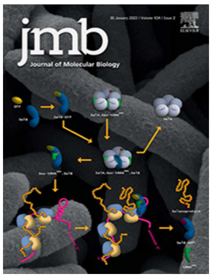

# Yeast Proteins may Reversibly Aggregate like Amphiphilic Molecules

**Pouria Dasmeh** [1,3,4*] **and Andreas Wagner** [1,2,4,5*]

*1 - **Institute for Evolutionary Biology and Environmental Studies,** University of Zurich, Zurich, Switzerland*
*2 - **The Santa Fe Institute,** Santa Fe, NM, USA*
*3 - **Department of Chemistry and Chemical Biology,** Harvard University, Cambridge, MA 02139, USA*
*4 - **Swiss Institute of Bioinformatics (SIB),** Switzerland*
*5 - **Stellenbosch Institute for Advanced Study (STIAS),** Wallenberg Research Centre at Stellenbosch University, Stellenbosch 7600, South Africa*

***Correspondence to Pouria Dasmeh and Andreas Wagner:*** Institute for Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland. *pouria.dasmeh@uzh.ch (P. Dasmeh), andreas.wagner@ieu.uzh.ch (A. Wagner)*
*@PouriaDasmeh (P. Dasmeh), @WagnerEvolution (A. Wagner)*
https://doi.org/10.1016/j.jmb.2021.167352
***Edited by Rita Casadio***

## Abstract

More than a hundred proteins in yeast reversibly aggregate and phase-separate in response to various stressors, such as nutrient depletion and heat shock. We know little about the protein sequence and structural features behind this ability, which has not been characterized on a proteome-wide level. To identify the distinctive features of aggregation-prone protein regions, we apply machine learning algorithms to genome-scale limited proteolysis-mass spectrometry (LiP-MS) data from yeast proteins. LiP-MS data reveals that 96 proteins show significant structural changes upon heat shock. We find that in these proteins the propensity to phase separate cannot be solely driven by disordered regions, because their aggregation-prone regions (APRs) are not significantly disordered. Instead, the phase separation of these proteins requires contributions from both disordered and structured regions. APRs are significantly enriched in aliphatic residues and depleted in positively charged amino acids. Aggregator proteins with longer APRs show a greater propensity to aggregate, a relationship that can be explained by equilibrium statistical thermodynamics. Altogether, our observations suggest that proteome-wide reversible protein aggregation is mediated by sequence-encoded properties. We propose that aggregating proteins resemble supra-molecular amphiphiles, where APRs are the hydrophobic parts, and non-APRs are the hydrophilic parts.

Proteins can aggregate reversibly or irreversibly. Irreversible aggregation is often pathological, indicates damaged cellular regulation,[1–3] and is involved in multiple diseases such as Alzheimer's and Parkinson's diseases.[4,5] Reversible aggregation, however, can be beneficial and help cells survive stressors. For example, in yeast more than a hundred proteins in multiple subcellular compartments form reversible aggregates in response to nutrient starvation, heat shock, or chemical stress.[6] These aggregated proteins are not misfolded or tagged for degradation.[7] Instead, they help increase cells re-initiate growth during recovery from stress by protecting metabolic enzymes from degradation.[6–9]

Reversible protein aggregation is a special case of a widespread phenomenon called protein phase-separation. In this process, a well-mixed solution of proteins de-mixes into two phases of high and low densities.[10] Proteins can phase separate with other proteins, but also with RNA or DNA molecules, into biomolecular condensates that regulate transcription,[11] chromatin states,[12,13] and RNA metabolism.[14] Phase-separating proteins often harbor intrinsically-disordered regions (domains) that do not fold into well-defined tertiary structures.[15] These regions can form a network of specific and non-specific protein interactions to promote phase separation.[16]

The role of disordered regions in stress-induced phase separation has been controversial. On the one hand, mammalian and yeast proteins that phase separate under stress are highly disordered and enriched in features that favor liquid–liquid phase separation. These features include many protein–protein interactions and low complexity regions with multiple polar and charged amino acids, as well as specific amino acids such as tyrosine.[17] These observations suggest that the assembly of low-complexity and disordered regions drive the stress-induced phase separation of these proteins. On the other hand, disordered domains are not essential for phase-separation in some well-studied proteins, where they only modulate the temperature at which phase separation begins. Examples include the poly-adenylate binding protein (Pab1), and the ATP-binding RNA helicase (Ded1) from yeast.[18,19] To understand the importance of disordered regions for stress-induced aggregation and phase separation comprehensively requires a proteome-wide view. In this work, we provide such a view. We also characterize sequence features that can help predict whether proteins will aggregate reversibly.

We took advantage of recent proteome-wide data from limited proteolysis-mass spectrometry (LiP-MS).[20,21] This method permits the detection of both pronounced and subtle changes that protein and peptide structures experience in response to stressors such as heat and osmotic shock. Recently, Cappelletti *et al.* used the method to characterize structural changes in the yeast proteome after heat shock.[22] They identified 96 proteins that show significant structural changes compared to unperturbed cell lysates.[22] These proteins were called *aggregators* and have diverse biological functions, such as telomeric DNA-binding, RNA-helicase activity, or ribosome assembly. The three most highly enriched gene ontology (GO) terms for molecular function among these proteins are RNA binding ($p \sim 10^{-14}$; GO:0003723), heterocyclic compound binding ($p \sim 10^{-13}$; GO:1901363), and translation regulator activity ($p \sim 10^{-12}$; GO:0008135). The proteins are also involved in the biogenesis of ribonucleoprotein complexes and of the ribosome ($p \sim 10^{-13}$; GO:0022613, GO:0042254). Because they are functionally diverse, they constitute an excellent dataset to examine sequence features that facilitate reversible protein aggregation. The length of these proteins varies from 163 to 4911 amino acids (average: ~828 amino acids). We studied regions within these proteins whose proteolysis resistance changes significantly upon aggregation[22] (Table S2, S3, and S5). We refer to these regions as aggregation-prone regions (APRs, Figure 1(A)). The total length of APRs in these proteins was ~42 ± 51 amino acids.

Our first analysis focuses on the question whether APRs preferentially comprise intrinsically disordered regions. To find out, we first calculated the extent of disorder for all amino acids within aggregator proteins, using the *IUpred* disorder predictor[23] (Dataset S1). This algorithm assigns a disorder score between 0 and 1 to every amino acid, based on the amino acid sequence surrounding this amino acid. The score is assigned by a statistical method that distinguishes globular domains from intrinsically disordered regions. Amino acids with a score exceeding 0.5 are predicted to be disordered. Remarkably, such amino acids were not significantly enriched in APRs ($p \sim 1$; Fisher's exact test). To the contrary, amino acids with disorder scores less than 0.5 were enriched ($p \sim 10^{-7}$; Fisher's exact test). We repeated this analysis with the *metapredict* software, which also predicts intrinsically disordered regions in protein sequences.[24] Its predictive algorithm uses neural networks that were trained on consensus disorder scores from several proteomes and, like the IUpred predictor, assigns a score between 0 and 1 to the protein's residues. This score represents the preference of different amino acids to occur in disordered regions. Using this predictor, we also found that amino acids with disorder scores greater than 0.5 were not enriched in APRs ($p \sim 1$; Fisher's exact test). Finally, we restricted our analysis to 17 proteins in yeast that aggregate the most in response to heat shock (also known as superaggregator proteins).[7] For these proteins too, we did not observe a significant enrichment of disordered residues in APRs ($p \sim 0.5$; Fisher's exact test). Altogether, these results suggest that APRs are not preferentially disordered.

Next, we studied the incidence of disorder in APRs for each member of our protein data set (Dataset S2). We found thatc15% of aggregator proteins were significantly more enriched with disordered amino acids than expected by chance (Figure 1(B), $p < 0.05$; Fisher's exact test, false discovery rate (FDR)-corrected for multiple testing).[25] Conversely, disordered amino acids were significantly depleted in ~11% of aggregators ($p < 0.05$; Fisher's exact test, corrected for multiple testing). The proteins whose APRs were significantly disordered are preferentially involved in RNA and nucleic acid binding (Figure 1(C); Dataset S3, GO terms of GO:0003723, and GO:0003676).

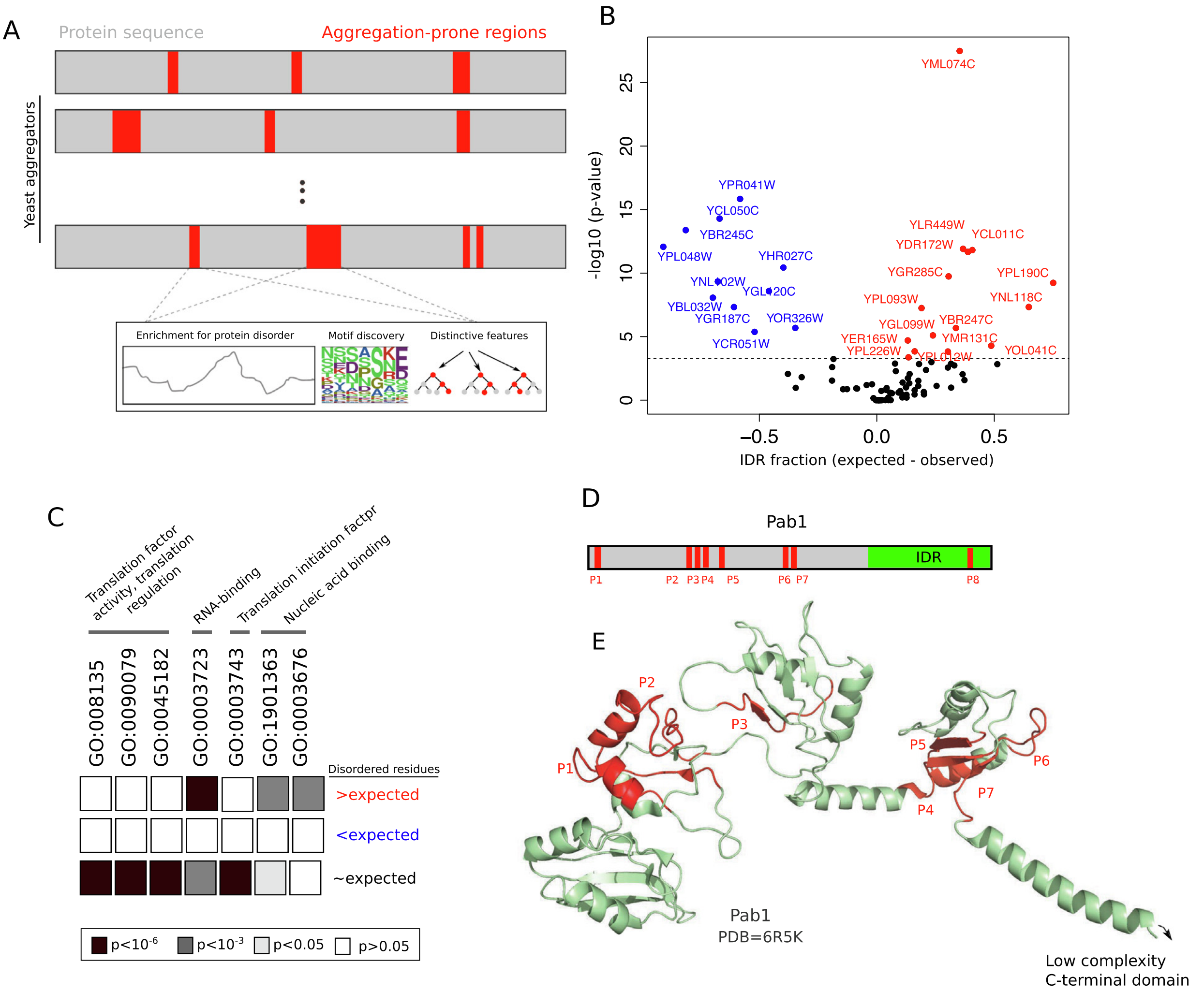

**Figure 1.** Intrinsically-disordered and structured regions both contribute to reversible aggregation in yeast aggregators. (A) Structure of data set and study design. We identified aggregation-prone regions within 96 aggregator proteins in yeast using a dataset of limited proteolysis-mass spectrometry by Cappelletti *et al.*[22] We performed several sequence-based statistical and predictive analyses and investigated the enrichment of protein disorder, specific sequence composition, and distinctive features of aggregation-prone regions in these proteins. (B) Volcano plot showing the preferential enrichment of intrinsically-disordered regions in aggregation-prone regions of the aggregator proteins. The y-axis represents the logarithm of the p-value calculated from Fisher's exact test and the x-axis shows the difference between the expected and observed fraction of intrinsically disordered amino acids. Proteins with significant enrichment and depletion of disordered amino acids in their aggregation-prone regions are shown in red, and blue, respectively. (C) The enrichment of Gene Ontology terms for molecular function in proteins whose aggregation-prone regions are enriched for (red), and depleted in (blue) disordered residues. (D) The sequence of Pab1 with its intrinsically disordered domains shown in green. Aggregation-prone regions are labeled P1 to P8. (E) The crystal structure of Pab1 (PDB ID = 6R5K).[27] Aggregation-prone regions are shown in red and labeled according to their definition in panel D. The region P8 in Pab1 falls in the low complexity domain, which is missing from the PDB X-ray structure.

For example, the protein with the highest disorder enrichment is the nuclear poly-adenylated RNA-binding protein 3 (NAB3; yeast gene identifier YPL190C), which is required for packaging pre-mRNAs into ribonucleoprotein structures for efficient RNA processing.[26]

To better understand the location of APRs in the sequence and in the protein structure we mapped these regions onto the 3D structure of the well-studied aggregator poly adenylate-binding protein Pab1.[18] Pab1 binds the poly(A) tail of mRNA, and regulates mRNA stability and translation. This protein forms reversible gel-like condensates upon heat shock.[18,19] Figure 1(D) shows APRs within the sequence of Pab1. Importantly, only one of eight APRs falls within the intrinsically disordered domain (green) of this protein. We then mapped APRs on the 3D structure of Pab1 (Figure 1(E); PDB ID = 6R5K).[27] From the figure, these regions form long loops which occur either within an APR or constitute an entire APR. The regions $^{151}$D-K$^{156}$, $^{167}$G-E$^{176}$, $^{186}$L-R$^{204}$, $^{242}$F-K$^{252}$, $^{383}$N-L$^{392}$ are examples of such loop-forming peptides. In sum, our disorder enrichment analysis of aggregator proteins, and the specific case of Pab1 suggests that reversible aggregation of yeast proteins may require a synergy between both disordered and structured regions

What other characteristic sequence features do APRs have? To answer this question, we first asked whether APRs comprise sequence motifs that may facilitate their interaction with each other and cause the aggregation of aggregator proteins. Specifically, we looked for linear sequence motifs in APRs using DALEL,[28] an algorithm for the exhaustive identification of degenerate sequence motifs, but found no such enriched motifs. We then compared the frequencies of amino acids, dipeptides, and tripeptides in APRs with the rest of the protein sequence in aggregators (Dataset S6). Here, we found a significant depletion of the positively charged residues Arg and Lys, and their dipeptides ($p \sim 10^{-16}$; Wilcoxon signed-rank test). We also observed that the sum of aliphatic residues Leu, Ile, Ala, Val, were significantly more frequent in APRs compared to the rest of the protein sequence ($p \sim 10^{-16}$; Wilcoxon signed-rank test). The fraction of positively charged to aliphatic residues was the strongest predictor of APRs compared to the fraction of all other amino acids (Figure 2(A); $p \sim 10^{-21}$; Wilcoxon signed-rank test).

To find out whether these patterns are linked to differences in physicochemical properties of amino acids, we used a random forest approach, a widely-used machine learning technique for the classification of two or more data sets.[29] Specifically, we used all 96 aggregators with experimentally characterized APRs,[22] and subdivided each protein into two sequence data sets, one comprising the APRs and one comprising the non-APRs. For each of these datasets, we calculated a feature matrix that consisted of 500 amino acid properties (Dataset S7, and S8). We took these properties from the AAindex database,[30] which curates various physicochemical and biochemical properties of amino acids. The classifier achieved an accuracy of ~93% in 100 independent runs, with the data split into a training set (80% of the data) and a testing set (20%) (Figure 2(B); Dataset S9). Our approach showed that two amino acid properties were most important in this classification (Figure 2(C)). The first is a low positive charge of amino acids in APRs compared to non-APRs. The second is a high amphiphilic propensity of these amino acids. The amphiphilic propensity measures the preference of amino acids to occur at protein-solvent interfaces, such as the end regions of transmembrane helices.[31] Amino acids like arginine and lysine have the highest preference for these environments, and the largest value of the amphiphilic index. In contrast, hydrophobic amino acids such as leucine, isoleucine, valine, and alanine prefer internal protein environments and have the lowest amphiphilic propensity.[31]

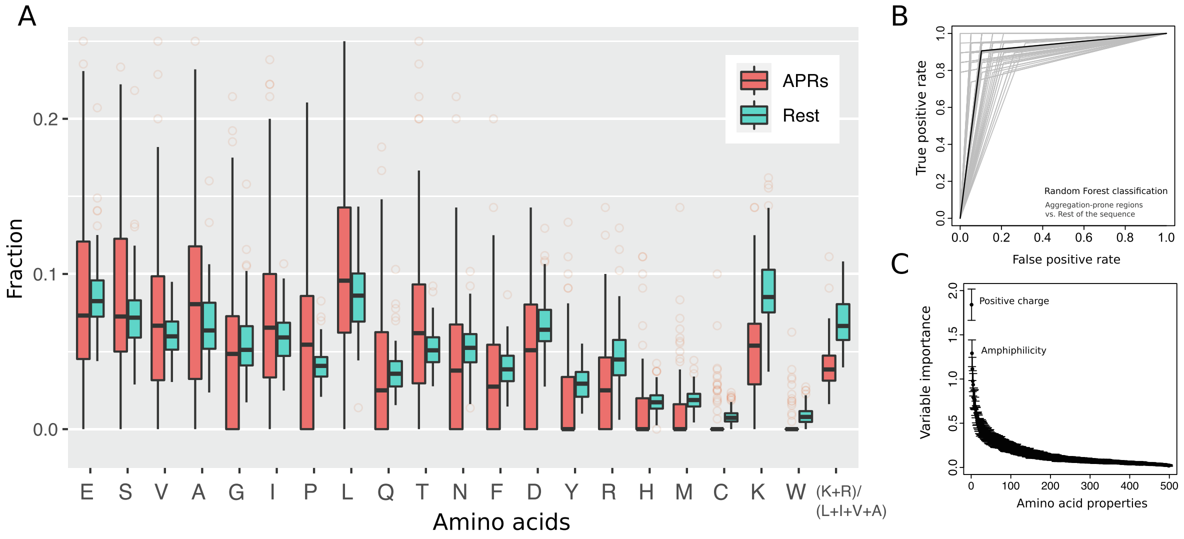

**Figure 2.** Sequence composition and physicochemical properties of aggregation-prone regions in the yeast aggregators. (A) The amino acid frequencies in APRs (shown in red) *versus* the rest of the protein sequence (shown in green) in 96 yeast aggregators. (B) The receiver-operating characteristic (ROC) curves for the clustering of APRs from non-APRs using random forest classification. ROC curves in grey are from 100 random forest clustering and their average is shown in black. (C) Ranked importance of physicochemical variables quantified as the average decrease in the Gini index for different physicochemical properties.

Because of the statistically significant difference between the amphiphilic nature of APRs and non-APRs, we argue that aggregator proteins resemble supramolecular amphiphiles. These molecular structures have distinct hydrophobic and hydrophilic parts.[32] In response to external stimuli such as temperature, pH, and ionic strength,[32] they can reversibly aggregate by non-covalent bonding forces, e.g., through electrostatic and hydrophobic interactions.

We next investigated the relevance of amphiphilic aggregation to stress-induced phase separation of our yeast aggregator proteins. An important feature of amphiphilic molecules is that they are more likely to occur in an aggregated phase if their hydrophobic chain is long.[33] We thus predicted that aggregator proteins with longer APRs will occur preferentially in the aggregated phase. To test this prediction, we asked whether aggregator proteins with long APRs preferentially occur in the pellet fraction of proteins extracted from heat-stressed yeast cells.[34] Specifically, we used for this purpose the $\log_2$ ratio of the protein's abundance in the pellet fraction of yeast cells to that of supernatant as a measure of protein's enrichment in the aggregated phase.[34] This enrichment data exists for 31 aggregated proteins in our dataset. Indeed, aggregator proteins with longer APRs preferentially accumulate in the aggregated pellet fraction (Figure 3(A), $R = 0.53$, $p = 0.0021$; Spearman's rank correlation). To see whether this observation extends to other stress-induced conditions, we also used a mass spectrometry dataset of nutrient-starved yeast cells reported by Narayanasamy *et al.*[35] We identified 56 aggregator proteins in this dataset and determined the association between APR lengths in these proteins and the Z-score of protein enrichment in the pellet fraction. In nutrient-starved cells too, the aggregator proteins with longer APRs preferentially accumulate in the aggregated pellet fraction (Figure 3(B), $R = 0.47$, $p = 0.00014$; Spearman's rank correlation). Importantly, we did not observe a significant association between the length of APRs and the protein length ($R = 0.02$, $p = 0.82$, Spearman's rank correlation), or between protein length and protein enrichment in the pellet fraction ($R = -0.20$, $p = 0.11$, Spearman's rank correlation). This indicates that the reversible aggregation of our study proteins is not significantly biased by sequence length.

We then investigated the relationship between APR length in aggregator proteins and their

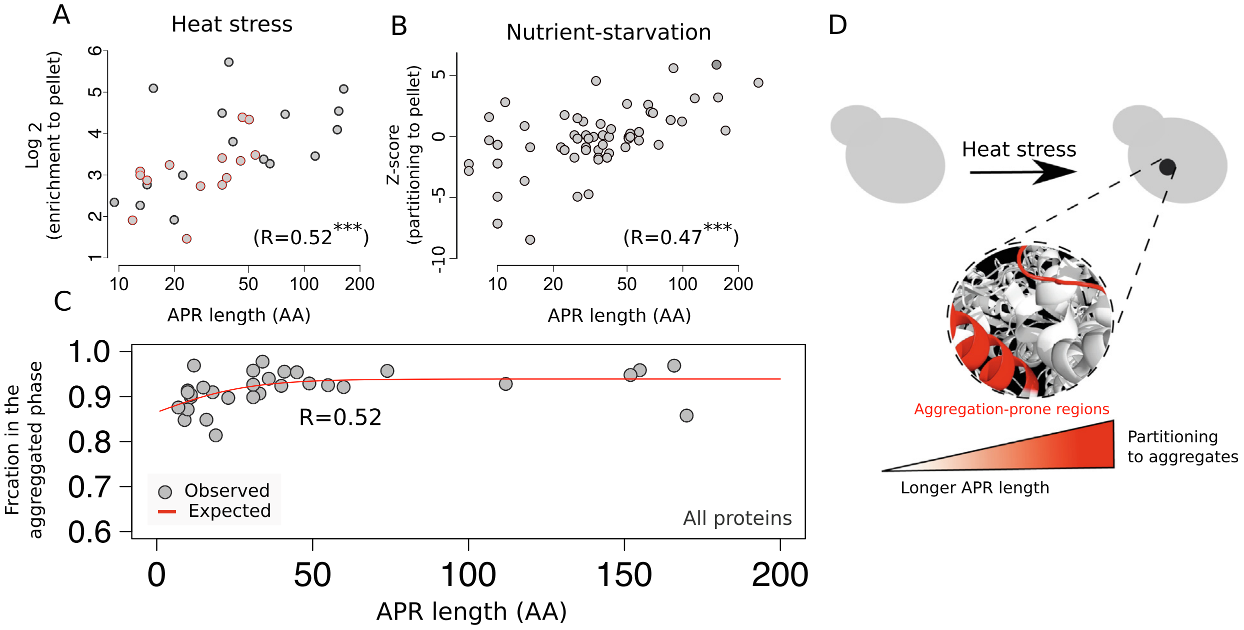

**Figure 3.** Aggregator proteins reversibly aggregate in a manner akin to amphiphilic molecules. (A) The log2 ratio of the protein's abundance in the pellet fraction to that of supernatant in heat-stressed yeast cells (30 out of 96 aggregator proteins[34] *versus* their APR length. (B) The aggregator proteins' enrichment in the pellet fraction upon nutrient starvation (59 out of 96 proteins[34] *versus* their APR length in log-scale. (C) The fraction of aggregator proteins in the aggregated phase calculated from Eq. (3) *versus* the length of APRs. The fitted curve to the observed fraction of proteins in the aggregated phase using Eq. (2) is shown in red. The R value is the Spearman's rank correlation between the expected (Eq. (2)) and observed (Eq. (3)) fractions of proteins in the aggregated phase. (D) Schematic of our two main findings. First, both disordered and structured regions contribute to stress-induced phase separation in yeast. Second, proteins whose aggregation-prone regions are longer or have a lower positive charge-to-aliphatic bias, accumulate preferentially in the aggregated phase.

differential enrichment in the aggregated phase using equilibrium statistical thermodynamics. We assumed that the reversible aggregation of the aggregator proteins proceeds with a mechanism akin to the reversible aggregation of amphiphilic molecules. We considered a two-state model where proteins exist either as a soluble monomer or an insoluble aggregate. The fraction $\mathscr{F}_{cc}$ of proteins occurring in the aggregate at the critical concentration (cc) for aggregation can be expressed as

$$\mathscr{F}_{cc}(theory) = \frac{(P_{agg})_{cc}}{(P_{agg})_{cc} + (P_{mon})_{cc}} = \frac{1}{1 + e^{-\beta \Delta G_{agg}}} \quad (1)$$

Here, $P_{agg}$ and $P_{mon}$ are the relative concentrations of proteins in the aggregated and the monomer phase, $\beta = 1/k_B T$ , where $k_B$ is the Boltzmann constant, and $\Delta G_{agg}$ is the free energy change upon aggregation. Because the aggregation free energy increases in proportion to the length of the hydrophobic chain,[36] we rearranged Eq. (1) as

$$\mathscr{F}_{cc}(theory) = \frac{1}{1 + e^{-\beta BL}} \quad (2)$$

where $B$ is a proportionality constant between the aggregation free energy and the APR length $L$. Eq. (2) states that proteins with longer APRs will partition preferentially to the aggregated phase, which is also observed in the aggregated phase of amphiphilic molecules with different hydrophobic chain lengths.[37] We then calculated the observed fraction of aggregator proteins in the aggregated phase from their reported $\log_2$ enrichment ratio in the pellet fraction of heat-stressed yeast ($log_2 \frac{[Pel}{[Supernatant]}$) as

$$\mathscr{F}_{cc}(experiment) = \left(1 + 2^{-log_2 \frac{[Pellet]}{[Supernatant]}}\right)^{-1} \quad (3)$$

Eq. (2) predicts the observed relationship between APR length and a protein's fraction in the aggregated phase well (Figures 3(C), $R = 0.53$, $p = 0.0021$; Spearman's rank correlation, see supplementary information for the fitted constants).

In summary, our work demonstrates that structured and disordered regions likely contribute to the reversible aggregation of yeast proteins upon heat shock. Both disordered and structured regions of proteins have been previously implicated in irreversible protein aggregation. For example, yeast proteins with more disordered regions are more likely to aggregate irreversibly.[38] Also, several intrinsically disordered proteins are involved in misfolding diseases.[39] In the case of structured regions, it is shown that aggregation-prone regions of proteins are generally enriched in structured residues.[40] For example, irreversible aggregation of superoxide dismutase 1 in the progression of Amyotrophic Lateral Sclerosis (ALS) can be caused by mutations in the dimeric interface of this protein.[41–43] However, the interplay between both structured and disordered regions in phase separation has thus far only been reported for few proteins, including poly-adenylate binding protein (Pab1)[44] and ATP-binding RNA helicase (Ded1) in yeast,[45] as well as members of the RNA-binding FET family of mammalian proteins.[46] In the case of the well-known aggregator proteins Pab1, structured aggregation-prone regions form long loops or short secondary structure elements. Our observations suggest that these elements can readily unfold in response to stressors, form hydrophobic long chains, and phase-separate in a manner akin to amphiphilic molecules. Although further structural studies will be necessary to generalize these observations, we propose that the partial unfolding of a protein's tertiary structure is a critical stage in stress-induced phase separation. Indeed, the secondary structure of Pab1 remains largely unchanged and only its tertiary structure changes when it forms heat-induces condensates.[18,45] We also observed that aggregator proteins with longer APRs are more prone to aggregation. This relationship suggests that aggregator proteins may leave the aggregated phase by masking their APRs from self-assembly or from interaction with other proteins. Molecular chaperones can facilitate this step: They have been co-purified with stress-induced aggregated proteins and can solubilize them.[47]

We related the length of APRs in our aggregating proteins to the fraction of a protein that exists in the aggregated phase, using a simple two-state equilibrium model. This model neglects the presence of protein oligomers in reversible aggregation. Oligomers can influence the relationship between APR length and the aggregated protein fraction, if oligomers aggregate by a mechanism different from amphiphilic aggregation (See Supplementary information for details). Amyloid formation is such a mechanism, and can proceed through the formation of soluble oligomers, such as in the amyloidogenic proteins amyloid- $\beta$ , huntingtin, $\alpha$ -synuclein, and Sup35.[48–51] Detection of such oligomers during the reversible aggregation of yeast proteins would require further experiments, such as experiments with fluorescence based methods that can detect and characterize small protein aggregates.[52,53]

We found that the depletion of positively charged residues, and their dipeptides in APRs are the most important feature that distinguishes APRs from non-APR segments in our proteins. This observation further extends previous results that the most significant factor that separates soluble and insoluble proteins is a higher fraction of positively charged amino acids in soluble proteins.[54] We also observed that the fractions of aspartic acid and glutamic acid, the two amino acids with acidic side chains, as well as of methionine are significantly lower in APRs compared to non-APRs. Methionine is a hydrophobic residue that is often buried in protein hydrophobic cores.[55] We conjecture that the lower frequency of methionine in APRs compared to non-APRs is caused by the higher preference of aggregation-prone regions to be on a protein's surface. In addition, surface exposed methionine residues are more likely oxidized than buried methionines.[56] Oxidation changes the hydrophobic methionine side-chain to a polar side-chain by creating a sulfoxide group, and the resulting gain of hydrophilicity may inhibit protein aggregation.[56] Indeed, previous studies have shown that such decreased hydrophobicity by methionine oxidation can attenuate the aggregation of amyloid-$\beta$ (A$\beta$) peptide.[57–59] Aspartic acid and glutamic acid play a role in chaperone-independent control of protein aggregation, which might help explain their different incidence in APRs and non-APRs.[60] These acidic residues are the most potent aggregation breakers,[61,62] and their lower frequency in APRs might promote reversible protein aggregation in a cell. Testing these hypotheses is an important task for future computational and experimental work.

We also compared the APRs of our proteins with protein regions that are involved in irreversible protein aggregation and amyloid formation. We found two of our study protein in the CPAD 2.0 database, which contains information on experimentally validated amyloidogenic proteins.[63] These proteins are the [NU+] prion formation protein 1, and the translation termination factor Sup35. For both proteins, APRs and regions involved in irreversible aggregation did not overlap (Figure S2). Interestingly, the aggregation-prone regions involved in amyloid formation comprised the disordered N-terminal domains in both proteins (Figure S2). We further performed a random forest classification of APRs for our proteins (positive set), and of regions involved in irreversible aggregation for yeast proteins available in the CPAD 2.0 database (negative set). We used amino acid compositions as the classification features. The method was able to classify sequences well, with an accuracy of ~88% using 5-fold cross validation. The most important features for this classification were the fractions of asparagine and glutamine in a protein's amino acid sequence. Specifically, the aggregation-prone regions that form irreversible aggregates contain a higher fraction of asparagine and glutamine ($p < 10^{-16}$, Wilcoxon signed-rank test). This observation is consistent with a high incidence of glutamine/asparagine-rich domains in amyloid-forming proteins.[64–66] Overall, these analyses suggest that different sequence elements drive reversible and irreversible protein aggregation

Altogether, our study demonstrates that stress-induced phase separation is a sequence-encoded phenotype. It suggests that proteins may act like amphiphilic molecules when they reversibly self-assemble into separate phases.

## Methods

We identified aggregation-prone regions in the 96 aggregator proteins from the dataset of Cappelletti *et al.*,[22] and downloaded their sequences from the Uniprot database[67] (Dataset S13). We chose 500 amino acid properties from the AAindex database to distinguish the physicochemical properties of APRs from non-APR segments in the aggregator protein.[30] For mapping APRs onto the 3D structure of proteins, we used the structure of Pab1 as determined by electron microscopy (protein database (PDB) id: 6R5K).[27]

We used a binary classification to group regions within an aggregator protein into aggregation-prone regions (class A) and the remainder of the protein (class B). More specifically, we used a random forest classification, splitting aggregator proteins into a random subset comprising 80% of aggregator proteins for training and 20% for testing. To build features for the classification, we calculated the average value of 500 physicochemical properties for each sequence in the APR and non-APR dataset. This yielded two feature matrices for the APR and non-APR sequences. To apply random forest classification, we used the randomForest package in R,[68] and evaluated the best number of trees (*nTree*) and the number of variables randomly sampled at each split (*mtry*), in the random forest algorithm. To do so, we systematically varied the *nTree* and *mtry*, and calculated the accuracy of classification with 10-fold cross-validation and 3 repeats. We defined accuracy as the percentage of correctly identified classes of proteins (APRs and non-APRs) out of all instances. The combination of *nTree* = 5000 trees and *mtry* = 10 variables achieved the highest accuracy of ~90%. We then used these parameters to perform 100 random forest clusterings, in which we randomly assigned proteins to the training and the testing datasets. To quantify the accuracy of classification we counted the number of true positive and false positive predictions and calculated the area under the curve for them (AUC). These values are shown as a receiver operating characteristic curve (ROC) in Figure 2(D). The most important physicochemical properties were the ones whose Gini index in the classification decreased the most compared to all other properties. This index is a widely-used measure of dispersion that reflects inequality in the values of a frequency distribution. Within the random forest framework, this index is calculated as the average probability that each of 500 physicochemical properties wrongly classifies APRs and non-APRs in the random forest algorithm.

We also calculated different measures of classification performance using 5-fold cross-validation, with a data split of 75% for the training set and 25% for the test set. We counted the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FP) from the confusion matrix. We then calculated sensitivity (TP/(TP + FN)), specificity (TN/(TN + FN)), accuracy ((TP + TN)/(TP + TN + FP

+ FN)), and the area under the receiver operating characteristic curve (AUC). These quantities were equal to 0.90, 0.93, 0.92, and 0.92, respectively.

We used the g:profiler server for Gene Ontology enrichment analysis.[69] We performed all models and statistical analyses using R. Scripts are available at: https://github.com/dasmeh/yeast_aggregators.

## Funding

This project has received funding from the European Research Council under Grant Agreement No. 739874. We would also like to acknowledge support by Swiss National Science Foundation grant 31003A_172887 and by the University Priority Research Program in Evolutionary Biology at the University of Zurich.

## CRediT authorship contribution statement

**Pouria Dasmeh:** Conceptualization, Writing – review & editing, Methodology, Data curation, Visualization, Investigation. **Andreas Wagner:** Conceptualization, Writing – review & editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2021.167352.

*Received 11 May 2021;*
*Accepted 7 November 2021;*
*Available online 10 November 2021*

***Keywords***:
reversible aggregation;
stress-induced phase separation;
amphiphilic molecules

## References

1. Aguzzi, A., Calella, A.M., (2009). Prions: protein aggregation and infectious diseases. *Physiol. Rev.* **89**, 1105–1152.
2. Aguzzi, A., O'connor, T., (2010). Protein aggregation diseases: pathogenicity and therapeutic perspectives. *Nature Rev. Drug Discovery* **9**, 237–248.
3. Ross, C.A., Poirier, M.A., (2004). Protein aggregation and neurodegenerative disease. *Nature Med.* **10**, S10–S17.
4. Scheibel, T., Buchner, J., (2006). Protein aggregation as a cause for disease. *Mol. Chaperones Health Dis.*, 199–219.
5. Shulman, J.M., De Jager, P.L., Feany, M.B., (2011). Parkinson's disease: genetics and pathogenesis. *Annu. Rev. Pathol.* **6**, 193–222.
6. O'Connell, J.D. et al, (2014). A proteomic survey of widespread protein aggregation in yeast. *Mol. BioSyst.* **10**, 851–861.
7. Wallace, E.W. et al, (2015). Reversible, specific, active aggregates of endogenous proteins assemble upon heat stress. *Cell* **162**, 1286–1298.
8. Saad, S. et al, (2017). Reversible protein aggregation is a protective mechanism to ensure cell cycle restart after stress. *Nature Cell Biol.* **19**, 1202–1213.
9. Cereghetti, G., Saad, S., Dechant, R., Peter, M., (2018). Reversible, functional amyloids: towards an understanding of their regulation in yeast and humans. *Cell Cycle* **17**, 1545–1558.
10. Hyman, A.A., Weber, C.A., Jülicher, F., (2014). Liquid-liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.* **30**, 39–58.
11. Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., Sharp, P.A., (2017). A phase separation model for transcriptional control. *Cell* **169**, 13–23.
12. Erdel, F., Rippe, K., (2018). Formation of chromatin subcompartments by phase separation. *Biophys. J.* **114**, 2262–2270.
13. Valsecchi, C.I.K. et al, (2021). RNA nucleation by MSL2 induces selective X chromosome compartmentalization. *Nature* **589**, 137–142.
14. Youn, J.-Y. et al, (2019). Properties of Stress Granule and P-Body Proteomes. *Mol. Cell* **76**, 286–294.
15. Uversky, V.N., (2017). Intrinsically disordered proteins in overcrowded milieu: membrane-less organelles, phase separation, and intrinsic disorder. *Curr. Opin. Struct. Biol.* **44**, 18–30.
16. Posey, A.E., Holehouse, A.S., Pappu, R.V., (2018). Phase separation of intrinsically disordered proteins. *Methods in Enzymology*, **Vol. 611** Elsevier, pp. 1–30.
17. Kuechler, E.R., Budzyńska, P.M., Bernardini, J.P., Gsponer, J., Mayor, T., (2020). Distinct features of stress granule proteins predict localization in Membraneless organelles. *J. Mol. Biol.* **432**, 2349–2368.
18. Riback, J.A. et al, (2017). Stress-triggered phase separation is an adaptive, evolutionarily tuned response. *Cell* **168** 1028-1040.e19.
19. Iserman, C. et al, (2020). Condensation of Ded1p Promotes a Translational Switch from Housekeeping to Stress Protein Production. *Cell*.
20. Gao, X. et al, (2005). High-throughput limited proteolysis/mass spectrometry for protein domain elucidation. *J. Struct. Funct. Genomics* **6**, 129–134.
21. Schopper, S. et al, (2017). Measuring protein structural changes on a proteome-wide scale using limited proteolysis-coupled mass spectrometry. *Nature Protoc.* **12**, 2391.
22. Cappelletti, V. et al, (2021). Dynamic 3D proteomes reveal protein functional alterations at high resolution in situ. *Cell* **184** 545–559.e22.
23. Dosztányi, Z., (2018). Prediction of protein disorder based on IUPred. *Protein Sci.* **27**, 331–340.
24. Emenecker, R.J., Griffith, D., Holehouse, A.S., (2021). metapredict: a fast, accurate, and easy-to-use cross-platform predictor of consensus disorder. *bioRxiv*.

25. Bonferroni, C.E., Bonferroni, C. & Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita'.
26. Sugimoto, K., Matsumoto, K., Kornberg, R.D., Reed, S.I., Wittenberg, C., (1995). Dosage suppressors of the dominant G1 cyclin mutant CLN3-2: identification of a yeast gene encoding a putative RNA/ssDNA binding protein. *Mol. General Genetics MGG* **248**, 712–718.
27. Schäfer, I.B. et al, (2019). Molecular basis for poly (A) RNP architecture and recognition by the Pan2-Pan3 deadenylase. *Cell* **177** 1619–1631.e21.
28. Kelil, A., Dubreuil, B., Levy, E.D., Michnick, S.W., (2017). Exhaustive search of linear information encoding protein-peptide recognition. *PLoS Comput. Biol.* **13**, e1005499
29. Ho, T.K., (1995). Random decision forests. *Proceedings of 3rd international conference on document analysis and recognition*, **Vol. 1** IEEE, pp. 278–282.
30. Kawashima, S., Kanehisa, M., (2000). AAindex: amino acid index database. *Nucleic Acids Res.* **28**, 374.
31. Mitaku, S., Hirokawa, T., Tsuji, T., (2002). Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane–water interfaces. *Bioinformatics* **18**, 608–616.
32. Zhang, X., Wang, C., (2011). Supramolecular amphiphiles. *Chem. Soc. Rev.* **40**, 94–101.
33. Shaul, B., Gelbart, W., (1985). Theory of chain packing in amphiphilic aggregates. *Annu. Rev. Phys. Chem.* **36**, 179–211.
34. Zhu, M. et al, (2020). Proteomic analysis reveals the direct recruitment of intrinsically disordered regions to stress granules in S. cerevisiae. *J. Cell Sci.* **133**
35. Narayanaswamy, R. et al, (2009). Widespread reorganization of metabolic enzymes into reversible assemblies upon nutrient starvation. *Proc. Natl. Acad. Sci.* **106**, 10147–10152.
36. Budin, I., Prywes, N., Zhang, N., Szostak, J.W., (2014). Chain-length heterogeneity allows for the assembly of fatty acid vesicles in dilute solutions. *Biophys. J.* **107**, 1582–1590.
37. Ruckenstein, E., Nagarajan, R., (1980). Aggregation of amphiphiles in nonaqueous media. *J. Phys. Chem.* **84**, 1349–1358.
38. Uemura, E. et al, (2018). Large-scale aggregation analysis of eukaryotic proteins reveals an involvement of intrinsically disordered regions in protein folding. *Sci. Rep.* **8**, 1–11.
39. Uversky, V.N., (2015). Intrinsically disordered proteins and their (disordered) proteomes in neurodegenerative disorders. *Front. Aging Neurosci.* **7**, 18.
40. Buck, P.M., Kumar, S., Singh, S.K., (2013). On the role of aggregation prone regions in protein evolution, stability, and enzymatic catalysis: insights from diverse analyses. *PLoS Comput. Biol.* **9**, e1003291
41. Alemasov, N.A., Ivanisenko, N.V., Ramachandran, S., Ivanisenko, V.A., (2018). Molecular mechanisms underlying the impact of mutations in SOD1 on its conformational properties associated with amyotrophic lateral sclerosis as revealed with molecular modelling. *BMC Struct. Biol.* **18**, 1–14.
42. Dasmeh, P., Kepp, K.P., (2017). Superoxide dismutase 1 is positively selected to minimize protein aggregation in great apes. *Cell. Mol. Life Sci.*, 1–15.
43. Broom, H.R., Rumfeldt, J.A., Vassall, K.A., Meiering, E.M., (2015). Destabilization of the dimer interface is a common consequence of diverse ALS-associated mutations in metal free SOD1. *Protein Sci.* **24**, 2081–2089.
44. Drummond, D.A., Wilke, C.O., (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352.
45. Iserman, C. et al, (2020). Condensation of Ded1p promotes a translational switch from housekeeping to stress protein production. *Cell* **181** 818–831.e19.
46. Martin, E.W. et al, (2021). Interplay of folded domains and the disordered low-complexity domain in mediating hnRNPA1 phase separation. *Nucleic Acids Res.* **49**, 2931–2945.
47. Cherkasov, V. et al, (2013). Coordination of translational control and protein homeostasis during severe heat stress. *Curr. Biol.* **23**, 2452–2462.
48. Lasmézas, C.I. et al, (1997). Transmission of the BSE agent to mice in the absence of detectable abnormal prion protein. *Science* **275**, 402–404.
49. Conway, K.A., Harper, J.D., Lansbury, P.T., (1998). Accelerated in vitro fibril formation by a mutant α-synuclein linked to early-onset Parkinson disease. *Nature Med.* **4**, 1318–1320.
50. Tzaban, S. et al, (2002). Protease-sensitive scrapie prion protein in aggregates of heterogeneous sizes. *Biochemistry* **41**, 12868–12875.
51. Dulle, J.E., Bouttenot, R.E., Underwood, L.A., True, H.L., (2013). Soluble oligomers are sufficient for transmission of a yeast prion but do not confer phenotype. *J. Cell Biol.* **203**, 197–204.
52. De, S., Klenerman, D., (2019). Imaging individual protein aggregates to follow aggregation and determine the role of aggregates in neurodegenerative disease. *Biochim. Biophys. Acta (BBA)-Proteins Proteomics* **1867**, 870–878.
53. Orte, A. et al, (2008). Direct characterization of amyloidogenic oligomers by single-molecule fluorescence. *Proc. Natl. Acad. Sci.* **105**, 14424–14429.
54. Chan, P., Curtis, R.A., Warwicker, J., (2013). Soluble expression of proteins correlates with a lack of positively-charged surface. *Sci. Rep.* **3**, 1–6.
55. Brosnan, J.T., Brosnan, M.E., Bertolo, R.F., Brunton, J.A., (2007). Methionine: a metabolically unique amino acid. *Livestock Science* **112**, 2–7.
56. Kim, G., Weiss, S.J., Levine, R.L., (2014). Methionine oxidation and reduction in proteins. *Biochim. Biophys. Acta (BBA)-General Subjects* **1840**, 901–905.
57. Johansson, A.-S. et al, (2007). Attenuated amyloid-β aggregation and neurotoxicity owing to methionine oxidation. *NeuroReport* **18**, 559–563.
58. Palmblad, M., Westlind-Danielsson, A., Bergquist, J., (2002). Oxidation of methionine 35 attenuates formation of amyloid β-peptide 1–40 oligomers. *J. Biol. Chem.* **277**, 19506–19510.
59. Bettinger, J., Ghaemmaghami, S., (2020). Methionine oxidation within the prion protein. *Prion* **14**, 193–205.
60. Houben, B. et al, (2020). Autonomous aggregation suppression by acidic residues explains why chaperones favour basic residues. *EMBO J.* **39**, e102864
61. Nadimpally, K.C., Paul, A., Mandal, B., (2014). Reversal of aggregation using β-breaker dipeptide containing peptides: application to Aβ (1–40) self-assembly and its inhibition. *ACS Chem. Neurosci.* **5**, 400–408.
62. Minicozzi, V. et al, (2014). Computational and experimental studies on β-sheet breakers targeting Aβ1–40 fibrils. *J. Biol. Chem.* **289**, 11242–11252.

63. Rawat, P. et al, (2020). CPAD 2.0: a repository of curated experimental data on aggregating proteins and peptides. *Amyloid* **27**, 128–133.
64. Zhang, Y., Man, V.H., Roland, C., Sagui, C., (2016). Amyloid properties of asparagine and glutamine in prion-like proteins. *ACS Chem. Neurosci.* **7**, 576–587.
65. Michelitsch, M.D., Weissman, J.S., (2000). A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions. *Proc. Natl. Acad. Sci.* **97**, 11910–11915.
66. Dasmeh, P., Wagner, A., (2021). Natural selection on the phase-separation properties of FUS during 160 My of mammalian evolution. *Mol. Biol. Evol.* **38**, 940–951.
67. U. Consortium, (2015). UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212.
68. Liaw, A., Wiener, M., (2002). Classification and regression by randomForest. *R news* **2**, 18–22.
69. Raudvere, U. et al, (2019). g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198.