFGRGGN FSGGGGFGGSRGGGGYGGSGDGYNGFGNDGSNFGGGGS YNDFGNYNNQSSNFGPMDGGNFGGRSSGPYGGGGQYFA DPRNQGGYGGSSSSSSYGSGDRF	-12.83	17.63	-22.40

[a] Standard molar enthalpy ($\Delta h°$) in units of kcal/mol. Values for Ddx4 CS, Ddx4 WT, A1-LCD Aro+, and A1-LCD Aro- were calculated from the temperature dependence to c_{sat} (see Methods) using

c_{sat} values digitally extracted from Figures 1C-D in Brady et al (11) and Figure 3F in Martin et al (12). Values for all other IDRs in this table were digitally extracted from Supplementary Figure 7D in Bremer et al (13).

[b] Standard molar entropy ($\Delta s°$) divided by the universal gas constant (R), and thus dimensionless. Values for Ddx4 CS, Ddx4 WT, A1-LCD Aro+, and A1-LCD Aro- were calculated from the temperature dependence to c_{sat} (see Methods) using c_{sat} values digitally extracted from Figures 1C-D in Brady et al (11) and Figure 3F in Martin et al (12). Values for all other IDRs in this table were digitally extracted from Supplementary Figure 7E in Bremer et al (13).

[c] Standard molar free energy ($\Delta g°$) in units of kcal/mol. Values were calculated from $\Delta h°$ and $\Delta s°$ using the equation, $\Delta g° = \Delta h° - T\Delta s°$, where T is the standard temperature (273.15 K).

Table S6. Saturation concentration (at 4 °C) of A1-LCD mutants.

Mutant	Primary Sequence	c_{sat} [a]
+23G-23S-12F+12Y	GSMAGAGGGQRGRGGGGNYGGGRGGGYGGNDNYGRGGNYGGRGGYGGGRG GGGYGGGGDGYNGYGNDGGNYGGGGGYNDYGNYNNQGGNYGPMKGGNYGG RGGGGGGGGGQYYAKPRNQGGYGGGGGGGGYGGGRRY	4.86E-07
+7R+12D	GSMASADSSQRDRDDRGNFGDGRGGGFGGNDNFGRGGNFSDRGGFGGSRG DGRYGGDGDRYNGFGNDGRNFGGGGSYNDFGNYNNQSSNFDPMKGGNFRD RSSGPYDRGGQYFAKPRNQGGYGGSSSSRSYGSDRRF	7.49E-07
-12F+12Y	GSMASASSSQRGRSGSGNYGGGRGGGYGGNDNYGRGGNYSGRGGYGGSRG GGGYGGSGDGYNGYGNDGSNYGGGGSYNDYGNYNNQSSNYGPMKGGNYGG RSSGGSGGGGQYYAKPRNQGGYGGSSSSSSYGSGRRY	2.74E-06
+4D	GSMASASSSQRDRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGDFGGSRG GGGYGGSGDGYNGFGNDGSNFGGGGSYNDFGNYNNQSSNFGPMKGGNFGG RSSDPYGGGGQYFAKPRNQGGYGGSSSSSSYDSGRRF	4.04E-06
-6R	GSMASASSSQGGRSGSGNFGGGRGGGFGGNDNFGGGGNFSGSGGFGGSRG GGGYGGSGDGYNGFGNDGSNFGGGGSYNDFGNYNNQSSNFGPMKGGNFGG SSSGPYGGGGQYFAKPGNQGGYGGSSSSSSYGSGGRF	7.34E-06
-2R-2K+3D	GSMASASSSQDGRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRG GGGYGGSGDGYNGFGNDGSNFGGGGSYNDFGNYNNQSSNFGPMDGGNFGG RSSGPYGGGGQYFADPRNQGGYGGSSSSSSYGSGGRF	9.91E-06
-20G+20S-12F+12Y	GSMASASSSQRSRSGSGNYSGSRSGSYSGNDNYGRSGNYSGRSGYGGSRS GGGYSGSGDSYNSYGNDGSNYSGSGSYNDYGNYNNQSSNYGPMKSGNYGG RSSGSSGGSGQYYAKPRNQGSYSGSSSSSSYGSSRRY	1.22E-05
WT+NLS	GSMASASSSQRGRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRG GGGYGGSGDGYNGFGNDGSNFGGGGSYNDFGNYNNQSSNFGPMKGGNFGG RSSGPYGGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF	1.25E-05
WT	GSMASASSSQRGRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRG GGGYGGSGDGYNGFGNDGSNFGGGGSYNDFGNYNNQSSNFGPMKGGNFGG RSSGGSGGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF	1.25E-05
-30G+30S-12F+12Y	GSMASASSSQRSRSSSGNYSGSRSGSYSGNDNYGRSGNYSGRSGYSGSRS GSGYSGSSDSYNSYGNDSSNYSGSSSYNDYGNYNNQSSNYGPMKSGNYSG RSSSSSGSSGQYYAKPRNQGSYSGSSSSSSYSSSRRY	1.47E-05
+2R	GSMASASSSQRGRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRG GGGYGGSGDGYNGFRNDGSNFGGGGRYNDFGNYNNQSSNFGPMKGGNFGG RSSGPYGGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF	1.81E-05
+8D	GSMASASSSQRDRSGSGNFGGGRDGGFGGNDNFGRGDNFSGRGDFGGSRD GGGYGGSGDGYNGFGNDGSNFGGGGSYNDFGNYNNQSSNFGPMKGGNFGG RSSDPYGGGGQYFAKPRNQDGYGGSSSSSSYDSGRRF	1.84E-05
-10G+10S	GSMASASSSQRSRSGSGNFGGGRSGGFGGNDNFGRSGNFSGRGGFGGSRG GGGYGGSGDSYNGFGNDGSNFGGSGSYNDFGNYNNQSSNFGPMKSGNFGG RSSGSSGGSGQYFAKPRNQGSYSGSSSSSSYGSGRRF	2.76E-05
-9F+6Y	GSMASASSSQRGRSGSGNFGGGRGGGYGGNDNYGRGGNYSGRGGFGGSRG GGGYGGSGDGYNGGGNDGSNYGGGGSYNDSGNYNNQSSNFGPMKGGNYGG RSSGGSGGGGQYGAKPRNQGGYGGSSSSSSYGSGRRY	2.80E-05
+7K+12D	GSMASADSSQRDRDDKGNFGDGRGGGFGGNDNFGRGGNFSDRGGFGGSRG DGKYGGDGDKYNGFGNDGKNFGGGGSYNDFGNYNNQSSNFDPMKGGNFKD RSSGPYDKGGQYFAKPRNQGGYGGSSSSKSYGSDRRF	4.31E-05
+7F-7Y	GSMASASSSQRGRSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRG GGGFGGSGDGFNGFGNDGSNFGGGGSFNDFGNFNNQSSNFGPMKGGNFGG RSSGGSGGGGQFFAKPRNQGGFGGSSSSSSFGSGRRF	4.94E-05
-20G+20S	GSMASASSSQRSRSGSGNFSGSRSGSFSGNDNFGRSGNFSGRSGFGGSRS GGGYSGSGDSYNSFGNDGSNFSGSGSYNDFGNYNNQSSNFGPMKSGNFGG RSSGSSGGSGQYFAKPRNQGSYSGSSSSSSYGSSRRF	5.39E-05
-8F+4Y	GSMASASSSQRGRSGSGNFGGGRGGGYGGNDNGGRGGNYSGRGGFGGSRG GGGYGGSGDGYNGGGNDGSNYGGGGSYNDSGNYNNQSSNFGPMKGGNYGG RSSGGSGGGGQYGAKPRNQGGYGGSSSSSSYGSGRRF	6.26E-05
+23G-23S+7F-7Y	GSMAGAGGGQRGRGGGGNFGGGRGGGFGGNDNFGRGGNFGGRGGFGGGRG GGGFGGGGDGFNGFGNDGGNFGGGGGFNDFGNFNNQGGNFGPMKGGNFGG RGGGGGGGGGQFFAKPRNQGGFGGGGGGGGFGGGRRF	7.63E-05

-3R+3K	GSMASASSSQRGKSGSGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSKG GGGYGGSGDGYNGFGNDGSNFGGGGSYNDFGNYNNQSSNFGPMKGGNFGG RSSGGSGGGGQYFAKPRNQGGYGGSSSSSSYGSGRKF	8.30E-05
-4D	GSMASASSSQRGRSGSGNFGGGRGGGFGGNGNFGRGGNFSGRGGFGGSRG GGGYGGSGGGYNGFGNSGSNFGGGGSYNGFGNYNNQSSNFGPMKGGNFGG RSSGPYGGGGQYFAKPRNQGGYGGSSSSSSYGSGRRF	8.69E-05
-30G+30S+7F-7Y	GSMASASSSQRSRSSSGNFSGSRSGSFSGNDNFGRSGNFSGRSGFSGSRS GSGFSGSSDSFNSFGNDSSNFSGSSSFNDFGNFNNQSSNFGPMKSGNFSG RSSSSSGSSSGQFFAKPRNQGSFSGSSSSSSFSSSRRF	8.98E-05
-20G+20S+7F-7Y	GSMASASSSQRSRSGSGNFSGSRSGSFSGNDNFGRSGNFSGRSGFGGSRS GGGFSGSGDSFNSFGNDGSNFSGSGSFNDFGNFNNQSSNFGPMKSGNFGG RSSGSSGGSGQFFAKPRNQGSFSGSSSSSSFGSSRRF	9.87E-05
+12D	GSMASADSSQRDRDDSGNFGDGRGGGFGGNDNFGRGGNFSDRGGFGGSRG DGGYGGDGDGYNGFGNDGSNFGGGGSYNDFGNYNNQSSNFDPMKGGNFGD RSSGPYDGGGQYFAKPRNQGGYGGSSSSSSYGSDRRF	9.96E-05
-4R-2K+5D	GSMASASSSQDGRSGSGNFGGGDGGGFGGNDNFGRGGNFSGGGGFGGSRG GGGYGGSGDGYNGFGNDGSNFGGGGSYNDFGNYNNQSSNFGPMDGGNFGG RSSGPYGGGGQYFADPRNQGGYGGSSSSSSYGSGDRF	1.06E-04
-9F+3Y	GSMASASSSQRGRSGSGNFGGGRGGGYGGNDNGGRGGNYSGRGGFGGSRG GGGYGGSGDGYNGGGNDGSNYGGGGSYNDSGNGNNQSSNFGPMKGGNYGG RSSGGSGGGGQYGAKPRNQGGYGGSSSSSSYGSGRRS	1.13E-04
-10R	GSMASASSSQGGSSGSGNFGGGGGGGFGGNDNFGGGGNFSGSGGFGGSGG GGGYGGSGDGYNGFGNDGSNFGGGGSYNDFGNYNNQSSNFGPMKGGNFGG SSSGPYGGGGQYFAKPGNQGGYGGSSSSSSYGSGGGF	1.34E-04
+7R	GSMASASSSQRGRSGRGNFGGGRGGGFGGNDNFGRGGNFSGRGGFGGSRG GGRYGGSGDRYNGFGNDGRNFGGGGSYNDFGNYNNQSSNFGPMKGGNFRG RSSGPYGRGGQYFAKPRNQGGYGGSSSSRSYGSGRRF	1.78E-04
+12E	GSMASAESSQREREESGNFGEGRGGGFGGNDNFGRGGNFSERGGFGGSRG EGGYGGEGDGYNGFGNDGSNFGGGGSYNDFGNYNNQSSNFEPMKGGNFGE RSSGPYEGGGQYFAKPRNQGGYGGSSSSSSYGSERRF	2.12E-04
Aro-	GSMASASSSQRGRSGSGNSGGGRGGGFGGNDNFGRGGNSSGRGGFGGSRG GGGYGGSGDGYNGFGNDGSNSGGGGSSNDFGNYNNQSSNFGPMKGGNFGG RSSGGSGGGGQYSAKPRNQGGYGGSSSSSSSGSGRRF	3.01E-04
-6R+6K	GSMASASSSQKGKSGSGNFGGGRGGGFGGNDNFGKGGNFSGRGGFGGSKG GGGYGGSGDGYNGFGNDGSNFGGGGSYNDFGNYNNQSSNFGPMKGGNFGG KSSGGSGGGGQYFAKPRNQGGYGGSSSSSSYGSGRKF	4.96E-04

[a] Saturation concentration (c_{sat}) in molarity (*M*). Values were digitally extracted from Figures 1-5 and Supplementary Figures 2, 4, 6, in Bremer et al (13) and Figure 3F in Martin et al (12).

Table S7. List of 500 proteins with the highest summed P classifier distance in the human proteome.

∑ P class. dist.	longest PS IDR	first residue	last residue	UniProt ID and protein
14249.16	2913	1714	4626	Q7Z5P9|MUC19_HUMAN Mucin-19 OS=Homo sapiens OX=9606 GN=MUC19 PE=1
10117.86	5705	995	6699	A0A0G2JR97|A0A0G2JR97_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
10109.08	5693	995	6687	A0A0G2JS65|A0A0G2JS65_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
10099.64	5441	995	6435	A0A0G2JR46|A0A0G2JR46_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
10092.09	5693	995	6687	A0A0G2JQK9|A0A0G2JQK9_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
10045.46	5453	995	6447	A0A0G2JRD8|A0A0G2JRD8_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
10040.92	5455	995	6449	A0A0G2JRY3|A0A0G2JRY3_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
10030.54	5469	995	6463	A0A0G2JRJ6|A0A0G2JRJ6_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
10021.77	5411	995	6405	A0A0G2JRS2|A0A0G2JRS2_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
10021.77	5521	995	6515	A0A0G2JQI2|A0A0G2JQI2_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
9926.97	2689	161	2849	Q86YZ3|HORN_HUMAN Hornerin OS=Homo sapiens OX=9606 GN=HRNR PE=1 SV
7899.66	222	8676	8897	Q8WXI7|MUC16_HUMAN Mucin-16 OS=Homo sapiens OX=9606 GN=MUC16 PE=1
7632.47	5095	251	5345	Q9UKN1|MUC12_HUMAN Mucin-12 OS=Homo sapiens OX=9606 GN=MUC12 PE=1
6761.51	3574	963	4536	A0A0G2JQT8|A0A0G2JQT8_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6752.73	3571	963	4533	A0A0G2JRA1|A0A0G2JRA1_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6744.49	3441	963	4403	A0A0G2JS91|A0A0G2JS91_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6735.74	3571	963	4533	A0A0G2JQC6|A0A0G2JQC6_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6690.45	3453	963	4415	A0A0G2JSB4|A0A0G2JSB4_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6685.78	3455	963	4417	A0A0G2JSD9|A0A0G2JSD9_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6676.71	2716	1884	4599	Q02817|MUC2_HUMAN Mucin-2 OS=Homo sapiens OX=9606 GN=MUC2 PE=1 SV=
6675.41	3469	963	4431	A0A0G2JS19|A0A0G2JS19_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6666.63	3521	963	4483	A0A0G2JR43|A0A0G2JR43_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6666.63	3411	963	4373	A0A0G2JRE6|A0A0G2JRE6_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6658.20	3565	990	4554	E7ENC5|E7ENC5_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC4 PE=1 S
6649.42	3562	990	4551	E9PDY6|E9PDY6_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC4 PE=1 S
6649.42	3565	963	4527	A0A0G2JMX1|A0A0G2JMX1_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6642.42	3556	891	4446	A0A0G2JM16|A0A0G2JM16_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
6640.64	3562	963	4524	A0A0G2JNM3|A0A0G2JNM3_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6633.80	3440	990	4429	E7EQG8|E7EQG8_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC4 PE=1 S

6632.43	3562	990	4551	E7EWN1|E7EWN1_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC4 PE=1 S
6626.06	3562	964	4525	A0A0G2JN54|A0A0G2JN54_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
6625.01	3430	963	4392	A0A0G2JQA9|A0A0G2JQA9_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6625.01	3440	963	4402	A0A0G2JS42|A0A0G2JS42_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6623.64	3562	963	4524	A0A0G2JRT1|A0A0G2JRT1_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6585.80	3452	990	4441	E7ERK0|E7ERK0_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC4 PE=1 S
6581.26	3454	990	4443	E7EUL9|E7EUL9_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC4 PE=1 S
6577.01	3452	963	4414	A0A0G2JRW6|A0A0G2JRW6_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6572.47	3454	963	4416	A0A0G2JRV5|A0A0G2JRV5_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6570.88	3468	990	4457	E7EQT2|E7EQT2_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC4 PE=1 S
6562.10	3520	990	4509	E7ETT5|E7ETT5_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC4 PE=1 S
6562.10	3410	990	4399	E7EW47|E7EW47_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC4 PE=1 S
6562.10	3468	963	4430	A0A0G2JQN9|A0A0G2JQN9_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6553.32	3520	963	4482	A0A0G2JRK4|A0A0G2JRK4_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6553.32	3410	963	4372	A0A0G2JRW3|A0A0G2JRW3_HUMAN Mucin-4 (Fragment) OS=Homo sapiens OX=
6366.10	1339	2223	3561	P98088|MUC5A_HUMAN Mucin-5AC OS=Homo sapiens OX=9606 GN=MUC5AC PE=
6226.32	2259	132	2390	Q5D862|FILA2_HUMAN Filaggrin-2 OS=Homo sapiens OX=9606 GN=FLG2 PE=
5454.25	3764	297	4060	P20930|FILA_HUMAN Filaggrin OS=Homo sapiens OX=9606 GN=FLG PE=1 SV
4726.82	689	4233	4921	Q9HC84|MUC5B_HUMAN Mucin-5B OS=Homo sapiens OX=9606 GN=MUC5B PE=1
4219.06	477	2087	2563	Q02505|MUC3A_HUMAN Mucin-3A OS=Homo sapiens OX=9606 GN=MUC3A PE=1
4033.40	857	442	1298	Q685J3|MUC17_HUMAN Mucin-17 OS=Homo sapiens OX=9606 GN=MUC17 PE=1
4032.16	857	442	1298	E7EPM4|E7EPM4_HUMAN Mucin-17 OS=Homo sapiens OX=9606 GN=MUC17 PE=1
2867.01	1659	1254	2912	Q02388|CO7A1_HUMAN Collagen alpha-1(VII) chain OS=Homo sapiens OX=
2854.13	1531	54	1584	P02462|CO4A1_HUMAN Collagen alpha-1(IV) chain OS=Homo sapiens OX=9
2815.86	1565	33	1597	P29400|CO4A5_HUMAN Collagen alpha-5(IV) chain OS=Homo sapiens OX=9
2778.12	30	4145	4174	P08519|APOA_HUMAN Apolipoprotein(a) OS=Homo sapiens OX=9606 GN=LPA
2585.77	596	933	1528	A8MXH5|A8MXH5_HUMAN Collagen alpha-6(IV) chain OS=Homo sapiens OX=
2562.66	596	917	1512	Q14031|CO4A6_HUMAN Collagen alpha-6(IV) chain OS=Homo sapiens OX=9
2532.07	582	916	1497	F5H851|F5H851_HUMAN Collagen alpha-6(IV) chain OS=Homo sapiens OX=
2521.38	1215	337	1551	P53420|CO4A4_HUMAN Collagen alpha-4(IV) chain OS=Homo sapiens OX=9
2514.09	581	916	1496	A0A087WZY5|A0A087WZY5_HUMAN Collagen alpha-6(IV) chain OS=Homo sap
2504.77	1256	302	1557	P08572|CO4A2_HUMAN Collagen alpha-2(IV) chain OS=Homo sapiens OX=9
2457.82	544	916	1459	F5H3Q5|F5H3Q5_HUMAN Collagen alpha-6(IV) chain OS=Homo sapiens OX=

2405.02	1265	91	1355	P02461|CO3A1_HUMAN Collagen alpha-1(III) chain OS=Homo sapiens OX=
2372.37	1056	492	1547	Q01955|CO4A3_HUMAN Collagen alpha-3(IV) chain OS=Homo sapiens OX=9
2262.45	244	435	678	A0A1B0GU24|A0A1B0GU24_HUMAN Trinucleotide repeat-containing gene 6
2231.96	325	391	715	Q8NDV7|TNR6A_HUMAN Trinucleotide repeat-containing gene 6A protein
2231.54	1302	92	1393	P05997|CO5A2_HUMAN Collagen alpha-2(V) chain OS=Homo sapiens OX=96
2194.23	712	355	1066	Q8N7X1|RMXL3_HUMAN RNA-binding motif protein, X-linked-like-3 OS=H
2129.73	1220	28	1247	P08123|CO1A2_HUMAN Collagen alpha-2(I) chain OS=Homo sapiens OX=96
2126.77	1214	31	1244	A0A087WTA8|A0A087WTA8_HUMAN Collagen alpha-2(I) chain OS=Homo sapi
2068.48	426	212	637	Q9UPQ9|TNR6B_HUMAN Trinucleotide repeat-containing gene 6B protein
2064.13	243	1188	1430	Q12816|TROP_HUMAN Trophinin OS=Homo sapiens OX=9606 GN=TRO PE=1 SV
2023.20	380	1309	1688	A0A6Q8NVI4|A0A6Q8NVI4_HUMAN AT-rich interactive domain-containing
2009.55	244	225	468	Q9HCJ0|TNR6C_HUMAN Trinucleotide repeat-containing gene 6C protein
1956.85	1232	106	1337	P02452|CO1A1_HUMAN Collagen alpha-1(I) chain OS=Homo sapiens OX=96
1956.80	1263	93	1355	P02458|CO2A1_HUMAN Collagen alpha-1(II) chain OS=Homo sapiens OX=9
1921.32	531	564	1094	Q9UMD9|COHA1_HUMAN Collagen alpha-1(XVII) chain OS=Homo sapiens OX
1910.96	593	1011	1603	Q07092|COGA1_HUMAN Collagen alpha-1(XVI) chain OS=Homo sapiens OX=
1909.43	380	1269	1648	A0A3F2YNW7|A0A3F2YNW7_HUMAN AT-rich interactive domain-containing
1890.60	1125	490	1614	A0A0G2JL35|A0A0G2JL35_HUMAN COL11A2 OS=Homo sapiens OX=9606 GN=COL
1890.43	1125	490	1614	A0A140TA43|A0A140TA43_HUMAN COL11A2 OS=Homo sapiens OX=9606 GN=COL
1884.98	1125	490	1614	P13942|COBA2_HUMAN Collagen alpha-2(XI) chain OS=Homo sapiens OX=9
1884.14	1125	490	1614	A0A0C4DFS1|A0A0C4DFS1_HUMAN COL11A2 OS=Homo sapiens OX=9606 GN=COL
1880.29	1125	377	1501	A0A140T9I7|A0A140T9I7_HUMAN Collagen alpha-2(XI) chain (Fragment)
1880.29	1125	404	1528	Q4VXY6|Q4VXY6_HUMAN Collagen alpha-2(XI) chain OS=Homo sapiens OX=
1880.12	1125	404	1528	A0A140T9N1|A0A140T9N1_HUMAN Collagen alpha-2(XI) chain OS=Homo sap
1879.97	1125	383	1507	H0YIS1|H0YIS1_HUMAN Collagen alpha-2(XI) chain OS=Homo sapiens OX=
1879.80	1125	383	1507	A0A140TA54|A0A140TA54_HUMAN Collagen alpha-2(XI) chain OS=Homo sap
1877.32	380	1173	1552	Q8NFD5|ARI1B_HUMAN AT-rich interactive domain-containing protein 1
1872.94	405	595	999	O14497|ARI1A_HUMAN AT-rich interactive domain-containing protein 1
1830.52	338	1752	2089	P35658|NU214_HUMAN Nuclear pore complex protein Nup214 OS=Homo sap
1827.89	338	1740	2077	A0A494C1F2|A0A494C1F2_HUMAN Nuclear pore complex protein Nup214 OS
1813.71	311	1	311	P23490|LORI_HUMAN Loricrin OS=Homo sapiens OX=9606 GN=LORICRIN PE=
1812.15	1105	476	1580	P25940|CO5A3_HUMAN Collagen alpha-3(V) chain OS=Homo sapiens OX=96
1809.72	1103	562	1664	P20908|CO5A1_HUMAN Collagen alpha-1(V) chain OS=Homo sapiens OX=96

1798.87	1127	532	1658	P12107|COBA1_HUMAN Collagen alpha-1(XI) chain OS=Homo sapiens OX=9
1789.69	1143	483	1625	Q8NFW1|COMA1_HUMAN Collagen alpha-1(XXII) chain OS=Homo sapiens OX
1777.68	554	1311	1864	A0A0G2JN42|A0A0G2JN42_HUMAN Mucin-6 OS=Homo sapiens OX=9606 GN=MUC
1776.98	554	1311	1864	Q6W4X9|MUC6_HUMAN Mucin-6 OS=Homo sapiens OX=9606 GN=MUC6 PE=1 SV=
1768.42	266	326	591	Q92804|RBP56_HUMAN TATA-binding protein-associated factor 2N OS=Ho
1767.36	213	1650	1862	A0A0G2JNJ8|A0A0G2JNJ8_HUMAN Mucin-6 OS=Homo sapiens OX=9606 GN=MUC
1705.57	338	1181	1518	A0A0A0MSW3|A0A0A0MSW3_HUMAN Nuclear pore complex protein Nup214 OS
1666.66	963	1	963	A0A087WYX9|A0A087WYX9_HUMAN Collagen alpha-2(V) chain OS=Homo sapi
1665.08	1071	489	1559	Q17RW2|COOA1_HUMAN Collagen alpha-1(XXIV) chain OS=Homo sapiens OX
1662.48	606	29	634	E2RYF6|MUC22_HUMAN Mucin-22 OS=Homo sapiens OX=9606 GN=MUC22 PE=1
1657.27	171	1	171	P35527|K1C9_HUMAN Keratin, type I cytoskeletal 9 OS=Homo sapiens O
1657.11	375	1	375	H0Y720|H0Y720_HUMAN Trinucleotide repeat-containing gene 6B protei
1593.26	314	1	314	P35637|FUS_HUMAN RNA-binding protein FUS OS=Homo sapiens OX=9606 G
1588.95	313	1	313	H3BPE7|H3BPE7_HUMAN RNA-binding protein FUS OS=Homo sapiens OX=960
1571.71	338	578	915	B7ZAV2|B7ZAV2_HUMAN Nuclear pore complex protein Nup214 OS=Homo sa
1566.62	163	1	163	P13645|K1C10_HUMAN Keratin, type I cytoskeletal 10 OS=Homo sapiens
1517.96	385	512	896	A0A0U1RQI7|KLF18_HUMAN Kruppel-like factor 18 OS=Homo sapiens OX=9
1506.20	787	625	1411	Q8IZC6|CORA1_HUMAN Collagen alpha-1(XXVII) chain OS=Homo sapiens O
1431.95	157	1	157	P04264|K2C1_HUMAN Keratin, type II cytoskeletal 1 OS=Homo sapiens
1401.00	94	1455	1548	Q9UPA5|BSN_HUMAN Protein bassoon OS=Homo sapiens OX=9606 GN=BSN PE
1390.15	290	605	894	H0Y837|H0Y837_HUMAN Nuclear pore complex protein Nup214 (Fragment)
1374.30	130	2108	2237	Q8NEZ4|KMT2C_HUMAN Histone-lysine N-methyltransferase 2C OS=Homo s
1352.52	291	1184	1474	P49790|NU153_HUMAN Nuclear pore complex protein Nup153 OS=Homo sap
1347.59	274	467	740	A0A0G2JNL3|A0A0G2JNL3_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
1343.38	581	27	607	A0A140T8X8|A0A140T8X8_HUMAN Mucin-21 OS=Homo sapiens OX=9606 GN=MU
1342.89	405	214	618	A0A1B0GTU5|A0A1B0GTU5_HUMAN AT-rich interactive domain-containing
1342.05	274	467	740	A0A0G2JPA4|A0A0G2JPA4_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
1339.93	405	212	616	H0Y488|H0Y488_HUMAN AT-rich interactive domain-containing protein
1339.80	207	137	343	Q99102|MUC4_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC4 PE=1 SV=
1323.56	475	1495	1969	P24928|RPB1_HUMAN DNA-directed RNA polymerase II subunit RPB1 OS=H
1317.29	791	510	1300	Q9NZW4|DSPP_HUMAN Dentin sialophosphoprotein OS=Homo sapiens OX=96
1298.72	193	1	193	P52948|NUP98_HUMAN Nuclear pore complex protein Nup98-Nup96 OS=Hom
1293.42	119	2552	2670	O14686|KMT2D_HUMAN Histone-lysine N-methyltransferase 2D OS=Homo s

1281.04	606	1397	2002	A0A0G2JR65|A0A0G2JR65_HUMAN Mucin-2 OS=Homo sapiens OX=9606 GN=MUC
1279.85	626	69	694	Q49AM6|Q49AM6_HUMAN COL4A5 protein OS=Homo sapiens OX=9606 GN=COL4
1273.45	210	1512	1721	Q5H9R4|ARMX4_HUMAN Armadillo repeat-containing X-linked protein 4
1264.84	404	212	615	A0A087WUV6|A0A087WUV6_HUMAN AT-rich interactive domain-containing
1264.40	627	417	1043	Q14993|COJA1_HUMAN Collagen alpha-1(XIX) chain OS=Homo sapiens OX=
1263.12	755	728	1482	P39060|COIA1_HUMAN Collagen alpha-1(XVIII) chain OS=Homo sapiens O
1259.34	547	27	573	A0A0G2JKD1|A0A0G2JKD1_HUMAN Mucin-21 OS=Homo sapiens OX=9606 GN=MU
1243.81	113	1	113	Q03164|KMT2A_HUMAN Histone-lysine N-methyltransferase 2A OS=Homo s
1238.63	193	1	193	A0A3B3ITD8|A0A3B3ITD8_HUMAN Nuclear pore complex protein Nup98-Nup
1238.18	273	39	311	Q15517|CDSN_HUMAN Corneodesmosin OS=Homo sapiens OX=9606 GN=CDSN P
1231.60	157	1	157	P35908|K22E_HUMAN Keratin, type II cytoskeletal 2 epidermal OS=Hom
1231.48	477	26	502	A0A182DWF7|A0A182DWF7_HUMAN Mucin-3A (Fragment) OS=Homo sapiens OX
1225.83	273	39	311	G8JLG2|G8JLG2_HUMAN Corneodesmosin OS=Homo sapiens OX=9606 GN=CDSN
1223.35	380	695	1074	H0Y7H8|H0Y7H8_HUMAN AT-rich interactive domain-containing protein
1223.15	273	39	311	Q2L6G8|Q2L6G8_HUMAN Corneodesmosin OS=Homo sapiens OX=9606 GN=CDSN
1209.74	114	642	755	Q9UGU0|TCF20_HUMAN Transcription factor 20 OS=Homo sapiens OX=9606
1199.54	82	2246	2327	Q96JG9|ZN469_HUMAN Zinc finger protein 469 OS=Homo sapiens OX=9606
1192.65	82	2274	2355	H3BS19|H3BS19_HUMAN Zinc finger protein 469 OS=Homo sapiens OX=960
1185.46	485	27	511	A0A0G2JJF7|A0A0G2JJF7_HUMAN Mucin-21 OS=Homo sapiens OX=9606 GN=MU
1180.20	485	27	511	Q5SSG8|MUC21_HUMAN Mucin-21 OS=Homo sapiens OX=9606 GN=MUC21 PE=1
1178.88	487	27	513	A0A0G2JHX4|A0A0G2JHX4_HUMAN Mucin-21 OS=Homo sapiens OX=9606 GN=MU
1177.93	109	597	705	A0A0A0MTL4|A0A0A0MTL4_HUMAN Neuron navigator 2 OS=Homo sapiens OX=
1164.24	487	27	513	A0A140TA38|A0A140TA38_HUMAN Mucin-21 OS=Homo sapiens OX=9606 GN=MU
1160.96	109	620	728	Q8IVL1|NAV2_HUMAN Neuron navigator 2 OS=Homo sapiens OX=9606 GN=NA
1160.96	109	620	728	A0A0A0MTE8|A0A0A0MTE8_HUMAN Neuron navigator 2 OS=Homo sapiens OX=
1141.13	374	1	374	Q01844|EWS_HUMAN RNA-binding protein EWS OS=Homo sapiens OX=9606 G
1132.79	157	863	1019	Q10571|MN1_HUMAN Transcriptional activator MN1 OS=Homo sapiens OX=
1131.54	61	388	448	A0A0J9YXN7|A0A0J9YXN7_HUMAN Perilipin-4 OS=Homo sapiens OX=9606 GN
1129.72	337	100	436	A0A494C0Y1|A0A494C0Y1_HUMAN Nuclear pore complex protein Nup214 (F
1127.65	61	373	433	Q96Q06|PLIN4_HUMAN Perilipin-4 OS=Homo sapiens OX=9606 GN=PLIN4 PE
1102.34	84	1157	1240	Q9Y566|SHAN1_HUMAN SH3 and multiple ankyrin repeat domains protein
1102.34	84	1165	1248	H9KV90|H9KV90_HUMAN SH3 and multiple ankyrin repeat domains protei
1099.42	109	519	627	P12035|K2C3_HUMAN Keratin, type II cytoskeletal 3 OS=Homo sapiens

1098.30	114	642	755	A0A6Q8PH68|A0A6Q8PH68_HUMAN Transcription factor 20 (Fragment) OS=
1097.91	70	1571	1640	Q68DE3|USF3_HUMAN Basic helix-loop-helix domain-containing protein
1081.47	352	1	352	B0QYK0|B0QYK0_HUMAN RNA-binding protein EWS OS=Homo sapiens OX=960
1072.95	1034	93	1126	A0A087WWM1|A0A087WWM1_HUMAN Mucin-1 OS=Homo sapiens OX=9606 GN=MUC
1067.22	262	1	262	H3BNZ4|H3BNZ4_HUMAN RNA-binding protein FUS OS=Homo sapiens OX=960
1060.10	64	3492	3555	A2VEC9|SSPO_HUMAN SCO-spondin OS=Homo sapiens OX=9606 GN=SSPOP PE=
1058.36	341	1495	1835	A0A6Q8PGB0|A0A6Q8PGB0_HUMAN DNA-directed RNA polymerase subunit OS
1056.51	140	1073	1212	Q15648|MED1_HUMAN Mediator of RNA polymerase II transcription subu
1047.86	1039	93	1131	P15941|MUC1_HUMAN Mucin-1 OS=Homo sapiens OX=9606 GN=MUC1 PE=1 SV=
1044.87	130	366	495	Q9Y6Q9|NCOA3_HUMAN Nuclear receptor coactivator 3 OS=Homo sapiens
1041.03	120	518	637	Q01546|K22O_HUMAN Keratin, type II cytoskeletal 2 oral OS=Homo sap
1040.79	662	27	688	Q14055|CO9A2_HUMAN Collagen alpha-2(IX) chain OS=Homo sapiens OX=9
1040.11	304	462	765	P20849|CO9A1_HUMAN Collagen alpha-1(IX) chain OS=Homo sapiens OX=9
1034.45	115	2011	2125	O75179|ANR17_HUMAN Ankyrin repeat domain-containing protein 17 OS=
1034.23	242	133	374	Q6E0U4|DMKN_HUMAN Dermokine OS=Homo sapiens OX=9606 GN=DMKN PE=1 S
1030.70	652	32	683	Q14050|CO9A3_HUMAN Collagen alpha-3(IX) chain OS=Homo sapiens OX=9
1024.56	284	529	812	Q14157|UBP2L_HUMAN Ubiquitin-associated protein 2-like OS=Homo sap
1016.98	153	189	341	Q92793|CBP_HUMAN CREB-binding protein OS=Homo sapiens OX=9606 GN=C
1014.49	67	4936	5002	Q9Y6V0|PCLO_HUMAN Protein piccolo OS=Homo sapiens OX=9606 GN=PCLO
1012.35	244	426	669	Q14686|NCOA6_HUMAN Nuclear receptor coactivator 6 OS=Homo sapiens
1002.13	64	4088	4151	Q2LD37|K1109_HUMAN Transmembrane protein KIAA1109 OS=Homo sapiens
999.38	174	831	1004	Q86UU0|BCL9L_HUMAN B-cell CLL/lymphoma 9-like protein OS=Homo sapi
999.37	190	1039	1228	O00512|BCL9_HUMAN B-cell CLL/lymphoma 9 protein OS=Homo sapiens OX
996.58	174	794	967	A0A087WZX0|A0A087WZX0_HUMAN B-cell CLL/lymphoma 9-like protein OS=
993.55	70	1034	1103	Q8IVL0|NAV3_HUMAN Neuron navigator 3 OS=Homo sapiens OX=9606 GN=NA
986.36	284	540	823	F8W726|F8W726_HUMAN Ubiquitin-associated protein 2-like OS=Homo sa
985.36	278	616	893	Q12906|ILF3_HUMAN Interleukin enhancer-binding factor 3 OS=Homo sa
976.33	115	1048	1162	Q9UQ35|SRRM2_HUMAN Serine/arginine repetitive matrix protein 2 OS=
975.22	319	1	319	C9JGE3|C9JGE3_HUMAN EWS RNA-binding protein variant 6 OS=Homo sapi
973.33	184	1038	1221	A8CG34|P121C_HUMAN Nuclear envelope pore membrane protein POM 121C
971.05	182	2733	2914	Q99715|COCA1_HUMAN Collagen alpha-1(XII) chain OS=Homo sapiens OX=
958.46	177	2733	2909	D6RGG3|D6RGG3_HUMAN Collagen alpha-1(XII) chain OS=Homo sapiens OX
958.16	459	245	703	Q6XPR3|RPTN_HUMAN Repetin OS=Homo sapiens OX=9606 GN=RPTN PE=1 SV=

952.77	54	2383	2436	Q9P2P6|STAR9_HUMAN StAR-related lipid transfer protein 9 OS=Homo s
951.29	102	656	757	P35568|IRS1_HUMAN Insulin receptor substrate 1 OS=Homo sapiens OX=
946.89	106	883	988	A0A2R8Y4T1|A0A2R8Y4T1_HUMAN Tensin-1 OS=Homo sapiens OX=9606 GN=TN
946.05	200	607	806	Q5T6F2|UBAP2_HUMAN Ubiquitin-associated protein 2 OS=Homo sapiens
945.89	106	837	942	A0A494C067|A0A494C067_HUMAN Tensin-1 (Fragment) OS=Homo sapiens OX
932.70	225	77	301	Q09472|EP300_HUMAN Histone acetyltransferase p300 OS=Homo sapiens
931.41	72	1629	1700	Q15911|ZFHX3_HUMAN Zinc finger homeobox protein 3 OS=Homo sapiens
930.65	317	623	939	Q96QC0|PP1RA_HUMAN Serine/threonine-protein phosphatase 1 regulato
930.13	380	476	855	A0A1B0GVK1|A0A1B0GVK1_HUMAN AT-rich interactive domain-containing
930.02	130	1815	1944	Q2M2H8|MGAL_HUMAN Probable maltase-glucoamylase 2 OS=Homo sapiens
929.68	293	1	293	A0A0D9SFL3|A0A0D9SFL3_HUMAN RNA-binding protein EWS OS=Homo sapien
929.14	238	20	257	Q17RH7|TPRXL_HUMAN Putative protein TPRXL OS=Homo sapiens OX=9606
923.64	54	1238	1291	P98160|PGBM_HUMAN Basement membrane-specific heparan sulfate prote
918.36	608	1	608	A0A0G2JNG3|A0A0G2JNG3_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
918.13	580	245	824	Q2UY09|COSA1_HUMAN Collagen alpha-1(XXVIII) chain OS=Homo sapiens
917.35	569	35	603	A0A0G2JLU8|A0A0G2JLU8_HUMAN Mucin-4 OS=Homo sapiens OX=9606 GN=MUC
911.60	225	77	301	A0A669KB12|A0A669KB12_HUMAN Histone acetyltransferase OS=Homo sapi
909.19	158	232	389	K7EQQ3|K7EQQ3_HUMAN Keratin, type I cytoskeletal 9 OS=Homo sapiens
906.71	33	2560	2592	O60494|CUBN_HUMAN Cubilin OS=Homo sapiens OX=9606 GN=CUBN PE=1 SV=
905.15	106	737	842	E9PGF5|E9PGF5_HUMAN Tensin-1 OS=Homo sapiens OX=9606 GN=TNS1 PE=1
904.25	62	1430	1491	I3L2J0|I3L2J0_HUMAN Protein capicua homolog OS=Homo sapiens OX=960
904.15	106	737	842	Q9HBL0|TENS1_HUMAN Tensin-1 OS=Homo sapiens OX=9606 GN=TNS1 PE=1 S
904.14	106	737	842	E9PF55|E9PF55_HUMAN Tensin-1 OS=Homo sapiens OX=9606 GN=TNS1 PE=1
903.24	55	194	248	P48634|PRC2A_HUMAN Protein PRRC2A OS=Homo sapiens OX=9606 GN=PRRC2
900.93	63	5827	5889	Q09666|AHNK_HUMAN Neuroblast differentiation-associated protein AH
900.11	106	388	493	A0A087WWW7|A0A087WWW7_HUMAN Tensin-1 OS=Homo sapiens OX=9606 GN=TN
896.87	287	195	481	A0A6I8PTU7|A0A6I8PTU7_HUMAN AT-rich interactive domain-containing
895.36	132	653	784	Q9H4A3|WNK1_HUMAN Serine/threonine-protein kinase WNK1 OS=Homo sap
892.75	119	2323	2441	P25054|APC_HUMAN Adenomatous polyposis coli protein OS=Homo sapien
887.57	115	1895	2009	H0YM23|H0YM23_HUMAN Ankyrin repeat domain-containing protein 17 (F
885.26	178	36	213	A0A075B7F4|A0A075B7F4_HUMAN TATA-binding protein-associated factor
880.06	129	432	560	Q9Y4H2|IRS2_HUMAN Insulin receptor substrate 2 OS=Homo sapiens OX=
878.19	182	1544	1725	A0A087X0A8|A0A087X0A8_HUMAN Collagen alpha-1(XII) chain OS=Homo sa

875.82	157	1059	1215	Q96HA1|P121A_HUMAN Nuclear envelope pore membrane protein POM 121
871.82	296	650	945	A6NCT7|A6NCT7_HUMAN Collagen alpha-1(XVI) chain OS=Homo sapiens OX
869.26	210	422	631	Q9ULL5|PRR12_HUMAN Proline-rich protein 12 OS=Homo sapiens OX=9606
868.87	31	34	64	A0A3B3ISX9|A0A3B3ISX9_HUMAN Tenascin-X OS=Homo sapiens OX=9606 GN=
865.35	539	55	593	Q03692|COAA1_HUMAN Collagen alpha-1(X) chain OS=Homo sapiens OX=96
861.45	61	2132	2192	Q5JSZ5|PRC2B_HUMAN Protein PRRC2B OS=Homo sapiens OX=9606 GN=PRRC2
861.04	181	191	371	P09651|ROA1_HUMAN Heterogeneous nuclear ribonucleoprotein A1 OS=Ho
855.43	521	80	600	P25067|CO8A2_HUMAN Collagen alpha-2(VIII) chain OS=Homo sapiens OX
854.14	92	910	1001	Q5VT52|RPRD2_HUMAN Regulation of nuclear pre-mRNA domain-containin
850.48	516	578	1093	Q9C0J8|WDR33_HUMAN pre-mRNA 3' end processing protein WDR33 OS=Hom
848.95	382	1397	1778	A0A0G2JM87|A0A0G2JM87_HUMAN Mucin-2 OS=Homo sapiens OX=9606 GN=MUC
841.89	145	653	797	F5GWT4|F5GWT4_HUMAN Non-specific serine/threonine protein kinase O
839.73	35	917	951	A0A140T902|A0A140T902_HUMAN Tenascin-X OS=Homo sapiens OX=9606 GN=
838.59	35	917	951	A0A140T9C0|A0A140T9C0_HUMAN Tenascin-X OS=Homo sapiens OX=9606 GN=
836.87	521	15	535	E9PP49|E9PP49_HUMAN Collagen alpha-2(VIII) chain OS=Homo sapiens O
833.67	380	432	811	A0A1B0GTJ8|A0A1B0GTJ8_HUMAN AT-rich interactive domain-containing
833.45	37	2223	2259	Q9Y6R7|FCGBP_HUMAN IgGFc-binding protein OS=Homo sapiens OX=9606 G
833.44	98	492	589	P13647|K2C5_HUMAN Keratin, type II cytoskeletal 5 OS=Homo sapiens
831.79	176	234	409	Q99081|HTF4_HUMAN Transcription factor 12 OS=Homo sapiens OX=9606
828.84	510	436	945	Q96P44|COLA1_HUMAN Collagen alpha-1(XXI) chain OS=Homo sapiens OX=
827.75	35	917	951	A0A140T8Y3|A0A140T8Y3_HUMAN Tenascin-X OS=Homo sapiens OX=9606 GN=
824.81	219	1010	1228	A0A2R8Y651|A0A2R8Y651_HUMAN PDZ domain-containing protein GIPC3 OS
824.62	30	843	872	Q8TCU4|ALMS1_HUMAN Alstrom syndrome protein 1 OS=Homo sapiens OX=9
824.07	141	39	179	A0A1B0GVR6|A0A1B0GVR6_HUMAN Transcription factor 4 OS=Homo sapiens
821.72	169	209	377	P51991|ROA3_HUMAN Heterogeneous nuclear ribonucleoprotein A3 OS=Ho
821.72	101	475	575	Q9BVL2|NUP58_HUMAN Nucleoporin p58/p45 OS=Homo sapiens OX=9606 GN=
821.66	212	387	598	Q15596|NCOA2_HUMAN Nuclear receptor coactivator 2 OS=Homo sapiens
820.61	205	1143	1347	Q15788|NCOA1_HUMAN Nuclear receptor coactivator 1 OS=Homo sapiens
820.52	35	917	951	A0A140TA33|A0A140TA33_HUMAN Tenascin-X OS=Homo sapiens OX=9606 GN=
820.51	141	131	271	E9PH57|E9PH57_HUMAN Transcription factor 4 OS=Homo sapiens OX=9606
820.08	505	439	943	F5GZK2|F5GZK2_HUMAN Collagen alpha-1(XXI) chain OS=Homo sapiens OX
819.38	35	917	951	A0A140TA41|A0A140TA41_HUMAN Tenascin-X OS=Homo sapiens OX=9606 GN=
817.56	141	29	169	P15884|ITF2_HUMAN Transcription factor 4 OS=Homo sapiens OX=9606 G

815.38	35	917	951	P22105|TENX_HUMAN Tenascin-X OS=Homo sapiens OX=9606 GN=TNXB PE=1
812.92	30	801	830	A0A087WTU9|A0A087WTU9_HUMAN Alstrom syndrome protein 1 OS=Homo sap
812.30	141	29	169	H3BTP3|H3BTP3_HUMAN Transcription factor 4 OS=Homo sapiens OX=9606
809.01	41	3038	3078	Q7Z407|CSMD3_HUMAN CUB and sushi domain-containing protein 3 OS=Ho
808.53	35	917	951	A0A140TA52|A0A140TA52_HUMAN Tenascin-X OS=Homo sapiens OX=9606 GN=
807.15	110	416	525	A0A2R8YDL9|A0A2R8YDL9_HUMAN Methyl-CpG-binding domain protein 5 OS
801.91	145	1	145	H3BPJ7|H3BPJ7_HUMAN Transcription factor 4 OS=Homo sapiens OX=9606
799.98	380	416	795	A0A1B0GWJ2|A0A1B0GWJ2_HUMAN AT-rich interactive domain-containing
798.99	127	2560	2686	O15417|TNC18_HUMAN Trinucleotide repeat-containing gene 18 protein
798.99	127	2560	2686	H9KVB4|H9KVB4_HUMAN Trinucleotide repeat-containing gene 18 protei
791.12	237	126	362	Q96F45|ZN503_HUMAN Zinc finger protein 503 OS=Homo sapiens OX=9606
791.12	93	165	257	O15027|SC16A_HUMAN Protein transport protein Sec16A OS=Homo sapien
789.84	149	579	727	A0A2R8YGI3|A0A2R8YGI3_HUMAN Collagen alpha-1(XIII) chain OS=Homo s
788.40	323	27	349	A0A0G2JMC4|A0A0G2JMC4_HUMAN Mucin-21 OS=Homo sapiens OX=9606 GN=MU
787.13	142	483	624	O14654|IRS4_HUMAN Insulin receptor substrate 4 OS=Homo sapiens OX=
786.68	141	29	169	A0A1B0GVB8|A0A1B0GVB8_HUMAN Transcription factor 4 OS=Homo sapiens
784.58	216	465	680	A0A6E1W314|A0A6E1W314_HUMAN Collagen alpha-1(XIII) chain OS=Homo s
784.23	205	992	1196	B5MCN7|B5MCN7_HUMAN Nuclear receptor coactivator 1 OS=Homo sapiens
783.69	149	568	716	Q5TAT6|CODA1_HUMAN Collagen alpha-1(XIII) chain OS=Homo sapiens OX
779.79	30	843	872	A0A087WV20|A0A087WV20_HUMAN Alstrom syndrome protein 1 OS=Homo sap
779.07	144	61	204	Q9UI36|DACH1_HUMAN Dachshund homolog 1 OS=Homo sapiens OX=9606 GN=
776.51	321	27	347	A0A140TA51|A0A140TA51_HUMAN Mucin-21 OS=Homo sapiens OX=9606 GN=MU
775.34	161	643	803	A6NF01|P121B_HUMAN Putative nuclear envelope pore membrane protein
774.85	124	713	836	Q5SYE7|NHSL1_HUMAN NHS-like protein 1 OS=Homo sapiens OX=9606 GN=N
772.87	67	27	93	Q8IWZ3|ANKH1_HUMAN Ankyrin repeat and KH domain-containing protein
771.21	47	2226	2272	O15018|PDZD2_HUMAN PDZ domain-containing protein 2 OS=Homo sapiens
771.17	188	1273	1460	E7EWN3|E7EWN3_HUMAN Histone-lysine N-methyltransferase SETD5 OS=Ho
770.13	216	383	598	A0A669KB55|A0A669KB55_HUMAN Collagen alpha-1(XIII) chain (Fragment
768.26	107	907	1013	Q8IZL2|MAML2_HUMAN Mastermind-like protein 2 OS=Homo sapiens OX=96
768.25	516	120	635	P27658|CO8A1_HUMAN Collagen alpha-1(VIII) chain OS=Homo sapiens OX
764.83	202	672	873	Q92585|MAML1_HUMAN Mastermind-like protein 1 OS=Homo sapiens OX=96
764.69	70	534	603	A0A2R8YFX5|A0A2R8YFX5_HUMAN Neuron navigator 3 OS=Homo sapiens OX=
764.55	137	327	463	P55197|AF10_HUMAN Protein AF-10 OS=Homo sapiens OX=9606 GN=MLLT10

764.39	54	352	405	Q9H195|MUC3B_HUMAN Mucin-3B (Fragments) OS=Homo sapiens OX=9606 GN
761.93	157	196	352	P22626|ROA2_HUMAN Heterogeneous nuclear ribonucleoproteins A2/B1 O
761.72	91	46	136	Q2KJY2|KI26B_HUMAN Kinesin-like protein KIF26B OS=Homo sapiens OX=
758.09	285	2119	2403	P12111|CO6A3_HUMAN Collagen alpha-3(VI) chain OS=Homo sapiens OX=9
757.78	156	716	871	Q9NTZ6|RBM12_HUMAN RNA-binding protein 12 OS=Homo sapiens OX=9606
755.58	126	75	200	Q6L8H1|KRA54_HUMAN Keratin-associated protein 5-4 OS=Homo sapiens
752.27	93	165	257	F1T0I1|F1T0I1_HUMAN Protein transport protein sec16 OS=Homo sapien
748.32	194	781	974	O94913|PCF11_HUMAN Pre-mRNA cleavage complex 2 protein Pcf11 OS=Ho
744.20	149	1691	1839	P16112|PGCA_HUMAN Aggrecan core protein OS=Homo sapiens OX=9606 GN
744.20	149	1691	1839	H0YMF1|H0YMF1_HUMAN Aggrecan core protein OS=Homo sapiens OX=9606
744.18	149	1672	1820	A0A087X1T7|A0A087X1T7_HUMAN Aggrecan core protein OS=Homo sapiens
743.54	155	1	155	Q2M2I5|K1C24_HUMAN Keratin, type I cytoskeletal 24 OS=Homo sapiens
738.90	188	1254	1441	Q9C0A6|SETD5_HUMAN Histone-lysine N-methyltransferase SETD5 OS=Hom
738.12	33	325	357	A0A087X0K4|A0A087X0K4_HUMAN CUB and sushi domain-containing protei
736.93	126	1718	1843	Q8IZD2|KMT2E_HUMAN Inactive histone-lysine N-methyltransferase 2E
736.65	203	408	610	A0A669KB28|A0A669KB28_HUMAN Collagen alpha-1(XIII) chain (Fragment
736.00	110	416	525	Q9P267|MBD5_HUMAN Methyl-CpG-binding domain protein 5 OS=Homo sapi
736.00	110	416	525	A0A1B0GW10|A0A1B0GW10_HUMAN Methyl-CpG-binding domain protein 5 OS
734.22	167	229	395	P15923|TFE2_HUMAN Transcription factor E2-alpha OS=Homo sapiens OX
732.11	73	88	160	A0A2R8Y5P9|A0A2R8Y5P9_HUMAN Protein Shroom3 OS=Homo sapiens OX=960
731.60	171	1455	1625	Q05707|COEA1_HUMAN Collagen alpha-1(XIV) chain OS=Homo sapiens OX=
730.82	463	182	644	A8MWQ5|A8MWQ5_HUMAN Collagen alpha-1(XXV) chain OS=Homo sapiens OX
730.29	111	1287	1397	A0A6Q8PFM0|A0A6Q8PFM0_HUMAN Serine/threonine-protein kinase WNK1 (
729.57	149	1215	1363	A0A5K1VW97|A0A5K1VW97_HUMAN Aggrecan core protein (Fragment) OS=Ho
728.34	474	196	669	A0A2R8Y760|A0A2R8Y760_HUMAN Collagen alpha-1(XXV) chain OS=Homo sa
727.41	174	230	403	B4DGI9|B4DGI9_HUMAN Transcription factor 12 (Fragment) OS=Homo sap
727.02	107	175	281	Q3L8U1|CHD9_HUMAN Chromodomain-helicase-DNA-binding protein 9 OS=H
725.68	115	815	929	Q9H2D6|TARA_HUMAN TRIO and F-actin-binding protein OS=Homo sapiens
725.42	285	1512	1796	E7ENL6|E7ENL6_HUMAN Collagen alpha-3(VI) chain OS=Homo sapiens OX=
725.23	95	360	454	Q96JK9|MAML3_HUMAN Mastermind-like protein 3 OS=Homo sapiens OX=96
725.18	62	733	794	Q9C0C2|TB182_HUMAN 182 kDa tankyrase-1-binding protein OS=Homo sap
718.21	145	401	545	Q9ULJ6|ZMIZ1_HUMAN Zinc finger MIZ domain-containing protein 1 OS=
717.59	94	484	577	Q7Z794|K2C1B_HUMAN Keratin, type II cytoskeletal 1b OS=Homo sapien

717.14	76	1145	1220	O43166|SI1L1_HUMAN Signal-induced proliferation-associated 1-like
715.72	125	365	489	P54259|ATN1_HUMAN Atrophin-1 OS=Homo sapiens OX=9606 GN=ATN1 PE=1
714.37	73	169	241	Q8TF72|SHRM3_HUMAN Protein Shroom3 OS=Homo sapiens OX=9606 GN=SHRO
711.74	71	944	1014	Q8IZF6|AGRG4_HUMAN Adhesion G-protein coupled receptor G4 OS=Homo
710.77	193	325	517	Q8NCA5|FA98A_HUMAN Protein FAM98A OS=Homo sapiens OX=9606 GN=FAM98
710.25	49	2698	2746	Q15751|HERC1_HUMAN Probable E3 ubiquitin-protein ligase HERC1 OS=H
707.49	216	409	624	A0A669KB16|A0A669KB16_HUMAN Collagen alpha-1(XIII) chain OS=Homo s
707.14	41	2384	2424	Q7Z7M0|MEGF8_HUMAN Multiple epidermal growth factor-like domains p
706.93	87	14	100	P08047|SP1_HUMAN Transcription factor Sp1 OS=Homo sapiens OX=9606
702.59	87	408	494	E9PNV5|E9PNV5_HUMAN Neuron navigator 2 (Fragment) OS=Homo sapiens
702.45	236	406	641	E7ES50|E7ES50_HUMAN Collagen alpha-1(XIII) chain OS=Homo sapiens O
701.61	122	993	1114	Q5TGY3|AHDC1_HUMAN AT-hook DNA-binding motif-containing protein 1
701.31	137	819	955	Q8IWN7|RP1L1_HUMAN Retinitis pigmentosa 1-like 1 protein OS=Homo s
700.85	126	638	763	Q5T1Z8|Q5T1Z8_HUMAN Pumilio homolog 1 OS=Homo sapiens OX=9606 GN=P
699.46	167	258	424	X6REB3|X6REB3_HUMAN Transcription factor E2-alpha OS=Homo sapiens
697.77	210	448	657	A0A669KAZ4|A0A669KAZ4_HUMAN Collagen alpha-1(XIII) chain OS=Homo s
696.33	37	2779	2815	O75592|MYCB2_HUMAN E3 ubiquitin-protein ligase MYCBP2 OS=Homo sapi
696.21	72	1290	1361	O15021|MAST4_HUMAN Microtubule-associated serine/threonine-protein
693.40	413	1	413	A0A3B3ITG7|A0A3B3ITG7_HUMAN Collagen alpha-1(IV) chain (Fragment)
691.43	233	291	523	P0CG12|DERPC_HUMAN Decreased expression in renal and prostate canc
689.68	93	325	417	Q8NF64|ZMIZ2_HUMAN Zinc finger MIZ domain-containing protein 2 OS=
689.68	93	325	417	A0A087X127|A0A087X127_HUMAN Zinc finger MIZ domain-containing prot
689.34	119	1540	1658	Q71F56|MD13L_HUMAN Mediator of RNA polymerase II transcription sub
689.34	119	1540	1658	A0A3B3IRX3|A0A3B3IRX3_HUMAN Mediator of RNA polymerase II transcri
685.05	33	325	357	Q7Z408|CSMD2_HUMAN CUB and sushi domain-containing protein 2 OS=Ho
684.01	62	514	575	A0A3B3IRW6|A0A3B3IRW6_HUMAN Glutamine and serine-rich protein 1 OS
683.54	263	100	362	H0Y2R3|H0Y2R3_HUMAN AT-rich interactive domain-containing protein
683.38	401	489	889	F8WDM8|F8WDM8_HUMAN Collagen alpha-1(XXIV) chain OS=Homo sapiens O
682.85	126	602	727	Q14671|PUM1_HUMAN Pumilio homolog 1 OS=Homo sapiens OX=9606 GN=PUM
682.85	126	603	728	Q5T1Z4|Q5T1Z4_HUMAN Pumilio homolog 1 OS=Homo sapiens OX=9606 GN=P
682.66	97	2039	2135	Q96RV3|PCX1_HUMAN Pecanex-like protein 1 OS=Homo sapiens OX=9606 G
680.87	59	1206	1264	P10071|GLI3_HUMAN Transcriptional activator GLI3 OS=Homo sapiens O
677.41	94	2461	2554	Q12830|BPTF_HUMAN Nucleosome-remodeling factor subunit BPTF OS=Hom

677.13	65	1	65	A0A088AWL3|A0A088AWL3_HUMAN Nuclear receptor corepressor 1 OS=Homo
676.89	458	196	653	Q9BXS0|COPA1_HUMAN Collagen alpha-1(XXV) chain OS=Homo sapiens OX=
676.19	56	2338	2393	O75376|NCOR1_HUMAN Nuclear receptor corepressor 1 OS=Homo sapiens
676.14	193	261	453	P02671|FIBA_HUMAN Fibrinogen alpha chain OS=Homo sapiens OX=9606 G
674.88	37	2741	2777	A0A499FJI4|A0A499FJI4_HUMAN RCR-type E3 ubiquitin transferase OS=H
671.51	34	270	303	P02751|FINC_HUMAN Fibronectin OS=Homo sapiens OX=9606 GN=FN1 PE=1
671.21	85	45	129	O94916|NFAT5_HUMAN Nuclear factor of activated T-cells 5 OS=Homo s
670.98	63	65	127	Q6L8H4|KRA51_HUMAN Keratin-associated protein 5-1 OS=Homo sapiens
670.25	140	334	473	A0A6Q8PH46|A0A6Q8PH46_HUMAN Mucin-19 (Fragment) OS=Homo sapiens OX
670.05	107	737	843	A0A087X0G5|A0A087X0G5_HUMAN Mastermind-like protein 2 OS=Homo sapi
667.88	84	2268	2351	Q9Y520|PRC2C_HUMAN Protein PRRC2C OS=Homo sapiens OX=9606 GN=PRRC2
667.59	77	1077	1153	P49792|RBP2_HUMAN E3 SUMO-protein ligase RanBP2 OS=Homo sapiens OX
667.39	89	1984	2072	Q8WYB5|KAT6B_HUMAN Histone acetyltransferase KAT6B OS=Homo sapiens
667.04	359	305	663	A0A0U1RRA7|A0A0U1RRA7_HUMAN Collagen alpha-1(XI) chain (Fragment)
665.90	84	2270	2353	E7EPN9|E7EPN9_HUMAN Protein PRRC2C OS=Homo sapiens OX=9606 GN=PRRC
665.83	59	1147	1205	A0A2R8YGX0|A0A2R8YGX0_HUMAN Transcriptional activator GLI3 OS=Homo
663.48	36	2997	3032	Q96JQ0|PCD16_HUMAN Protocadherin-16 OS=Homo sapiens OX=9606 GN=DCH
663.44	62	521	582	A0A0A0MQR4|A0A0A0MQR4_HUMAN Protein capicua homolog OS=Homo sapien
661.18	76	809	884	P10070|GLI2_HUMAN Zinc finger protein GLI2 OS=Homo sapiens OX=9606
660.58	62	521	582	Q96RK0|CIC_HUMAN Protein capicua homolog OS=Homo sapiens OX=9606 G
658.62	77	1	77	A0A3F2YNZ0|A0A3F2YNZ0_HUMAN Protein transport protein sec16 OS=Hom
656.28	110	416	525	A0A0D9SG23|A0A0D9SG23_HUMAN Methyl-CpG-binding domain protein 5 OS
656.18	50	3262	3311	Q9NYQ7|CELR3_HUMAN Cadherin EGF LAG seven-pass G-type receptor 3 O
656.10	77	1266	1342	Q68CP9|ARID2_HUMAN AT-rich interactive domain-containing protein 2
656.10	77	1240	1316	F8WCU9|F8WCU9_HUMAN AT-rich interactive domain-containing protein
654.13	69	552	620	Q86YV5|PRAG1_HUMAN Inactive tyrosine-protein kinase PRAG1 OS=Homo
650.98	84	240	323	Q15714|T22D1_HUMAN TSC22 domain family protein 1 OS=Homo sapiens O
648.63	179	572	750	A0A087X0K0|A0A087X0K0_HUMAN Collagen alpha-1(XV) chain OS=Homo sap
647.08	77	876	952	F8W108|F8W108_HUMAN AT-rich interactive domain-containing protein
646.71	66	774	839	P55198|AF17_HUMAN Protein AF-17 OS=Homo sapiens OX=9606 GN=MLLT6 P
646.50	77	1	77	Q92945|FUBP2_HUMAN Far upstream element-binding protein 2 OS=Homo
645.31	36	4719	4754	Q96RW7|HMCN1_HUMAN Hemicentin-1 OS=Homo sapiens OX=9606 GN=HMCN1 P
644.29	182	371	552	H0Y5N9|H0Y5N9_HUMAN Collagen alpha-1(XII) chain (Fragment) OS=Homo

643.38	29	1393	1421	E5RIG2|E5RIG2_HUMAN CUB and sushi domain-containing protein 1 OS=H
642.98	122	906	1027	A0A669KBM4|A0A669KBM4_HUMAN DNA-binding protein RFX7 OS=Homo sapie
642.06	77	1	77	A0A3F2YNX0|A0A3F2YNX0_HUMAN Protein transport protein sec16 OS=Hom
640.12	29	1392	1420	Q96PZ7|CSMD1_HUMAN CUB and sushi domain-containing protein 1 OS=Ho
639.00	179	586	764	P39059|COFA1_HUMAN Collagen alpha-1(XV) chain OS=Homo sapiens OX=9
637.56	91	109	199	P31942|HNRH3_HUMAN Heterogeneous nuclear ribonucleoprotein H3 OS=H
636.90	66	1	66	Q6ZRS2|SRCAP_HUMAN Helicase SRCAP OS=Homo sapiens OX=9606 GN=SRCAP
635.17	29	1393	1421	F8W9C3|F8W9C3_HUMAN CUB and sushi domain-containing protein 1 OS=H
635.14	244	426	669	F6M2K2|F6M2K2_HUMAN Nuclear receptor coactivator 6 OS=Homo sapiens
635.13	41	728	768	E7EVZ1|E7EVZ1_HUMAN Zinc finger homeobox protein 4 OS=Homo sapiens
631.47	171	416	586	A0A0A0MQT7|A0A0A0MQT7_HUMAN Collagen alpha-1(XIV) chain OS=Homo sa
630.89	227	163	389	Q96E39|RMXL1_HUMAN RNA binding motif protein, X-linked-like-1 OS=H
630.73	354	1403	1756	A8TX70|CO6A5_HUMAN Collagen alpha-5(VI) chain OS=Homo sapiens OX=9
630.73	354	1403	1756	E9PAL5|E9PAL5_HUMAN Collagen alpha-5(VI) chain OS=Homo sapiens OX=
630.59	40	874	913	C9JG08|C9JG08_HUMAN Uncharacterized protein C2orf16 OS=Homo sapien
630.14	41	728	768	Q86UP3|ZFHX4_HUMAN Zinc finger homeobox protein 4 OS=Homo sapiens
628.75	103	1275	1377	O14513|NCKP5_HUMAN Nck-associated protein 5 OS=Homo sapiens OX=960
628.75	103	1275	1377	A0A0A0MS79|A0A0A0MS79_HUMAN Nck-associated protein 5 OS=Homo sapie
628.64	63	382	444	G3V5H7|G3V5H7_HUMAN SKI family transcriptional corepressor 1 OS=Ho
628.47	74	54	127	H7C269|H7C269_HUMAN Trinucleotide repeat-containing gene 6A protei
625.32	63	410	472	P84550|SKOR1_HUMAN SKI family transcriptional corepressor 1 OS=Hom
625.11	244	426	669	F6M2K4|F6M2K4_HUMAN Nuclear receptor coactivator 6 OS=Homo sapiens
624.99	122	809	930	Q2KHR2|RFX7_HUMAN DNA-binding protein RFX7 OS=Homo sapiens OX=9606
624.64	114	844	957	Q8NET4|RTL9_HUMAN Retrotransposon Gag-like protein 9 OS=Homo sapie
624.48	87	875	961	A0A2R8YDS2|A0A2R8YDS2_HUMAN Ras/Rap GTPase-activating protein SynG
623.57	462	32	493	A0A087X1E1|A0A087X1E1_HUMAN Collagen alpha-1(XXV) chain OS=Homo sa
620.50	110	507	616	Q6AI39|BICRL_HUMAN BRD4-interacting chromatin-remodeling complex-a
620.35	64	788	851	Q9ULM3|YETS2_HUMAN YEATS domain-containing protein 2 OS=Homo sapie
620.16	77	1	77	A0A087WTP3|A0A087WTP3_HUMAN Far upstream element-binding protein 2
619.50	104	267	370	Q96PE2|ARHGH_HUMAN Rho guanine nucleotide exchange factor 17 OS=Ho
618.95	35	917	951	A0A140T956|A0A140T956_HUMAN Tenascin-X (Fragment) OS=Homo sapiens
618.51	379	294	672	A0A1B0GV63|A0A1B0GV63_HUMAN AT-rich interactive domain-containing
616.85	71	713	783	A0A669KBC5|A0A669KBC5_HUMAN Protein unc-80 homolog OS=Homo sapiens

616.16	90	194	283	H0Y4U1|H0Y4U1_HUMAN Tensin-1 OS=Homo sapiens OX=9606 GN=TNS1 PE=1
615.33	87	920	1006	B7ZCA0|B7ZCA0_HUMAN Ras/Rap GTPase-activating protein SynGAP OS=Ho
614.92	37	2223	2259	A0A087WXI2|A0A087WXI2_HUMAN IgGFc-binding protein OS=Homo sapiens
614.79	71	713	783	Q8N2C7|UNC80_HUMAN Protein unc-80 homolog OS=Homo sapiens OX=9606
614.79	71	713	783	A0A669KAW8|A0A669KAW8_HUMAN Protein unc-80 homolog OS=Homo sapiens
614.14	167	178	344	A0A0A0MRB7|A0A0A0MRB7_HUMAN Transcription factor E2-alpha OS=Homo
613.57	291	1	291	F8WC90|F8WC90_HUMAN RNA-binding protein EWS (Fragment) OS=Homo sap
612.12	87	934	1020	Q96PV0|SYGP1_HUMAN Ras/Rap GTPase-activating protein SynGAP OS=Hom
611.74	87	934	1020	A0A2R8Y6T2|A0A2R8Y6T2_HUMAN Ras/Rap GTPase-activating protein SynG
611.53	83	179	261	Q9P2D1|CHD7_HUMAN Chromodomain-helicase-DNA-binding protein 7 OS=H
611.28	87	919	1005	A0A0A0MQZ2|A0A0A0MQZ2_HUMAN Ras/Rap GTPase-activating protein SynG
610.35	174	64	237	F5GY10|F5GY10_HUMAN Transcription factor 12 OS=Homo sapiens OX=960
610.21	63	371	433	G3V3E1|G3V3E1_HUMAN SKI family transcriptional corepressor 1 OS=Ho
610.13	94	203	296	Q5JU85|IQEC2_HUMAN IQ motif and SEC7 domain-containing protein 2 O
608.34	61	1364	1424	A0A494C0D3|A0A494C0D3_HUMAN Protein PRRC2B (Fragment) OS=Homo sapi
608.25	122	119	240	E7EX21|E7EX21_HUMAN Collagen alpha-1(XIII) chain OS=Homo sapiens O
607.09	37	331	367	Q9UGM3|DMBT1_HUMAN Deleted in malignant brain tumors 1 protein OS=
606.34	60	1153	1212	A0A1U9X989|A0A1U9X989_HUMAN NOTCH4 OS=Homo sapiens OX=9606 GN=NOTC
604.97	60	1154	1213	Q99466|NOTC4_HUMAN Neurogenic locus notch homolog protein 4 OS=Hom
604.29	59	2398	2456	F5GXF5|F5GXF5_HUMAN Nucleosome-remodeling factor subunit BPTF (Fra
604.18	100	7935	8034	A6NGQ3|A6NGQ3_HUMAN Non-specific serine/threonine protein kinase O
603.98	29	1254	1282	F5GZ18|F5GZ18_HUMAN CUB and sushi domain-containing protein 1 OS=H
603.98	114	966	1079	Q9NZP6|NPAP1_HUMAN Nuclear pore-associated protein 1 OS=Homo sapie
603.41	60	1156	1215	A0A140T9R5|A0A140T9R5_HUMAN NOTCH4 OS=Homo sapiens OX=9606 GN=NOTC
603.14	37	331	367	A0A590UJ76|A0A590UJ76_HUMAN Deleted in malignant brain tumors 1 pr
603.11	50	104	153	P15502|ELN_HUMAN Elastin OS=Homo sapiens OX=9606 GN=ELN PE=1 SV=4
602.98	228	163	390	P38159|RBMX_HUMAN RNA-binding motif protein, X chromosome OS=Homo
600.93	122	809	930	H0YLX2|H0YLX2_HUMAN DNA-binding protein RFX7 OS=Homo sapiens OX=96
600.77	93	293	385	E7EWM3|E7EWM3_HUMAN Zinc finger MIZ domain-containing protein 2 OS
600.11	133	1	133	A0A2R8YET7|A0A2R8YET7_HUMAN Eyes absent homolog OS=Homo sapiens OX
599.91	41	2308	2348	H7BXX0|H7BXX0_HUMAN CUB and sushi domain-containing protein 3 (Fra
599.66	65	499	563	P04259|K2C6B_HUMAN Keratin, type II cytoskeletal 6B OS=Homo sapien
597.89	65	499	563	P02538|K2C6A_HUMAN Keratin, type II cytoskeletal 6A OS=Homo sapien

597.69	96	1291	1386	A6NEM2|A6NEM2_HUMAN Host cell factor 1 OS=Homo sapiens OX=9606 GN=
597.42	66	126	191	Q9ULD9|ZN608_HUMAN Zinc finger protein 608 OS=Homo sapiens OX=9606
597.34	84	203	286	A0A6Q8PFR7|A0A6Q8PFR7_HUMAN IQ motif and SEC7 domain-containing pr
597.13	303	49	351	A0A2R8Y6K8|A0A2R8Y6K8_HUMAN Mucin-19 (Fragment) OS=Homo sapiens OX
596.79	69	940	1008	A0A590UJ96|A0A590UJ96_HUMAN Uncharacterized protein OS=Homo sapien
596.13	94	2322	2415	A0A2R8Y7Q1|A0A2R8Y7Q1_HUMAN Nucleosome-remodeling factor subunit B
595.61	97	1519	1615	Q9UHV7|MED13_HUMAN Mediator of RNA polymerase II transcription sub
594.85	30	2415	2444	P46531|NOTC1_HUMAN Neurogenic locus notch homolog protein 1 OS=Hom
594.48	260	1	260	H7BXV5|H7BXV5_HUMAN Collagen alpha-1(XVIII) chain (Fragment) OS=Ho
594.14	65	499	563	P48668|K2C6C_HUMAN Keratin, type II cytoskeletal 6C OS=Homo sapien
593.71	56	122	177	P08151|GLI1_HUMAN Zinc finger protein GLI1 OS=Homo sapiens OX=9606
592.95	129	164	292	A0A2R8YGM9|A0A2R8YGM9_HUMAN Eyes absent homolog OS=Homo sapiens OX
592.50	224	261	484	A0A669KB39|A0A669KB39_HUMAN Collagen alpha-1(XIII) chain (Fragment
590.14	96	1291	1386	P51610|HCFC1_HUMAN Host cell factor 1 OS=Homo sapiens OX=9606 GN=H
589.63	101	467	567	Q9ULI3|HEG1_HUMAN Protein HEG homolog 1 OS=Homo sapiens OX=9606 GN
589.16	103	2461	2563	O75962|TRIO_HUMAN Triple functional domain protein OS=Homo sapiens
589.08	40	2269	2308	Q5T1R4|ZEP3_HUMAN Transcription factor HIVEP3 OS=Homo sapiens OX=9
588.79	77	146	222	Q9UIF8|BAZ2B_HUMAN Bromodomain adjacent to zinc finger domain prot
587.63	119	1350	1468	H0YHC1|H0YHC1_HUMAN Mediator of RNA polymerase II transcription su
587.09	74	600	673	Q8N2Y8|RUSC2_HUMAN Iporin OS=Homo sapiens OX=9606 GN=RUSC2 PE=1 SV
586.73	29	4680	4708	Q6V0I7|FAT4_HUMAN Protocadherin Fat 4 OS=Homo sapiens OX=9606 GN=F
586.72	109	981	1089	H7BY37|H7BY37_HUMAN Histone-lysine N-methyltransferase 2C (Fragmen
586.39	100	1407	1506	Q8NEV8|EXPH5_HUMAN Exophilin-5 OS=Homo sapiens OX=9606 GN=EXPH5 PE
586.29	97	124	220	Q9HCD6|TANC2_HUMAN Protein TANC2 OS=Homo sapiens OX=9606 GN=TANC2
585.86	74	1790	1863	Q5HYC2|K2026_HUMAN Uncharacterized protein KIAA2026 OS=Homo sapien
584.09	61	206	266	Q7Z5J4|RAI1_HUMAN Retinoic acid-induced protein 1 OS=Homo sapiens
583.26	60	1154	1213	A0A140T8Y6|A0A140T8Y6_HUMAN NOTCH4 OS=Homo sapiens OX=9606 GN=NOTC
581.94	29	4682	4710	A0A6Q8JR05|A0A6Q8JR05_HUMAN Protocadherin Fat 4 OS=Homo sapiens OX
581.32	99	480	578	Q99700|ATX2_HUMAN Ataxin-2 OS=Homo sapiens OX=9606 GN=ATXN2 PE=1 S
580.27	62	385	446	Q2KHR3|QSER1_HUMAN Glutamine and serine-rich protein 1 OS=Homo sap
579.45	72	247	318	O60299|LZTS3_HUMAN Leucine zipper putative tumor suppressor 3 OS=H
577.86	81	1549	1629	Q6N021|TET2_HUMAN Methylcytosine dioxygenase TET2 OS=Homo sapiens
577.86	81	1570	1650	E7EQS8|E7EQS8_HUMAN Methylcytosine dioxygenase TET OS=Homo sapiens

577.43	71	64	134	Q96T58|MINT_HUMAN Msx2-interacting protein OS=Homo sapiens OX=9606
577.02	184	347	530	Q86VE3|SATL1_HUMAN Spermidine/spermine N(1)-acetyltransferase-like
577.02	184	347	530	A0A2R8YFQ0|A0A2R8YFQ0_HUMAN Spermidine/spermine N(1)-acetyltransfe
575.93	119	935	1053	A0A3B3IS46|A0A3B3IS46_HUMAN Mediator of RNA polymerase II transcri
575.60	307	51	357	O95429|BAG4_HUMAN BAG family molecular chaperone regulator 4 OS=Ho
575.27	64	657	720	P15822|ZEP1_HUMAN Zinc finger protein 40 OS=Homo sapiens OX=9606 G
575.13	77	833	909	A0A590UJW6|A0A590UJW6_HUMAN Zinc finger CCHC domain-containing pro
573.38	132	173	304	Q13151|ROA0_HUMAN Heterogeneous nuclear ribonucleoprotein A0 OS=Ho
573.38	143	1	143	O15534|PER1_HUMAN Period circadian protein homolog 1 OS=Homo sapie

Supporting Figures

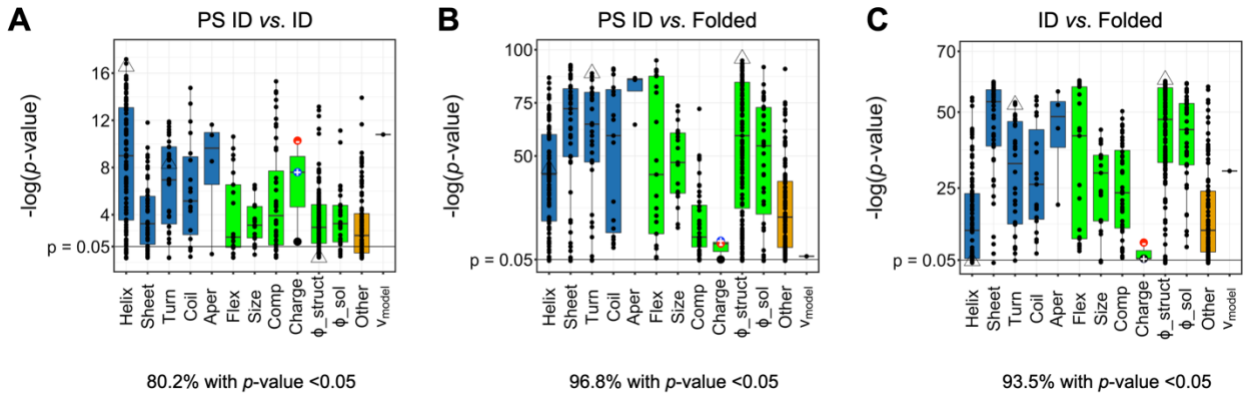

Figure S1. Comparing means in the sequence sets using a nonparametric test. *A-C*, *p*-values calculated by the Mann-Whitney *U*-test, shown as -log(*p*-value), compares set means in 567 amino acid scales and v_{model}. Here, the use of colors and symbols are identical to that used in Figure 2, where conformation-based scales are grouped by type and highlighted by blue boxplots, and physicochemical-based scales are grouped by type and highlighted by green boxplots. Scales (e.g., refractivity, crystal melting point) that did not easily map into a conformation-based or physicochemical-based group were combined separately (Other; orange boxplot). Boxplots show the dataset median (50th percentile) with the central bar, and the vertical width spans the 25th to 75th percentiles. Open triangles highlight the smallest *p*-value from Welch's *t*-test when comparing means in the PS ID and ID sets, which was from an α-helix propensity scale, the smallest *p*-value from Welch's *t*-test when comparing means in either ID set with the folded set, which was from a structure-based hydrophobicity scale, and the β-turn propensity scale used in ParSe (also provided for reference).

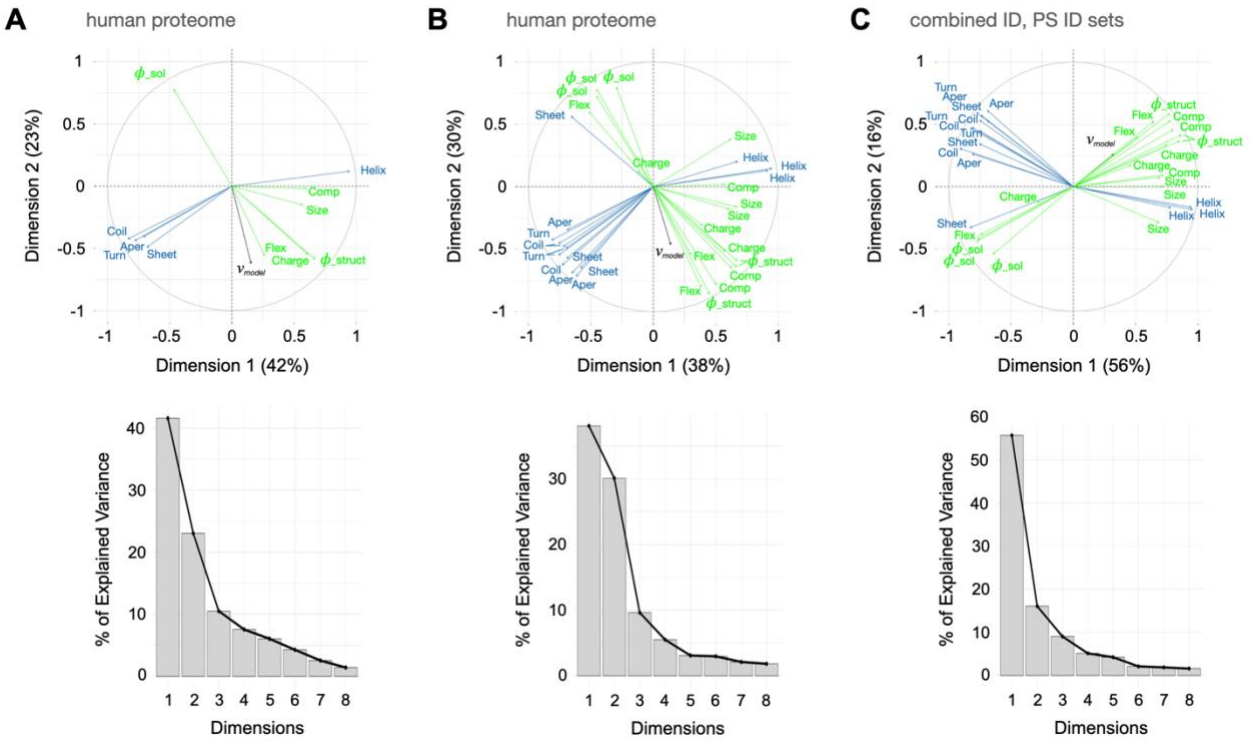

Figure S2. Modes of variance in the sequence sets arising from different amino acid property scales. *A-C*, bidimensional plots (top figures) from PCA showing the modes of variance in the human proteome and the combined ID sets (PS ID and ID) arising from conformation- (blue arrows) and physicochemical-based (green arrows) scales relative to the two principal components of variance, given as Dimension 1 and Dimension 2. Scree plots (bottom figures) showing the percent of the total variance in the human proteome and in the combined set of ID sequences that is captured by each principal component (i.e., dimension).

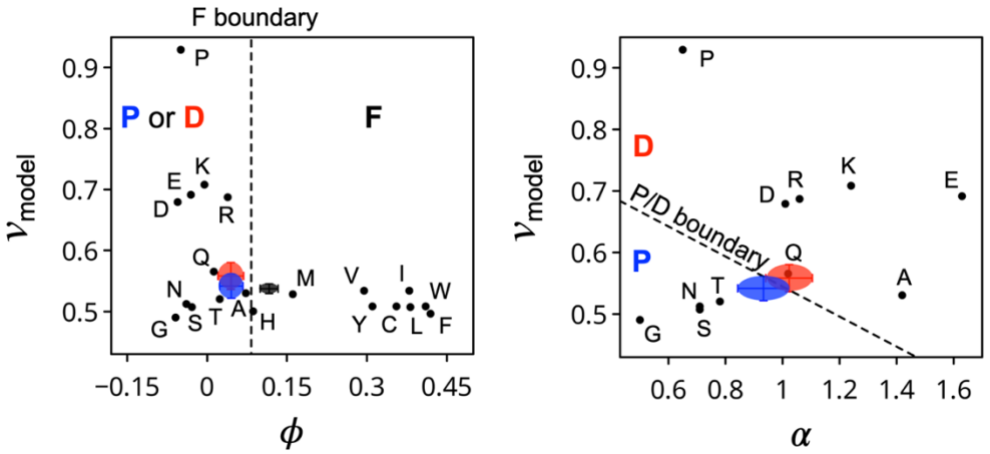

Figure S3. Comparing hydrophobicity, α-helix propensity, and ν_{model} in homopolymers. Hydrophobicity (ϕ) and α-helix propensity (α) were calculated using the scales from Vendruscolo and coworkers (14) and Tanaka and Scheraga (15), respectively, in homopolymers (N = 100) where amino acid type is identified by its one-letter code. Filled circles show the mean and standard deviation in ϕ, α, and ν_{model} in the PS ID (blue), ID (red), and folded sets (black).

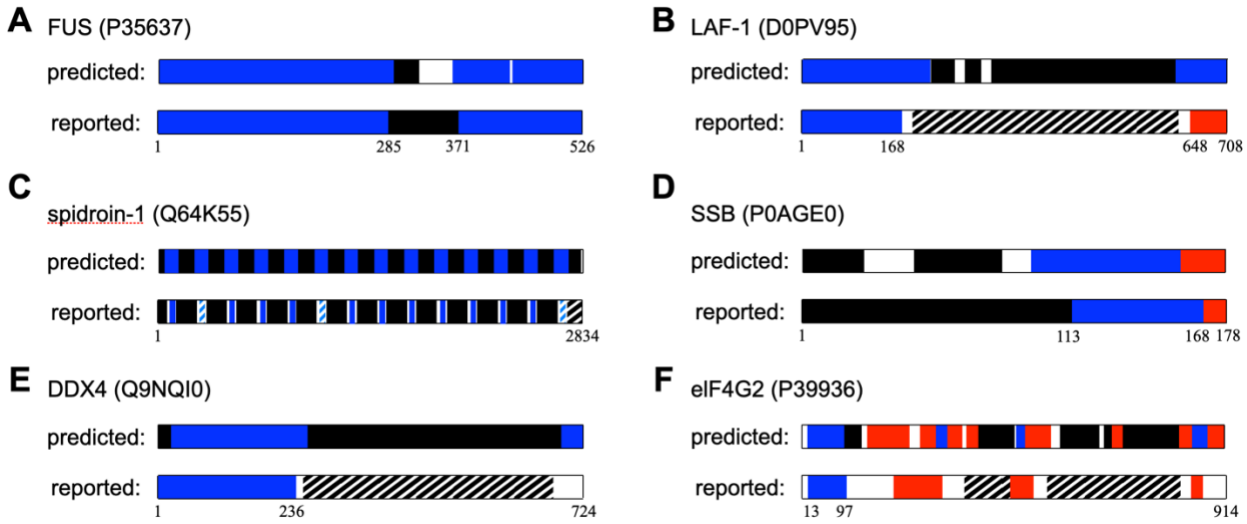

Figure S4. Predicting protein regions that drive LLPS. ParSe v2 was applied to the whole sequences of proteins with diverse reported mechanisms driving LLPS. The proteins are identified by name and UniProt accession number. Contiguous regions ($N \geq 20$) that were 90% of only one label, P, D, or F were colored blue, red, or black, respectively, to represent predicted PS, ID, or folded regions. Striped represents ≥50% identity to a known PS IDR (blue) or folded protein (black).

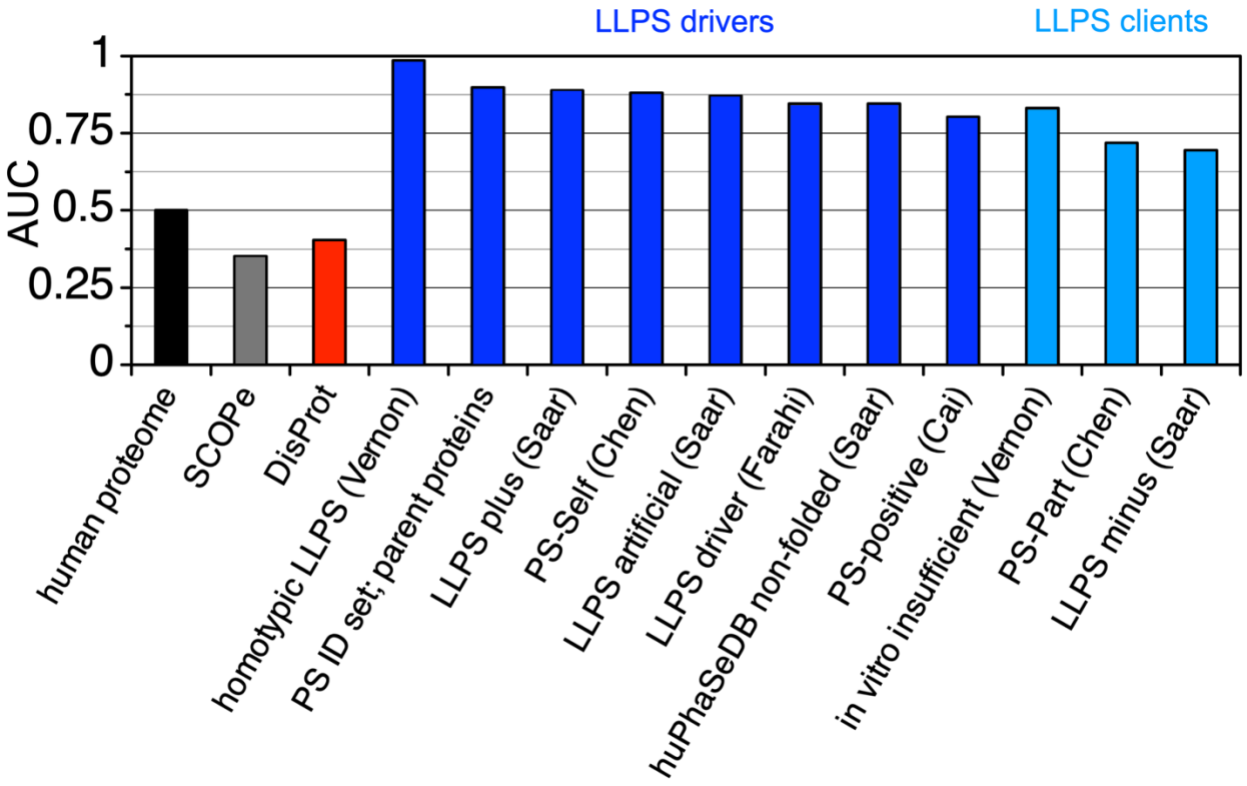

Figure S5. LLPS driver sequences have AUC >0.8 when compared against the human proteome. AUC calculations used the human proteome as the comparison set, and recall was based on the summed P classifier distance, as described in Figure 4. AUC values for SCOPe (grey) and DisProt (red) are reproduced from Figure 4C. LLPS driver sets (blue) are from Vernon et al (3), representing a set of proteins that have been verified *in vitro* to exhibit homotypic phase separation behavior (referred to as "*in vitro* sufficient" by Vernon), the parent proteins of the PS ID sequence set from the current study, and LLPS driver sets from Saar et al (16), Chen et al (17), Farahi et al (18), and Cai et al (19), where the sets are identified by the names used for these sets in each study. For comparison, light blue shows AUC for protein sets thought to have lower potential for phase separation (compared to the driver sets) because these proteins require partners (Vernon (3) and Chen sets (17)) and/or relatively high protein concentrations (>100 µM; Saar set (16)) for LLPS.

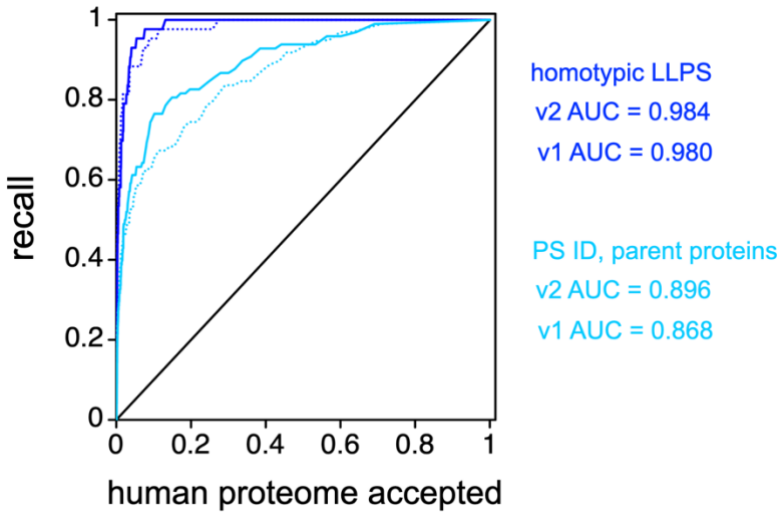

Figure S6. ParSe v2 shows improved recall compared to the original version. Homotypic LLPS is the Vernon et al set of proteins that have been verified *in vitro* to exhibit homotypic phase separation behavior (3). Solid lines are ParSe v2 results, while stippled lines are from the original ParSe algorithm. Data in this figure is a reproduction of the results in Figure 4A. Calculated AUC values are indicated to the right of the figure.

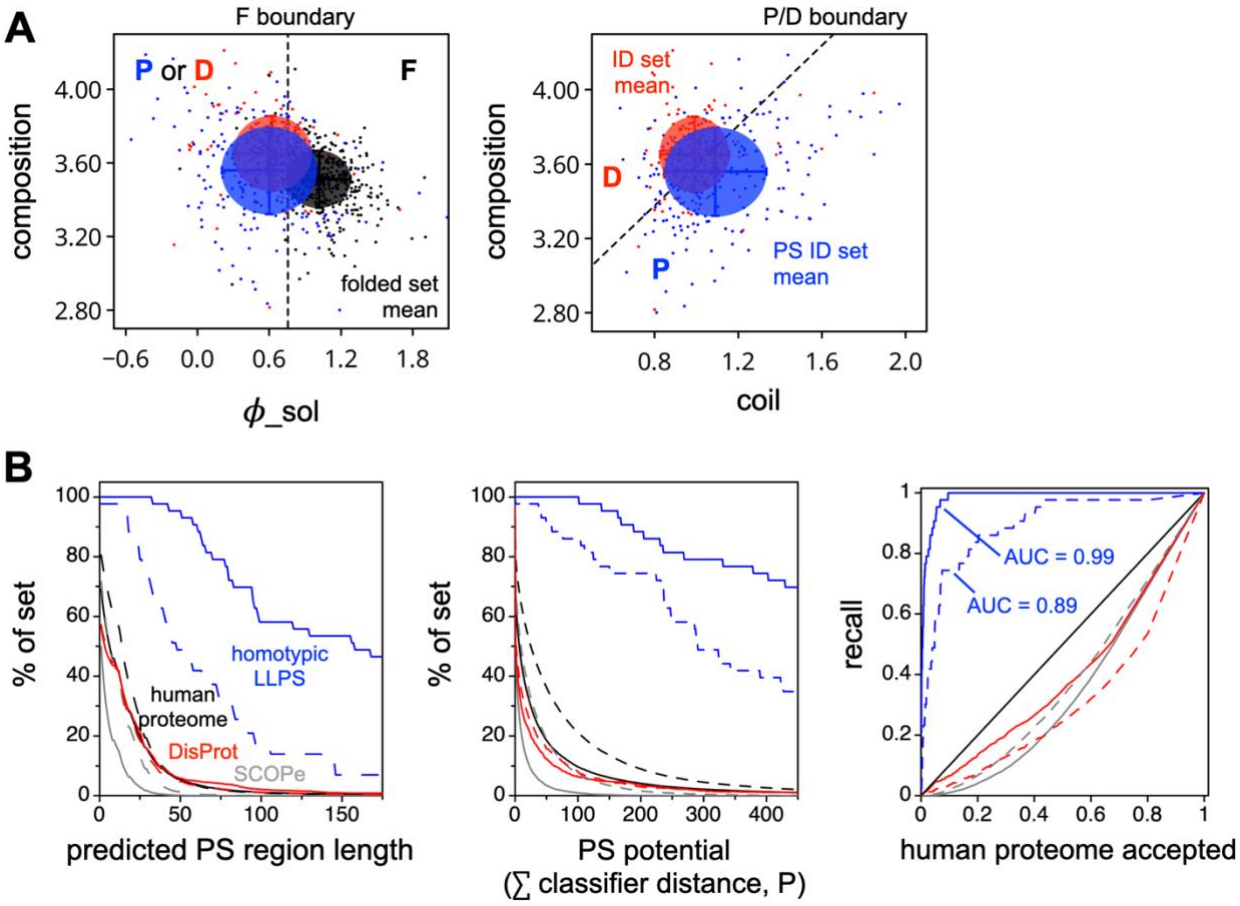

Figure S7. ParSe v2 shows reduced recall when using scales with weaker predictive value. *A*, the structure-based hydrophobicity scale from Vendruscolo and coworkers (14) was substituted for a solution-based hydrophobicity scale from Wilce et al (20) with *t*-test *p*-values of 3.4E-21 and 1.7E-18 when comparing means in the folded and PS ID and folded and ID sets, respectively. This solution-based hydrophobicity scale was used to identify F windows from P or D. A composition-based scale from Jukes et al (21), with a *t*-test *p*-value of 7.4E-08 when comparing means in the PS ID and ID sets, and a coil propensity scale from Isogai et al (22), with a *t*-test *p*-value of 4.6E-08 when comparing means in the PS ID and ID sets, were used to identify P windows from D. *B*, when using these weaker scales with ParSe v2 (dashed lines), the overall predictive value, as judged by AUC (left-most figure), decreased relative to ParSe v2 when using the top-performing scales (solid lines).

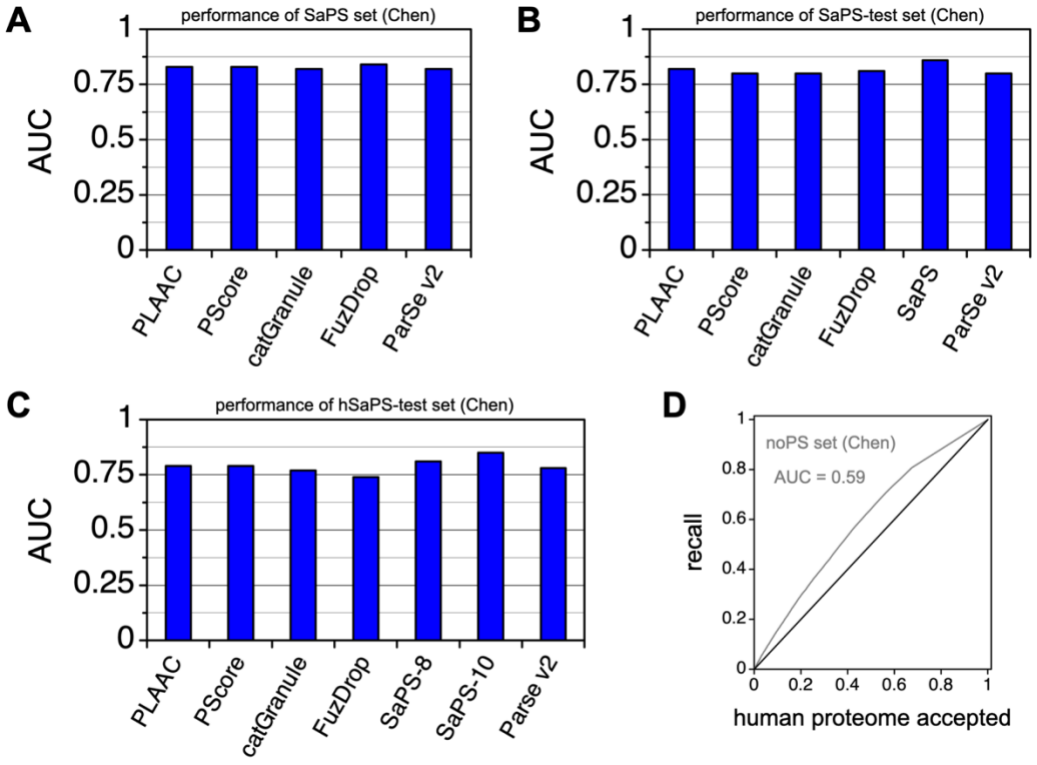

Figure S8. ParSe v2 shows similar predictive accuracy as other LLPS predictors. AUC values for PLAAC, PScore, catGranule, FuzDrop, SaPS, SaPS-8, SaPS-10 are reproduced from the scores given in Figures 1D, 2E, and S2B in Chen et al (17). AUC values for ParSe v2 used recall based on the summed P classifier distance, as described in Figure 4. *A*, SaPS, *B*, SaPS-test, and *C*, hSaPS-test sets were evaluated against the NoPS set. These sequence sets were obtained from Chen et al (17). *D*, recall in the NoPS set compared to the human proteome gives AUC >0.5, indicating that ParSe v2 predicts the NoPS set is enriched in PS.

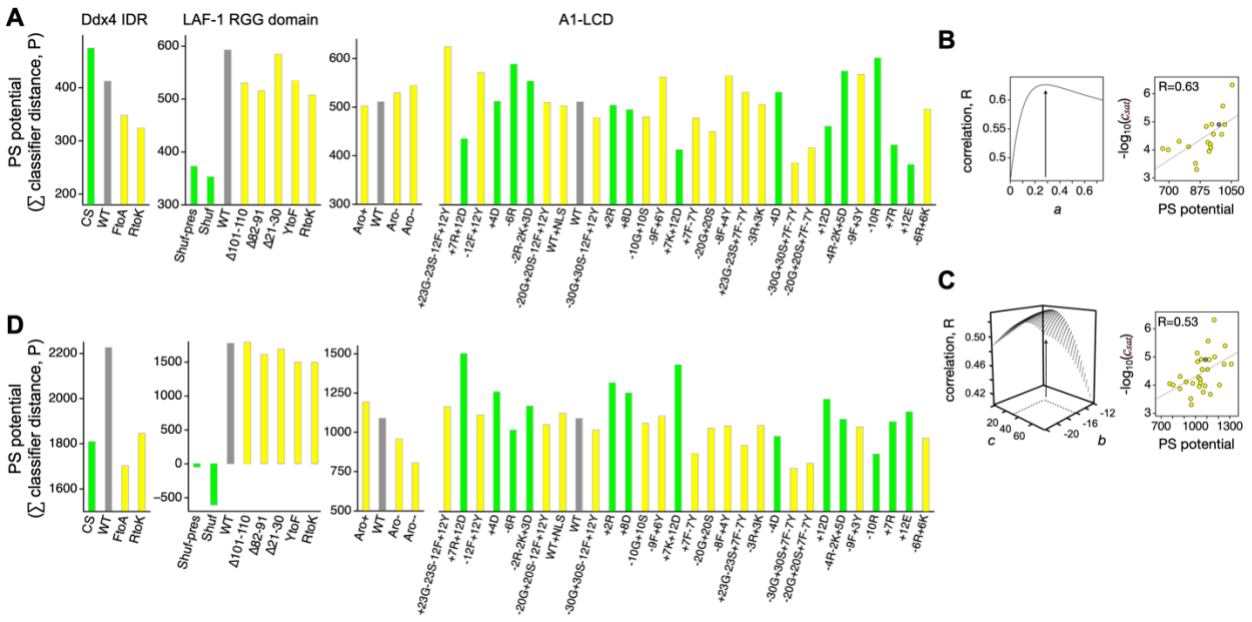

Figure S9. Predicting mutation effects on phase separation behavior by training against c_{sat}. *A*, the summed classifier distance of P-labeled positions was used to calculate a phase-separating (PS) potential from sequence. Mutants were grouped by experimental study and colored grey for wildtype (WT), yellow for mutants with both *NCPR* and *SCD* identical to the WT values, and green otherwise (non-WT *NCPR* and *SCD*). Placement left-to-right within a study follows the reported PS potential in rank, from high-to-low, for comparison to the predicted PS potential. A1-LCD mutants used c_{sat} to establish rank. *B*, A1-LCD mutants with *NCPR* and *SCD* matching the WT values were used to fix a in Equation 3 by optimizing the correlation of Parse-calculated PS potential (including U_π) to -log$_{10}$(c_{sat}); the right figure shows the optimal correlation. *C*, similarly, all A1-LCD mutants with experimental c_{sat} were then used to fix b and c in Equation 4 by optimizing the correlation of ParSe-calculated PS potential (including U_π and U_q) to -log$_{10}$(c_{sat}); the right figure shows the optimal correlation. *D*, ParSe-calculated PS potentials (including U_π and U_q optimized to -log$_{10}$(c_{sat})) for the mutant and WT sequences.

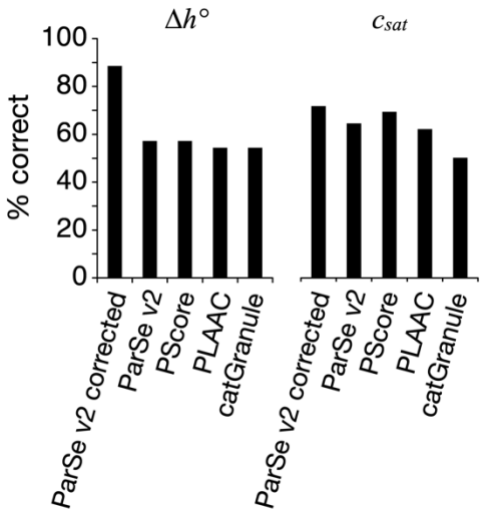

Figure S10. ParSe v2 and other LLPS predictors show similar accuracy for predicting mutation effects. Each predictor was used to rank the mutant sequences in order of phase separation potential, for the set of mutants shown in Figures 5 (rank determined by Δh°) and S9 (rank determined by c_{sat}). Percent correct is the number of mutant sequences that correctly predicted an increase or decrease relative to the wildtype sequence, divided by the total number of mutants and given as a percentage. Granule propensity was used for the catGranule score and LLR was used for the PLAAC score. "ParSe v2 corrected" refers to PS potential (sum of P-labeled windows) including U_π and U_q.

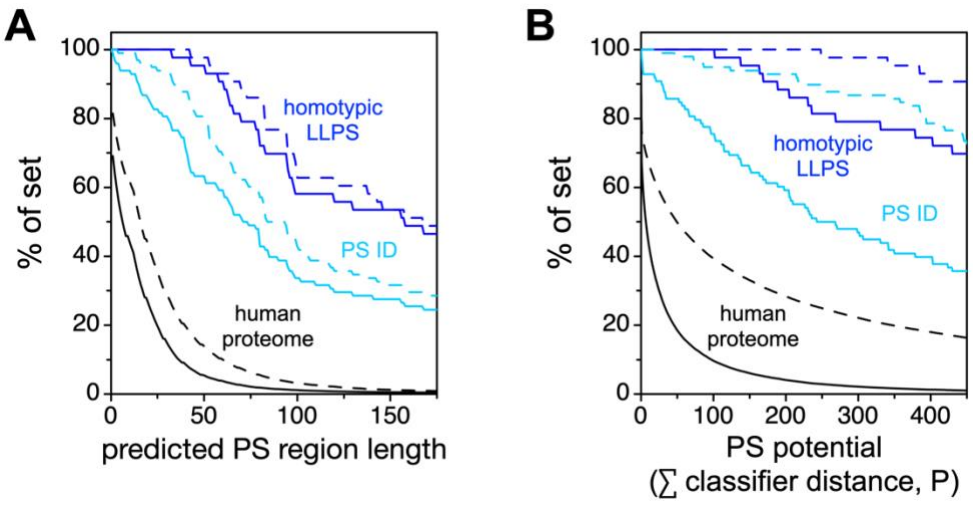

Figure S11. U_π and U_q effects on ParSe predicted PS regions and potential. *A*, ParSe v2 (solid lines) and ParSe v2 including U_π and U_q in the calculations (dashed lines) were used to identify regions in proteins that were ≥90% labeled P, which are referred to as phase-separating, PS, regions. Shown by the y-axis is the percent of proteins in a set with PS regions at least as long as the length indicated by the x-axis. The human proteome (UniProt reference proteome UP000005640) is given by black lines; a set of *in vitro* sufficient homotypic LLPS proteins by blue lines; and the full sequences of the proteins in the PS ID set by light blue lines. *B*, the summed P classifier distance was calculated for the proteins sets in panel A, using both ParSe v2 (solid lines) and ParSe v2 including U_π and U_q (dashed lines). Shown by the y-axis is the percent of proteins in a set with a summed P classifier distance at least as much as the value indicated by the x-axis. Lines were colored using the same coloring scheme as in panel A.

Supporting References

1. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Research*. **28**, 235–242
2. Paiz, E. A., Allen, J. H., Correia, J. J., Fitzkee, N. C., Hough, L. E., and Whitten, S. T. (2021) Beta turn propensity and a model polymer scaling exponent identify intrinsically disordered phase-separating proteins. *Journal of Biological Chemistry*. **297**, 101343
3. Vernon, R. M., Chong, P. A., Tsang, B., Kim, T. H., Bah, A., Farber, P., Lin, H., and Forman-Kay, J. D. (2018) Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife*. 10.7554/eLife.31486
4. Mészáros, B., Erdős, G., Szabó, B., Schád, É., Tantos, Á., Abukhairan, R., Horváth, T., Murvai, N., Kovács, O. P., Kovács, M., Tosatto, S. C. E., Tompa, P., Dosztányi, Z., and Pancsa, R. (2020) PhaSePro: the database of proteins driving liquid–liquid phase separation. *Nucleic Acids Res*. **48**, D360–D367
5. Hatos, A., Hajdu-Soltész, B., Monzon, A. M., Palopoli, N., Álvarez, L., Aykac-Fas, B., Bassot, C., Benítez, G. I., Bevilacqua, M., Chasapi, A., Chemes, L., Davey, N. E., Davidović, R., Dunker, A. K., Elofsson, A., Gobeill, J., Foutel, N. S. G., Sudha, G., Guharoy, M., Horvath, T., Iglesias, V., Kajava, A. V., Kovacs, O. P., Lamb, J., Lambrughi, M., Lazar, T., Leclercq, J. Y., Leonardi, E., Macedo-Ribeiro, S., Macossay-Castillo, M., Maiani, E., Manso, J. A., Marino-Buslje, C., Martínez-Pérez, E., Mészáros, B., Mičetić, I., Minervini, G., Murvai, N., Necci, M., Ouzounis, C. A., Pajkos, M., Paladin, L., Pancsa, R., Papaleo, E., Parisi, G., Pasche, E., Barbosa Pereira, P. J., Promponas, V. J., Pujols, J., Quaglia, F., Ruch, P., Salvatore, M., Schad, E., Szabo, B., Szaniszló, T., Tamana, S., Tantos, A., Veljkovic, N., Ventura, S., Vranken, W., Dosztányi, Z., Tompa, P., Tosatto, S. C. E., and Piovesan, D. (2020) DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* **48**, D269–D276
6. Wang, S., Gu, J., Larson, S. A., Whitten, S. T., and Hilser, V. J. (2008) Denatured-State Energy Landscapes of a Protein Structural Database Reveal the Energetic Determinants of a Framework Model for Folding. *Journal of Molecular Biology*. **381**, 1184–1201
7. Fitzkee, N. C., and Rose, G. D. (2004) Reassessing random-coil statistics in unfolded proteins. *PNAS*. **101**, 12497–12502
8. Panja, A. S., Maiti, S., and Bandyopadhyay, B. (2020) Protein stability governed by its structural plasticity is inferred by physicochemical factors and salt bridges. *Sci Rep*. **10**, 1822
9. Chen, N., Das, M., LiWang, A., and Wang, L.-P. (2020) Sequence-Based Prediction of Metamorphic Behavior in Proteins. *Biophysical Journal*. **119**, 1380–1390
10. Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Kent Wenger, R., Yao, H., and Markley, J. L. (2008) BioMagResBank. *Nucleic Acids Research*. **36**, D402–D408
11. Brady, J. P., Farber, P. J., Sekhar, A., Lin, Y.-H., Huang, R., Bah, A., Nott, T. J., Chan, H. S., Baldwin, A. J., Forman-Kay, J. D., and Kay, L. E. (2017) Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E8194–E8203

12. Martin, E. W., Holehouse, A. S., Peran, I., Farag, M., Incicco, J. J., Bremer, A., Grace, C. R., Soranno, A., Pappu, R. V., and Mittag, T. (2020) Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*. **367**, 694–699
13. Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., and Mittag, T. (2022) Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat. Chem.* **14**, 196–207
14. Bastolla, U., Porto, M., Roman, H. E., and Vendruscolo, M. (2005) Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. *Proteins*. **58**, 22–30
15. Tanaka, S., and Scheraga, H. A. (1977) Statistical Mechanical Treatment of Protein Conformation. 5. Multistate Model for Specific-Sequence Copolymers of Amino Acids. *Macromolecules*. **10**, 9–20
16. Saar, K. L., Morgunov, A. S., Qi, R., Arter, W. E., Krainer, G., Lee, A. A., and Knowles, T. P. J. (2021) Learning the molecular grammar of protein condensates from sequence determinants and embeddings. *Proc Natl Acad Sci U S A*. **118**, e2019053118
17. Chen, Z., Hou, C., Wang, L., Yu, C., Chen, T., Shen, B., Hou, Y., Li, P., and Li, T. (2022) Screening membraneless organelle participants with machine-learning models that integrate multimodal features. *Proc Natl Acad Sci U S A*. **119**, e2115369119
18. Farahi, N., Lazar, T., Wodak, S. J., Tompa, P., and Pancsa, R. (2021) Integration of Data from Liquid-Liquid Phase Separation Databases Highlights Concentration and Dosage Sensitivity of LLPS Drivers. *Int J Mol Sci*. **22**, 3017
19. Cai, H., Vernon, R. M., and Forman-Kay, J. D. (2022) An interpretable machine learning algorithm to predict disordered protein phase separation based on biophysical interactions. bioRxiv doi: 10.1101/2022.07.06.499043
20. Wilce, M. C. J., Aguilar, M.-Isabel., and Hearn, M. T. W. (1995) Physicochemical Basis of Amino Acid Hydrophobicity Scales: Evaluation of Four New Scales of Amino Acid Hydrophobicity Coefficients Derived from RP-HPLC of Peptides. *Anal. Chem.* **67**, 1210–1219
21. Jukes, T. H., Holmquist, R., and Moise, H. (1975) Amino acid composition of proteins: Selection against the genetic code. *Science*. **189**, 50–51
22. Isogai, Y., Némethy, G., Rackovsky, S., Leach, S. J., and Scheraga, H. A. (1980) Characterization of multiple bends in proteins. *Biopolymers*. **19**, 1183–1210