Computational identification of prion-like RNA-binding proteins that form liquid phase-separated condensates

Orlando, Gabriele; Raimondi, Daniele; Tabaro, Francesco; Codicé, Francesco; Moreau, Yves; Vranken, Wim

Published in:
Bioinformatics

DOI:
10.1093/bioinformatics/btz274

Publication date:
2019

Bioinformatics, 35(22), 2019, 4617–4623
doi: 10.1093/bioinformatics/btz274
Advance Access Publication Date: 17 April 2019
Original Paper

OXFORD

Sequence analysis

Computational identification of prion-like RNA-binding proteins that form liquid phase-separated condensates

Gabriele Orlando[1,2,†], Daniele Raimondi[3,†], Francesco Tabaro[4], Francesco Codicè[5], Yves Moreau[3] and Wim F. Vranken[1,2,6,*]

[1]Interuniversity Institute of Bioinformatics in Brussels, ULB-VUB, 1050 Brussels, Belgium, [2]Structural Biology Brussels, Department of Bioengineering Sciences, Vrije Universiteit Brussel, 1050 Brussels, Belgium, [3]ESAT-STADIUS, KU Leuven, Leuven 3001, Belgium, [4]Institute of Biosciences and Medical Technology, Tampere 33520, Finland, [5]Department of Computer Science and Engineering, University of Bologna, Bologna 40127, Italy and [6]Center for Structural Biology, VIB, 1050 Brussels, Belgium

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: John Hancock

Received on December 14, 2018; revised on April 6, 2019; editorial decision on April 8, 2019; accepted on April 12, 2019

Abstract

Motivation: Eukaryotic cells contain different membrane-delimited compartments, which are crucial for the biochemical reactions necessary to sustain cell life. Recent studies showed that cells can also trigger the formation of membraneless organelles composed by phase-separated proteins to respond to various stimuli. These condensates provide new ways to control the reactions and phase-separation proteins (PSPs) are thus revolutionizing how cellular organization is conceived. The small number of experimentally validated proteins, and the difficulty in discovering them, remain bottlenecks in PSPs research.

Results: Here we present PSPer, the first *in-silico* screening tool for prion-like RNA-binding PSPs. We show that it can prioritize PSPs among proteins containing similar RNA-binding domains, intrinsically disordered regions and prions. PSPer is thus suitable to screen proteomes, identifying the most likely PSPs for further experimental investigation. Moreover, its predictions are fully interpretable in the sense that it assigns specific functional regions to the predicted proteins, providing valuable information for experimental investigation of targeted mutations on these regions. Finally, we show that it can estimate the ability of artificially designed proteins to form condensates ($r = -0.87$), thus providing an *in-silico* screening tool for protein design experiments.

Availability and implementation: PSPer is available at bio2byte.com/psp.

Contact: wim.vranken@vub.be

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Eukaryotic cells are divided into different membrane-delimited compartments, whose varying degrees of compound selectivity supports a plethora of biochemical reactions (Nott *et al.*, 2015; Weber and Brangwynne, 2012). The universe of known cellular compartments recently experienced a drastic expansion with the discovery of membraneless organelles, termed biomolecular condensates (Wang *et al.*, 2018) or ribonucleoprotein granules (Lin *et al.*, 2015; Weber and

Brangwynne, 2012; Youn *et al.*, 2018). The cell can trigger the temporary formation of these condensates to provide fast responses to a wide range of stimuli (Banani *et al.*, 2017; Shin and Brangwynne, 2017; Weber and Brangwynne, 2012). To do so, it exploits the concentration-driven liquid–liquid phase-separation (LLPS) properties of some proteins (Wang *et al.*, 2018), enabling functionality similar to membrane-bound organelles, but without the burden of building and preserving an actual membrane (Banani *et al.*, 2016; Feric *et al.*, 2016). While our knowledge on membraneless organelles is evolving extremely fast, information on the prevalence and spread of phase-separation proteins (PSP) is lacking. This creates a bottleneck for further experimental research, which can only be solved by tools for the rapid *in-silico* screening of proteomes. The small number of available PSPs, however, impedes the application of *supervised* machine learning approaches to identify putative protein candidates. A recent thorough analysis pinpointed common residue-level determinants of the LLPS behavior of 22 proteins belonging to the FUS family (Wang *et al.*, 2018). Although these proteins are mentioned as ‘family’, they have variable domain organization (Wang *et al.*, 2018) composed by various arrangements of long disordered regions and regions capable of non-specific RNA interactions (Boeynaems *et al.*, 2018; Lin *et al.*, 2015; Weber and Brangwynne, 2012). As shown in Supplementary Section S1, these proteins generally do not share evolutionary relationships and sequence similarity: they are often so diverse that it is likely that they arose from convergent evolution.

Analyzing these proteins, the authors of Wang *et al.* (2018) identified three key regions for LLPS behavior in FUS-like PSPs: Prion-like domains (PLD), RNA-recognition motifs (RRMs) and disordered, arginine (Arg) rich regions that we here call Spacers. Additionally, there may be other regions or domains with low impact on the LLPS behavior (Other). Wang *et al.* showed that all these regions have peculiar biophysical characteristics, and they defined conceptual *biophysical rules* governing LLPS behavior (Wang *et al.*, 2018) in FUS proteins. For example, the PLD is a disordered region with low sequence complexity (SC) (e.g. only few types of residues are present) with an enrichment of tyrosines (Tyrs), while the RNA-binding Domain (RBD) region contains one or more structured RRM separated by Spacers. The interaction between the Tyr in the PLD and the Arg in the Spacers has shown to be crucial for LLPS in FUS-like proteins. In this paper, we present the first *in-silico* method for the detection of such FUS-like PSPs. Our approach, called PSPer, uses an unsupervised, rule-based, mathematical model that hardwires the residue-level and domain-level characteristics of the proteins described in Wang *et al.* (2018) as biophysical rules into a flexible probabilistic model. The model is not limited to the detection of members of the ‘FUS family’, as it has been built to recognize all the proteins with similar LLPS behavior (FUS-like PSPs), which are expected to be a not negligible portion of eukaryote proteomes. Moreover, the main goal of this study is not to build a biophysical model for the LLPS behavior *per se*, but to allow the prioritization of the most likely FUS-like PSPs among candidates for *in-vitro* experiments, thus helping the experimental discovery and validation of such proteins. PSPer can indeed rapidly and accurately screen entire proteomes to identify proteins that have the potential to form condensates. We show it can prioritize FUS-like PSPs among random and selected proteins containing similar RBDs, intrinsically disordered regions and prions, even though the resemblance to the FUS family goes beyond homology or sequence similarity between proteins. We screen proteomes *in-silico* to identify the most likely FUS-like PSPs for further experimental investigation, and show that the prediction scores correlate with the saturation concentration for condensate formation.

Table 1. The expected properties of the FUS-like PSP regions, which are translated into emission probabilities for the correspondent states of the HMM

States	Arg	Tyr	SC	Disorder	RRM
PLD	Low	High	Low	High	Low
Spacer	High	Low	High	High	Low
RRM	Low	Low	High	Low	High
Other	Low	Low	High	Low	Low

Note: The columns represent for each region the local Arg enrichment, the local Tyr enrichment, the local SC, the disorder and the probability of belonging to a RRM.

Moreover, the tool annotates the functional role of the PSP-relevant regions, providing means to interpret the prediction results, suggesting to experimentalists where mutations are more likely to affect the LLPS behavior. The method is available at http://bio2byte.com/psp.

2 Materials and methods

2.1 Datasets

Although our model is rule-based and unsupervised, in this study we used several datasets to analyze the performances and to investigate the predictions obtained. These datasets are described in Supplementary Section S5 and we briefly summarize them here. The Background dataset (BGD) contains 5000 non-redundant proteins from Uniprot. The RNA-binding proteins (RBPs) dataset contains 1724 non-redundant RBPs extracted from Uniprot. The Prions dataset contains 27 non-redundant prions and the Disordered dataset contains intrinsically disordered proteins (IDPs) taken from Walsh *et al.* (2012). The 144 granule forming proteins have been extracted from Youn *et al.* (2018). The 22 ‘FUS family’ PSPs and the 16 proteins with corresponding experimentally determined saturation concentration have been obtained from Wang *et al.* (2018).

2.2 The molecular rules driving LLPS in FUS-like proteins

We used the empirical rules $\mathcal{R}$ described in Wang *et al.* (2018) to create a probabilistic model for the identification of putative FUS-like PSPs. These rules list the possible domain arrangements within this class of proteins, with the respective biophysical characteristics. They can be summarized in the following way:

- FUS-like PSPs always contain a PLD (Hennig *et al.*, 2015) and an RBD.
- PLDs have low SC, are disordered (Disorder) and are enriched in Tyrs; each protein has only one PLD.
- RBDs contain at least one RRM (Pfam: PF00076), and are separated by disordered ‘Spacer’ regions rich in Args. Such linkers may be present elsewhere in the protein. Our use of the Spacer name is different from the ‘stickers vs. spacers’ mentioned in Wang *et al.* (2018), since they refer to specific residues in the context of the associative polymers theory.
- FUS-like PSPs may contain additional, unspecified domains or motifs irrelevant for condensate formation (designated as ‘Other’ regions).

From these rules we thus summarize which biophysical characteristics belong to each region, as shown in Table 1, which shows the four regions identified by the rules $\mathcal{R}$ and their expected characteristics.

2.3 A probabilistic description of FUS-like PSPs regions

The biophysical characteristics corresponding to the columns of Table 1 can be translated in mathematical terms by using standard sequence analysis approaches to infer probability distributions describing these aspects on the general population of proteins. We estimated these distributions in the following way from the BGD:

- *Local abundance of Arg and Tyr residues:* We ran a sliding window of 11 residues along all the proteins in the BGD, computing the distribution of the content of Arg and Tyr residues in the windows. Normalizing for the total number of observed windows, we obtained two discrete distributions corresponding to the probability of observing, in a window of 11 residues, a certain number of Arg or Tyr residues. We refer to these distributions as $R(x_i)$ and $Y(x_i)$, where x_i is respectively the number of Arg and Tyr in the window centered in position *i*.
- *Local SC:* Using the same sliding window technique on the BGD, we computed the distribution of the SC (which is the number of different residue types in each sliding window) over 11-residue windows. SC is a commonly used property of protein sequences that is known to correlate with their intrinsic disorder (Romero *et al.*, 2001). We called this distribution $SC(x_i)$, where x_i is the SC in the window centered in *i*.
- *Disorder prediction (Disorder):* We computed the predicted disorder profile for each protein in the BGD using Disomine, a neural network-based in-house disorder predictor (under review, available at http://bio2byte.ibsquare.be/disomine/). We discretized the probability-like Disomine predictions into 10 bins. We refer to this distribution as $D(x_i)$, where x_i is the predicted disorder at position *i*.
- *RRM probability (RRM):* We used HMMer (Finn *et al.*, 2011) to determine the presence of the RRM described by the PFAM family PF00076 (Finn *et al.*, 2014; Wang *et al.*, 2018) on each target sequence *S*. For each residue $j \in S$, HMMer provides the score (from 0 to 10) that *j* belongs to PF00076. We consider the scores as bins of a discrete distribution were the probability *j* being part of a RRM increases linearly. We call this distribution $M(x_i)$, where x_i is the probability provided by hmmer of residue *i* being part of an RRM.

2.4 From molecular rules to an HMM-like model

We encoded the empirical rules described in Table 1 within a mathematical framework, so creating a Hidden Markov Model (HMM) with a relaxed definition of emission probabilities and Markov property. We hard-encoded the domain organization of FUS-like PSPs in the model topology, ensuring that all the needed regions are present and in an allowed position. Our approach is unsupervised because only the knowledge-based rules $\mathcal{R}$ have been used to define the HMM-like model, without using any data to fit it. The model contains 16 logical states (see Supplementary Fig. S4) representing instances of the four conceptual regions (RRM, PLD, Spacer and Other). The RBD is represented as meta-state involving alternations of RRM and Spacers (see red boxes in Supplementary Fig. S1), and the transition probabilities have been set according to the expected domain lengths and abundance (see Supplementary Section S3). The architecture of our HMM-like model is shown in Supplementary Figure S1.

HMMs are particularly suitable for the task presented in this paper because they are well-known models that can natively deal with protein sequences. Several optimal algorithms have been already developed to score sequences and to identifying different regions within them, thus allowing a deeper understanding of the probabilistic score prediction, considering the sequence as a whole. Last, with minimal changes to the standard HMM framework, we could implement an unsupervised model based on biophysically-derived empirical rules, which helped us overcome the data scarcity in the PSPs field.

2.5 Emission probabilities as deviations from background

Since the rules $\mathcal{R}$ are qualitative and not enough data is available for properly training the parameters, emission probabilities were calculated in an unsupervised way by exploiting the unique biophysical characteristics of FUS-like PSPs with respect to the rest of the protein realm. For each biophysical characteristic described in the rules and summarized in Table 1, we defined 'high' as '*higher than what is expected from the background*' and 'low' as '*lower than what is expected from the background*'.

Within our HMM-like model, this is translated into mathematical terms by using the normalized cumulative distributions $F_B(x_i)$ of the distributions $B = \{R, Y, SC, D\}$, learned on the BGD (see Section 2.3). These distributions correspond to the Arg and Tyr enrichment, the SC and the disorder, plus the normalized probability *M*, obtained from HMMer, that a residue belongs to a RRM. x_i indicates the value of characteristic *B* at the position *i* in the protein S under scrutiny.

If a biophysical characteristic *B* in a certain PSP region is supposed to be *high* (see Table 1), we describe its emission probability as $F_B(x_i)$. On the other hand, a characteristic *B* which is supposed to be low is described by $1 - F_B(x_i)$.

For example, the emission probability for the PLD state (see first row in Table 1) is obtained by implementing a logic AND between the normalized cumulative distributions as: $(1 - F_R(x_i)) \times F_Y(x_i) \times F_D(x_i) \times (1 - F_{SC}(x_i)) \times (1 - M(x_i))$, where *i* is the current position in the protein under scrutiny *S*. In this way, the emission for the PLD state will be high if the number of Arg in around *i* is low *and* the number of Tyr residues around *i* is high *and* disorder x_i is high *and* complexity around *i* is low *and* residue *i* does not belongs to an RRM, as evinced from the *R*, *Y*, *D* and *CS* cumulative background distributions and from the HMMer predictions. These cumulative distributions are shown in Supplementary Figure S5 and more details are provided in Supplementary Section S4.1.

The final emission probability of a state is obtained from the multiplication of the probabilities coming from these five distributions, as summarized in Table 1. This strategy allows us to define emission probabilities based on empirical rules related to biophysical aspects of FUS-like PSPs, similar to *soft constraints* that each region should respect. Our HMM-like model uses the Backward algorithm to quantify how well each target protein *fits* in these rules and uses this score to predict the likelihood of LLPS behavior. The fact that we always multiply the same number of factors ensures that every state has emission scores with the same scale.

2.6 Computing the prediction scores

To predict the score of a protein to exhibit phase-separation behavior, we then compute the backward score from the HMM-like model, which provides a value indicating how likely it is that the observed sequence is generated by the model. For the same sequence, we compute also the likelihood of it being generated by a dummy

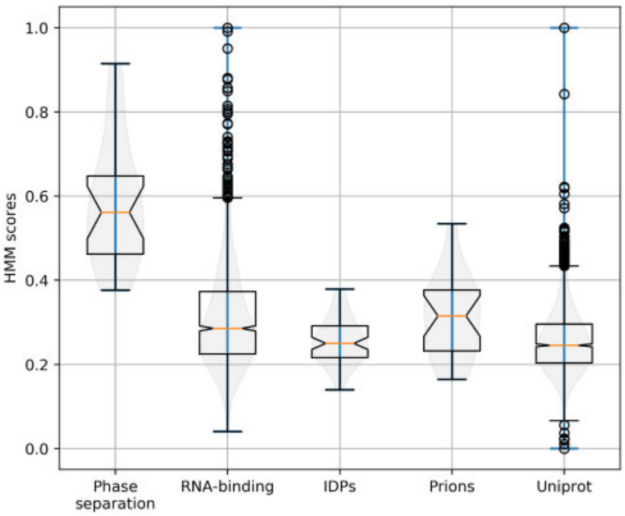

Fig. 1. Boxplots showing the distributions of the HMM scores for FUS-like PSPs, RBPs, IDPs and randomly selected proteins from Uniprot

model where the only possible hidden state is ‘Other’ (so the expected distributions of the background) and we then compute the log-odds between the two. Finally, we apply a min–max scaling to these values to provide probability-like scores, which facilitate the interpretation of the results. Supplementary Section S6 shows that similar results can be achieved with a Bayesian treatment of the scores.

Although the model cannot be considered to be truly a HMM, it is still expressed as a product of event scores. As a result, the Viterbi algorithm can still be used to find the highest scoring state path. Similarly, the forward and backward algorithm can be used to compute the sum of scores of a sequence over all possible paths, which can be considered more informative than the score of the sequence and its single best path.

2.7 The Viterbi decoding can annotate the functional regions on the target proteins

One of the advantages of using generative models is that they can help the interpretation of the results obtained. In the case of PSPer, we used the Viterbi algorithm to determine the most likely sequence of hidden states that could have generated each sequence under scrutiny. With this procedure we can pinpoint the location of the biologically relevant hidden states, namely the PLD, RRM and spacers, on the input sequence, providing valuable information to the experimentalists willing, e.g. to investigate the effect of targeted mutations on the proteins under scrutiny.

3 Results

3.1 Discriminating FUS-like PSPs from other proteins

To investigate the ability of our model to prioritize FUS-like PSPs among sets of proteins with different characteristics, we first showed that our tool ranks the 22 PSPs from Wang *et al.* (2018) significantly higher (Wilcoxon rank sums $=1.7 \times 10^{-15}$) than 2000 randomly selected non-redundant proteins extracted from SwissProt (Apweiler *et al.*, 2004), labeled as ‘Uniprot’ in Figure 1. Since this prioritization may present a simplified challenge for our method, we selected three classes of proteins, generic RBPs, IDPs and proteins annotated as Prions in UniprotKB (Apweiler *et al.*, 2004) (see Section 2 and Supplementary Section S5), that contain regions which are also present in FUS-like PSPs and therefore present a greater challenge for our method. The difference between the PSP and these datasets are significant, with *P*-value $=1.2 \times 10^{-12}$ for RBPs, *P*-value $=2.6 \times 10^{-12}$ for IDPs and *P*-value $=3.5 \times 10^{-8}$ for prions (Fig. 1). The 22 PSPs from Wang *et al.* (2018) are ranked higher than the proteins in these sets, with the exception for few outliers.

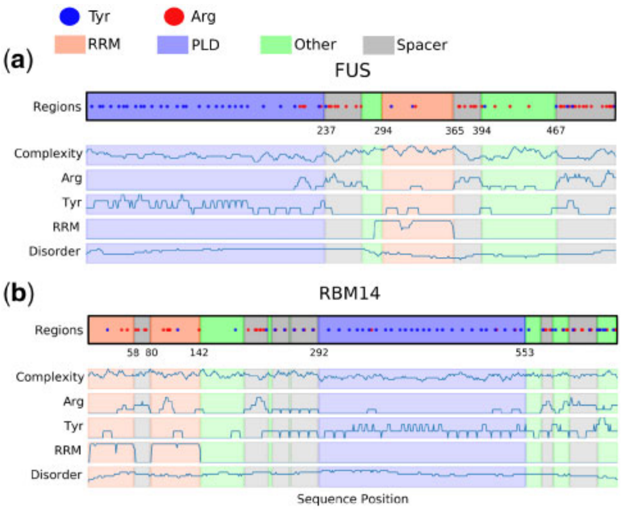

Fig. 2. Predicted regions of FUS (**a**) and RBM14 (**b**). The RRM (red), PLD (blue) and spacer (grey) regions are indicated, as well as areas of the protein with likely a limited impact on phase separation (green). Blue and red dots highlight the Tyrs and Args in the sequence. For each protein we also report the observed features: SC, Arg enrichment, Tyr enrichment, RRM likelihood and disorder

Nevertheless, the true negatives are not known, and each of the ‘test’ datasets in Figure 1 may contain few FUS-like PSPs proteins by chance, so the above procedure is not a conventional validation in that sense. This is reflected by the top hits in the RBPs dataset: the first hit is SRRM1 (Uniprot ID: Q52KI8, score: 1.0), for which a recent study (Rai *et al.*, 2018) describes how this protein is observed to be involved in the formation of liquid condensates. The third hit (Q14152, 0.95) is implicated in yeast stress granules (Grousl *et al.*, 2009; Rinnerthaler *et al.*, 2013), while the eighth hit (Q8C2Q3, 0.85) has the PLD-RRMs domain architecture and can form hydrogels (Hennig *et al.*, 2015). In the Uniprot dataset, the first hit is an adhesive protein used by mussels to cling to surfaces (Q25434, 1.0), which is known to undergo liquid–liquid phase separation (Waite, 2017). The full list of predictions is available in Supplementary Section S5.

3.2 PSPer assigns regions with key roles in FUS-like PSPs

Two examples illustrate how PSPer can provide molecular biologists with insights into the putative PSP sequences under scrutiny (Fig. 2a). In the FUS protein (P35637, score 0.91), PSPer identifies an initial PLD of 237 residues (blue), an RRM region from position 294–365 (red), and Other (green) and Spacer (grey) regions throughout the protein. This matches the annotations from Wang *et al.* (2018), as well as those from Uniprot. The ‘Other’ states are assigned to regions with biophysical characteristics similar to the background distributions. These regions are therefore not necessarily independent domains, but they should be less relevant for the formation of the condensates.

The second example, shown in Figure 2b, is the human RBM14 protein (Uniprot ID: Q96PK6, score 0.90), which has been

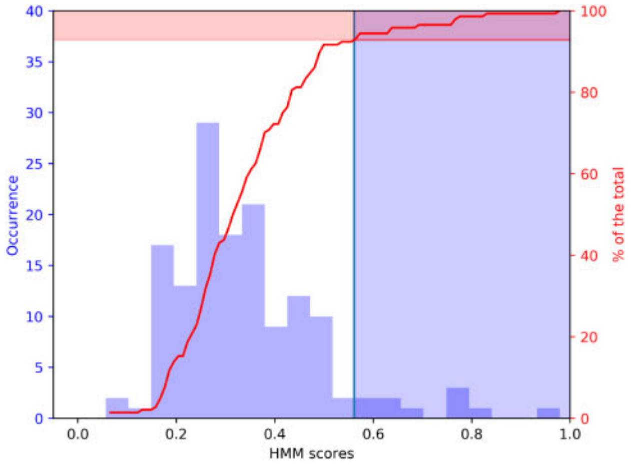

Fig. 3. The distribution of PSPer scores on 144 proteins found in cytosolic granules in Youn *et al.* (2018a). Only nine of these, or 6.3%, have scores >0.56, indicating that specific proteins are the drivers for the formation of these condensates

experimentally shown to exhibit concentration-dependent phase-separation behavior (Courchaine and Neugebauer, 2015; Hennig *et al.*, 2015). This protein is not present in the 22 proteins studied in Wang *et al.* (2018), and PSPer identifies two RRM regions, 0–50 and 81–140, as annotated in InterPro on Uniprot. The PLD is predicted as part of the largest low complexity region of the protein (position 224–599).

3.3 Predicted prevalence of FUS-like PSPs in cystosolic granules

A recent study (Youn *et al.*, 2018) identified 144 yeast proteins as putative core components of stress granule and P-body condensates. Despite such efforts, the number and type of proteins that actively promote phase separation is still unclear. In particular, it is not known if this emergent behavior derives from cooperation between most of the proteins in the organelles, or if specialized scaffold proteins are responsible (Weber and Brangwynne, 2012; Youn *et al.*, 2018). The most recent hypothesis implies that the remaining proteins, called client proteins, have co-evolved with their scaffold partners in order to be recruited in the membraneless organelles and correctly perform their task (Wang *et al.*, 2018; Youn *et al.*, 2018). The distribution of the PSPer scores on the 144 proteins from Youn *et al.* (2018) is shown in Figure 3. The light blue area contains scores >0.56, which is the median of the 22 PSPs scores in Figure 1, and thus comprises the most likely FUS-like PSPs. Only 10 proteins (6.9%) fall in this this region, with the second ranked protein (Q9Y520, score 0.83) reported to be critical for the formation of stress granules (Youn *et al.*, 2018). If PSPer indeed identifies most of the FUS-like PSPs in this dataset, then our results support the hypothesis of an heterogeneous composition of these condensates, where few scaffold proteins are responsible for the formation and maintenance of the membraneless organelles, while the others are client proteins, that use the droplets to perform their task, but do not contribute to its formation (see Supplementary Section S5.5 for the full list of proteins). Interestingly, when we analyze with GOrilla (Eden *et al.*, 2009) which GO-terms are over-represented in the top-ranked proteins, then they emerge as significantly related to processes involved in mRNA-preprocessing and splicing regulation (P-values between $< 10^{-3}$ and $< 10^{-4}$).

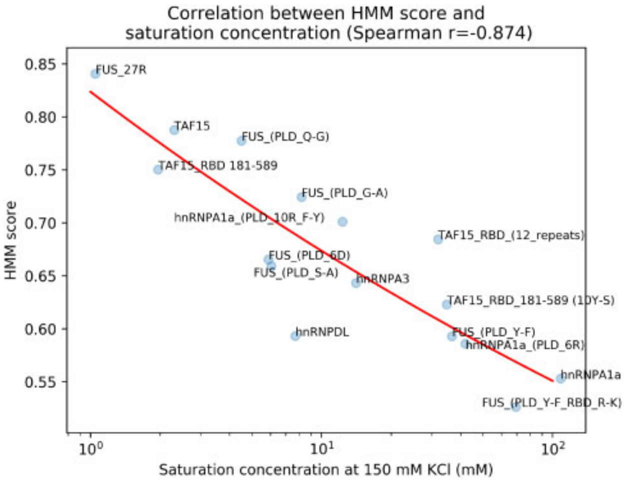

Fig. 4. Correlation between the PSPer score and the saturation concentration at 150 mM of salt concentration for 16 PSPs

3.4 Predicted FUS-like prevalence among proteomes

Supplementary Figure S3 shows the PSPer scores for the proteome of five well studied model organisms (*Saccharomyces cerevisiae*, *Homo sapiens*, *Arabidopsis thaliana*, *Drosophila melanogaster* and *Caenorhabditis elegans*). The distributions are quite similar, with each proteome presenting 0.2–1% of proteins scored higher than 0.56, which is the median of the scores of the 22 FUS PSPs. To identify GO term enrichment with GOrilla (Eden *et al.*, 2009) on these full proteomes, we first tried to reduce bias by removing all proteins not annotated as 'RNA binding proteins' in Uniprot (keyword: KW-0694) and then applying a 20% sequence identity filter on the remaining proteins. All organisms, except for *S.cerevisiae*, show significant enrichment of GO-terms related to splicing and mRNA processing (P-values between $< 10^{-3}$ and $< 10^{-14}$) for the proteins with high PSPer scores. The list of identified proteins is available in Supplementary Material to enable their experimental validation.

3.5 PSPer predictions correlate with experimentally determined saturation concentrations

We retrieved the saturation concentration of 16 FUS-like PSPs from Wang *et al.* (2018). All the selected proteins have been tested in the same experimental condition, thus providing an unbiased indication of their LLPS propensity. For a detailed description of the experimental procedures and conditions, refer to Wang *et al.* (2018). We compared these concentrations to the corresponding PSPer scores (Fig. 4). The Spearman correlation between them is −0.87, with a P-value of 9.87×10^{-6}. This is a surprising result, as our model was developed completely blind to these concentrations: the PSPer score purely reflects the presence of regions with particular biophysical characteristics, as well as how strongly the regions reflect those characteristics. The correlation therefore indicates that, if the right regions are present, the ease with which phase separation occurs purely depends on the biophysical 'typicality' of these regions. Moreover, PSPer seems to detect quite subtle changes in the ability to form condensates among slightly different versions of the same protein. In particular, the proteins listed in Figure 4 contain different versions of the FUS and TAF15 proteins, and in Wang *et al.* (2018) the authors showed that mutations of residues in the PLD or RBD regions of FUS-like PSPs can influence the ability of the proteins to form condensates. The high correlation between PSPer scores and the experimental saturation concentration indicates that our model can identify mutations likely to improve or reduce the concentration of protein necessary to trigger the liquid–liquid phase separation, opening possibilities for the *in-silico* design of FUS-like PSPs with tuned properties.

4 Discussion

4.1 Prioritizing FUS-like candidates with PSPer

We here present PSPer, the first *in-silico* method for the identification of proteins driving the formation of cytosolic condensates acting as membraneless organelles. PSPer relies on an unsupervised method that is based on a set of simple empirical rules extracted from Wang *et al.* (2018) for 22 PSPs of the FUS family. Despite the name, the FUS proteins often have no evolutionary relationship (see Supplementary Section S1 for more details), and the arrangement of their PLD and RBD domains can be drastically different (Wang *et al.*, 2018), requiring a specific approach to describe their common traits. PSPer can natively deal with changes in domain organization, and can precisely indicate the location of the PLD, RRM and Spacer regions crucial for phase-separation behavior in the target proteins. This level of semantic annotation surpasses the computational analysis performed so far in literature. For example, in Wang *et al.* (2018) the authors based their model on the observed frequencies of Tyr and Arg residues, (i) without incorporating aspects related to the modular nature of FUS-like PSPs into their model and (ii) without being able to annotate the protein sequences besides the enrichment for these two amino acids.

The architecture and the interchangeability of the PLD, RRM, Spacer and Other regions in FUS-like PSPs is quite unusual (Weber and Brangwynne, 2012); e.g. FUS has a PLD-Spacer-RRM-Spacer-Other-Spacer, while hnRNPH1 has the RBD first (with three RRMs) and the PLD after. The ability of our method to deal with this variable organization is essential to recognize FUS-like PSPs and it is not a trivial task for classical homology detection or alignment methods, such as HMMer (Finn *et al.*, 2011) or BLAST, since they are based on a more 'sequential' model of the protein sequences.

Moreover, PSPer looks for characteristics suitable for phase-separation based on a more fundamental level compared to sequence similarity or amino acid composition, with its ranking not focusing on specific similarities between the RRMs, Others or PLDs regions, but on their relative position, number and basic biophysical properties. The ability to search for regions with particular biophysical characteristics, as observed in the FUS family proteins, enables PSPer to prioritize FUS-like PSPs among proteomes and other classes of proteins. The LLPS process seems indeed to be driven by fundamental biophysical characteristics that are conserved beyond sequence similarity and domain arrangement.

Wang *et al.* (2018) suggest that other residues, such as glycine and tryptophan, can sometimes substitute the effect of Tyrs and Args as sticky residues, even if they result in weaker interaction. Given the small amount of data available, we decided to be conservative and limit the complexity of our model by excluding the contribution of other amino acids. Since the scope of this work is to allow the identification of novel FUS-like PSPs, we are more interested in finding high confidence candidates for experimental validation (high specificity of the predictions) than including all possible variants of PSP (high sensitivity of the predictions).

4.2 Insights on the LLPS propensity

We tested our method on various scenarios, investigating its ranking ability with the goal to find indications that PSPer could be used to help scientists in this field to select the most promising proteins to be investigated with *in-vitro* experiments. Although very few FUS-like PSPs have been experimentally validated, we tested it against 16 PSPs for which also the saturation concentration has been determined, and PSPer scores showed a very high correlation with the actual experimental values. Given this results, we hypothesize that the application of PSPer to the known 144 proteins composing cytosolic condensates (Bolognesi *et al.*, 2016; Youn *et al.*, 2018) should help identifying which proteins are scaffold or client proteins, and so distinguishing between proteins driving the droplets formation and proteins recruited in it to perform molecular duties. Our analysis shows that only 10 of those proteins are indeed likely to be scaffold proteins that drive the phase-separation process, with evidence in literature confirming the phase-separation properties of a few of the highest ranking proteins. This is in agreement with the hypothesis that condensates are heterogeneous, with only a few scaffold proteins necessary. At the proteome level we ran a GO-terms enrichment analysis showing that the highest scoring proteins predicted by PSPer are enriched for GO-terms related to splicing and mRNA processing, indicating the centrality of the RNA-binding aspect in the LLPS mechanism detected by PSPer.

4.3 Current limitations and future perspectives in the LLPS field

Even if it is unsupervised, our approach is intrinsically bounded by the very low number of currently known FUS-like PSPs. While developing PSPer, we thus limited its complexity by modeling only the most striking biophysical aspects driving LLPS mentioned in Wang *et al.* (2018), without considering e.g. the contributions of other residues besides the Arg–Tyr combination. We think indeed that PSPer should be considered as an early effort to detect FUS-like behavior from their sequence alone, and its main goal is indeed to help a better selection of putative PSPs for *in-vitro* experimental validation, and not the precise biophysical modeling of the LLPS behavior, which will be possible once more data will be available.

The FUS-like PSPs is only one class of proteins capable of LLPS that are currently being investigated by experimentalists. General biophysical approaches to the prediction of PSPs are being developed, such as in Vernon *et al.* (2018). Unfortunately, our knowledge about phase separation is still very limited and such methods may fail to deal with classes of proteins with peculiar condensation characteristics, such as the FUS-like PSPs. An indication of this can be found in Supplementary Figure S9, where we show how the aforementioned tool fails to find a relationship (Spearman $r = 0.14$) between its predicted score and the saturation concentration on the same experimental data showed in Figure 4.

Other proteins, such as DDX4 (Q9NQI0) (Nott *et al.*, 2015) and LAF-1 (D0PV95) (Elbaum-Garfinkle *et al.*, 2015) use different RBDs (such as the DEAD-box helicase) and have disordered, Arg-rich, *N*-terminal RGG domains playing a central role in the droplet formation. These proteins are not currently picked up by PSPer, which is specifically built for the detection of PSPs with FUS-like behavior, and they are scored respectively 0.28 and 0.33 by our model. If we change the RRM used in PSPer from PF00076 (RRM) to PF00270, which corresponds to the DEAD-box helicase, the predicted probability of LLPS behavior for both DDX4 and LAF-1 rises to 0.45 and 0.57, which is respectively very close and above the median of the 22 FUS-like PSPs used in this study. This indicates, when more data will be available, PSPer model could be extended in order to encompass different LLPS mechanisms and thus making it more general in the detection of this behavior.

5 Conclusion

Cytosolic condensates experienced a recent surge in interest, and they are now increasingly recognized as crucial for the cell's ability to provide a fast response to external stress conditions by triggering the formation of temporary membraneless organelles. Only few proteins have been experimentally shown to be able to form such condensates (Wang *et al.*, 2018), and while there are studies investigating the composition of the *membraneless proteome* (Bolognesi *et al.*, 2016; Youn *et al.*, 2018), very little is known about the formation of such organelles. In this paper, we present PSPer, the first *in-silico* method able to identify phase-separating proteins of the FUS family. PSPer is based on identifying the biophysical characteristics of protein regions, and was tested on various scenarios. It is consistently able to prioritize FUS-like PSPs, and we hope that the ranking ability that our model provides will inspire experimentalist to focus on the most interesting and likely PSPs, so steering our understanding of these proteins and the formation of condensates.

Moreover, interesting insights might also be obtained from the analysis of false positives hits, leading to the detection of new characteristics that exclude these proteins from the theoretical model described by Wang *et al.* (2018) and, maybe, allow the definition of new PSP classes. Additionally, because of its generic nature, our HMM-like framework can potentially be applied to many different constrained problems of structural biology with an underlying common mechanism, not only to PSP detection. A potential application could be, for instance, the detection of amyloidosis regions, where aggregation prone regions with very peculiar amino acidic composition and high beta propensity are surrounded by gatekeeper residues (Greenwald and Riek, 2010).

Acknowledgements

D.R. and G.O. are grateful to Anna Laura Mascagni and Angela Motz for support and helpful discussion.

Funding

G.O. acknowledges funding by the Research Foundation Flanders (FWO) [project nr. G.0328.16N]. D.R. is funded by a KU Leuven post-doctoral mandate and by an FWO post-doctoral fellowship.

Conflict of Interest: none declared.

References

Apweiler,R. *et al.* (2004) Uniprot: the universal protein knowledgebase. *Nucleic Acids Res.*, **32** (Suppl. 1), D115–D119.

Banani,S.F. *et al.* (2016) Compositional control of phase-separated cellular bodies. *Cell*, **166**, 651–663.

Banani,S.F. *et al.* (2017) Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.*, **18**, 285.

Boeynaems,S. *et al.* (2018) Protein phase separation: a new phase in cell biology. *Trends Cell Biol.*, **28**, 420–435.

Bolognesi,B. *et al.* (2016) A concentration-dependent liquid phase separation can cause toxicity upon increased protein expression. *Cell Rep.*, **16**, 222–231.

Courchaine,E. and Neugebauer,K.M. (2015) Paraspeckles: paragons of functional aggregation. *J. Cell Biol.*, **210**, 527–528.

Eden,E. *et al.* (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.

Elbaum-Garfinkle,S. *et al.* (2015) The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proc. Natl. Acad. Sci. USA*, **112**, 7189–7194.

Feric,M. *et al.* (2016) Coexisting liquid phases underlie nucleolar subcompartments. *Cell*, **165**, 1686–1697.

Finn,R.D. *et al.* (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39** (Suppl. 2), W29–W37.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Greenwald,J. and Riek,R. (2010) Biology of amyloid: structure, function, and regulation. *Structure*, **18**, 1244–1260.

Grousl,T. *et al.* (2009) Robust heat shock induces eIF2alpha-phosphorylation-independent assembly of stress granules containing eIF3 and 40S ribosomal subunits in budding yeast, *Saccharomyces cerevisiae*. *J. Cell Sci.*, **122**, 2078–2088.

Hennig,S. *et al.* (2015) Prion-like domains in RNA binding proteins are essential for building subnuclear paraspeckles. *J. Cell Biol.*, **210**, 529–539.

Lin,Y. *et al.* (2015) Formation and maturation of phase-separated liquid droplets by RNA-binding proteins. *Mol. Cell*, **60**, 208–219.

Nott,T.J. *et al.* (2015) Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell*, **57**, 936–947.

Rai,A.K. *et al.* (2018) Kinase-controlled phase transition of membraneless organelles in mitosis. *Nature*, **559**, 211.

Rinnerthaler,M. *et al.* (2013) Mmi1, the yeast homologue of mammalian TCTP, associates with stress granules in heat-shocked cells and modulates proteasome activity. *PLoS One*, **8**, e77791.

Romero,P. *et al.* (2001) Sequence complexity of disordered protein. *Proteins*, **42**, 38–48.

Shin,Y. and Brangwynne,C.P. (2017) Liquid phase condensation in cell physiology and disease. *Science*, **357**, eaaf4382.

Vernon,R.M. *et al.* (2018) Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife*, **7**, e31486.

Waite,J.H. (2017) Mussel adhesion–essential footwork. *J. Exp. Biol.*, **220**, 517–530.

Walsh,I. *et al.* (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.

Wang,J. *et al.* (2018) A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell*, **174**, 688–699.

Weber,S.C. and Brangwynne,C.P. (2012) Getting RNA and protein in phase. *Cell*, **149**, 1188–1191.

Youn,J.-Y. *et al.* (2018) High-density proximity mapping reveals the subcellular organization of mRNA-associated granules and bodies. *Mol. Cell*, **69**, 517–532.