


RESEARCH ARTICLE

Charged sequence motifs increase the propensity towards liquid–liquid phase separation

 András László Szabó¹, Anna Sánta¹, Rita Pancsa² and Zoltán Gáspári¹ 
¹ Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary

² Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary

Correspondence

 Z. Gáspári, Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Práter utca 50/A, 1083 Budapest, Hungary
 Tel: +36 1 8864780
 E-mail: gaspari.zoltan@itk.pkpe.hu

(Received 16 November 2021, revised 12 January 2022, accepted 14 January 2022, available online 3 February 2022)

doi:10.1002/1873-3468.14294

Edited by Lukas Alfons Huber

Protein phase separation is a major governing factor in multiple cellular processes, such as RNA metabolism and those involving RNA-binding proteins. Despite many key observations, the exact structural characteristics of proteins involved in the phase separation process are still not fully deciphered. In this work, we show that proteins harbouring sequence regions with specific charged residue patterns are significantly associated with liquid–liquid phase separation. In particular, regions with repetitive arrays of alternating charges show the strongest association, whereas segments with generally high charge density and single α -helices also show detectable but weaker connections.

Keywords: charged residue repeat; charge-dense region; liquid–liquid phase separation; membraneless organelle; single alpha-helix

Protein phase separation

Cells are continuously conducting various complex biochemical processes. These need to be efficiently performed and finely regulated in space and time, therefore cells employ a molecular process wherein macromolecules, mainly multi-domain proteins and RNAs, establish multivalent weak interactions to form functionally specialized liquid compartments, the so-called membraneless organelles (MLOs). Interestingly, this smart solution for the reversible and finely tuned compartmentalization of biochemical processes has only come to the focus of intensive research relatively recently. The general name for the phenomenon is phase separation, and it has been shown to have a role in critical processes of the living cell, such as chromatin regulation, RNA transcription, organization of the postsynaptic density (PSD) and so on [1–3]. Phase separation can result in different states of MLOs from liquids through gels to solids [4].

The molecular assemblies formed by these processes are typically micron-sized objects, containing multivalent proteins with multiple modular domains and disordered regions, often of low sequence complexity [5]. Regarding their functions, there are two main types of molecules participating in phase separation, as distinguished by Banani et al. [6], the ‘scaffolds’ and the ‘clients’. The former type consists of molecules essential to the integrity of the MLOs. The latter type contains the majority of the components; however, they only participate in functionalities under certain circumstances. P bodies are a good example of this duality, they are scaffolded by a few critical RNA-binding proteins and store mRNAs and most protein components of the mRNA degradation machinery as clients. Scaffold–scaffold interactions are more persistent than scaffold–client interactions, and the composition changes according to a set of factors such as stress and the cell cycle [1,7].

Abbreviations

SAH, single alpha-helix; CRR, charged residue repeat; CDR, charge-dense region; uCDR, unsigned charge-dense region; sCDR, signed charge-dense region; MLO, membraneless organelle; PSD, postsynaptic density; MCD, mixed-charged domain; OT, overrepresentation test; SV, synaptic vesicle; IDP, intrinsically disordered protein; LLPS, liquid–liquid phase separation.

The characteristic feature of liquid–liquid phase separation (LLPS) that distinguishes it from other processes related to phase separation is that the solution transitions into distinct phases where certain solutes are present in highly elevated concentrations, and these phases exhibit liquid-like properties. In this specific type of phenomenon, the term ‘scaffolds’ is often replaced by ‘drivers’, referring to sets of proteins that are able to drive LLPS on their own. Smaller molecules and ions are omitted from this category, even if they are required for the initiation of the condensation process. Clients in this context are molecules that may partition into MLOs, but they hold no influence over their formation [8].

A category of proteins whose members are especially prone to such processes is intrinsically disordered proteins (IDPs). Their low-complexity, prion-like subsequences govern LLPS, making the process prone to undergo material state transitions, such as the liquid–solid transition exhibited by RNA-binding protein fused in sarcoma (FUS), as well as TAR DNA-binding protein 43 (TDP-43). Liquid–solid phase transitions – in specific cases called aggregation – are often associated with severe diseases such as amyotrophic lateral sclerosis (ALS), making the examination of IDPs and LLPS rigorously researched fields [9,10].

Pak et al. [11] described the phase separation of the intrinsically disordered intracellular domain of Nephric (NICD) as complex coacervation. Coacervates are dense liquid droplets of macromolecules. While simple coacervation requires only one type of polymer, complex coacervation requires associative interactions between multiple soluble molecules (of which at least one is a macromolecule), resulting in de-mixing into a polymer-dense and a polymer-depleted phase that are in equilibrium with each other. When expressed as a soluble protein, NICD formed droplets in HeLa cells, and LLPS was also observed *in vitro* when mixed with positively supercharged GFP. While no specific sequence was identified in NICD responsible for LLPS, a pattern of blocks of negatively charged residues was observed. This implies phase separation to be robust in withstanding mutations, as shuffle and deletion mutants were still able to drive phase separation as long as charged blocks were retained.

The role of arginine-rich structural components in nuclear speckle condensation has also been shown by Greig et al. [12] where such charged components are referred to as mixed-charged domains (MCDs). Their investigations highlight the importance of arginine and aspartic acid (RD)-enriched MCDs in certain fungi, where their condensation serves to form gates for cell-to-cell channels.

Many proteins involved in the phase separation phenomena have been identified in the last 20 years, and thus it has become a relevant issue to catalogue them. PhaSepDB [13] is a manually curated database that has been created with the specific purpose of providing researchers with a detailed, reliable collection of information about the proteins connected to the phenomenon. There are other, more specific databases of phase-separating protein such as the stress granule protein database constructed by Nunes et al. [14], or PhaSePro that specifically collects LLPS drivers [15], but as of today one of the most comprehensive databases is PhaSepDB. Data collection and processing for such a library of proteins consists of several levels, and new entries are added into the database after their respective publications show evidence on the localization of the proteins to MLOs.

There are three categories of protein sets available in PhaSepDB: The first one is the ‘Reviewed’ category that only contains articles published after January 1st of 2000, constituting the most reliable subset. The second category is called ‘UniProt reviewed’ that includes a wider range of articles, not restricted by publication date, although its results must be confirmed by more recent studies. The third and final category is the ‘High-throughput’ research data, generated by high-throughput methods categorized as either organelle purification, proximity labelling, immunofluorescence image-based screen or affinity purification.

Single α -helices and other charged residue repeats

Single α -helices (SAHs) are protein segments that form stable and rigid helical structures even in isolation [16,17]. These regions are rich in arginine, lysine and glutamate and exhibit a characteristic repetitive pattern of oppositely charged residues such as glutamic acid and lysine. Their length may vary from a few dozen up to about 200 residues, and they are stabilized by intrahelical salt bridges [18]. These regions show similar characteristics to coiled coils, as Peckham et al. [19] showed that $\sim 4\%$ of human proteins that were previously predicted to contain coiled-coils actually have SAHs instead. SAH domains have been found to be prevalent in RNA-binding proteins [16], although the SAH region itself is unlikely to directly contribute to the interaction with RNA.

Inspired by the prominent role of RNAs and RNA-binding proteins in phase separation, we intended to investigate whether proteins involved in LLPS are enriched in SAH segments. As SAH segments constitute a special case of sequences with high charge

density, we have extended our analysis to segments with repeating patterns of charged residues, termed charged residue repeats, (CRRs), as well as with segments enriched in charged residues but without any further specifications, denoted charge-dense regions (CDRs) below. We differentiate between ‘unsigned’ and ‘signed’ CDRs on the basis of their net charge. In this classification, most CRRs are special cases of CDRs, and SAHs constitute a sub-class of CRRs (Fig. 1).

CRRs are identified here with the program FT_CHARGE, developed for the identification of SAHs but also allowing the identification of patterns with repeat frequencies outside of the range characteristic for SAHs. Thus, CRRs can be defined as segments for which the Fourier spectrum of the charge correlation function (Eq. 1. In ref. [16]) contains a peak with an amplitude significantly higher than expected from a random sequence with similar content of positively and negatively charged residues. Thus, the approach can be categorized as a ‘pattern recognition strategy’ according to the classification by Luo & Nijveen [20]. The minimum length of the CRRs is defined by the window size (16, 32 or 64) used by FT_CHARGE, the maximum length is not limited as consecutive windows with detected CRRs can be combined.

It is important to note that while the majority of CRRs satisfy at least one of the two definitions for CDRs, not all of them fall under those categories. The cleavage stimulation factor subunit 2 (CSTF2, UniProtKB ID: P33240) for example contains 12×5 AA tandem repeats on a 60 residue-long segment that was detected as a CRR by the FT_CHARGE algorithm – a 105 residue-long sub-sequence encompassing the annotated region was highlighted (Fig. 2).

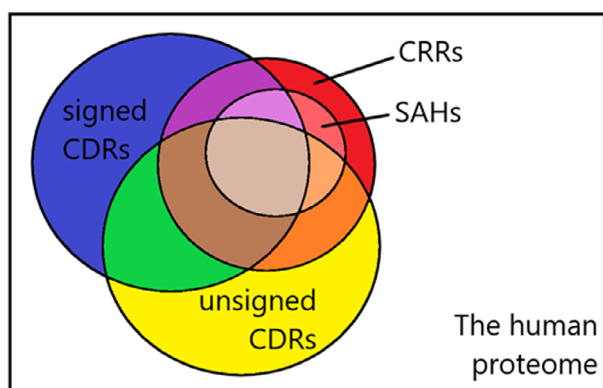


Fig. 1. Venn diagram illustrating the relations between single α -helices, CRRs, as well as signed and unsigned CDRs within the human proteome.

The region does not qualify as a SAH due to its slightly higher Fourier frequency, although there are multiple helical structures along the entire sequence of this protein. While up to this point, we could not find a region that qualifies as a SAH while remaining outside our definitions for CDRs, we would reserve the possibility for their existence as the herein applied representation of the human proteome only included a single isoform per protein.

To strengthen the specificity of our investigations, we have performed all our evaluations by excluding proteins with transmembrane segments. The rationale behind this is that the presence of charged regions is expected to be characteristic of soluble proteins. Without excluding transmembrane proteins, any possible association between charged regions and LLPS might simply reflect that phase separation is primarily occurring between soluble biomacromolecules.

Results

CDRs, CRRs and SAHs in the human proteome

We have restricted our studies to the human proteome as specified at the beginning of the [Methods](#) section. For the prediction of SAHs, we used our previously established methodology, involving the FT_CHARGE algorithm. CRR detection was carried out by relaxing the requirement for the charge frequency characteristic for SAHs. CDRs were detected by an algorithm enumerating Asp, Glu, Arg and Lys residues in sequence windows of different lengths (see [Methods](#) section for details). Since the vast majority of human proteins were found to contain a CDR according to our original detection threshold, we have applied a stricter criterion and investigated only the top 1% of CDRs (Table 1, Fig. 3, Table S1).

CRR-containing proteins show significant enrichment in proteins involved in LLPS, while proteins with SAHs show a weaker but still detectable association

Fisher’s exact test of independence showed that there is an exceptionally high probability that CRRs contribute to the LLPS of human proteins (Figs 4 and S1, Tables S1 and S2).

SAHs can be regarded as special subsets of CRRs detected with FT_CHARGE with the frequency region between $1/9$ and $1/6$. SAH regions still exhibit a statistically significant enrichment in LLPS-associated proteins, although this is considerably weaker than for CRRs in general (Fig. 4 and Tables S1 and S2).

magltvrpavdrslrsvfvgnipyateeqldkfsevgpvvsfrlvydretgkpkgyg
 fceyqdqetalsamrnlngrefsgralrvdnaaseknkeelkslgtgapviespygetis
 pedapesiskavaslppeqmfelmkqmklcvnspqearnmlqnpqlayallqaqvvmr
 ivdpeialkihrqtniptliagnpqpvhgagpgsgsnvsmnqqnpqapqaqslggmhvn
 gapplmqasmqggvppagqmpaavtggpgslapgggmqaqvgmpgsgpvsmerggqvpmq
 dpraamqrgslpanvptprgllgdapndprggtllsvtgeveprgylgpphqqppmhvp
 ghesrgppphelrggplpeprplmA**EPRGPMLDQRGPPLDGRGGRDPRGIDARGMEARAM**
EARGLDARGLEARAMEARAMEARAMEARAMEARAMEVVRGMEARGMDTRGPVPGPRGPIPS
GMQGPSPINmgavvpqgsrqvpmqgtgmqgasiqggsqpggfspgqnqvtpqdhekaal
 imqvlqltadqiamlppeqrqsililkeqiqkstgap

Fig. 2. The CRR of CSTF2 is not recognized as a SAH.

Table 1. Summary table exhibiting the number of sequences associated with different sequence motifs, as well as the fraction of residues encompassed by those motifs. Includes data for the full proteome and its variants with reduced redundancy.

Sequence motifs	Full proteome	Redundancy-filtered proteome		
		90%	70%	50%
Number of proteins				
All proteins	20659	19638	18294	15672
CDRs (unsigned)	9731	9314	8792	7669
CDRs (signed)	14065	13471	12757	11097
CRRs (all)	1054	1025	985	910
CRRs (>=90% overlap)	782	763	738	688
with uCDRs				
CRRs (>=90% overlap)	227	221	214	208
with sCDRs				
SAHs (all)	134	131	126	118
Percentage of residues				
All proteins	100%	96.74%	92.04%	80.73%
CDRs (unsigned)	10.02%	9.63%	9.16%	8.22%
CDRs (signed)	14.53%	13.98%	13.30%	11.58%
CRRs (all)	0.82%	0.80%	0.77%	0.72%
CRRs (>=90% overlap)	0.54%	0.52%	0.50%	0.47%
with uCDRs				
CRRs (>=90% overlap)	0.14%	0.13%	0.13%	0.12%
with sCDRs				
SAHs (all)	0.11%	0.11%	0.10%	0.10%

Another way to illustrate the level of association between LLPS and the presence of sequence motifs is to compare their enrichment in LLPS-related proteins. For example, the ratio of sequences with CRRs to other proteins is 1010 : 19418 in the full proteome in the case of entries that are unrelated to LLPS. The same ratio in the case of related sequences is 44 : 187. This means that the ratio increases 4.52-fold between the two functional categories (Fig. 5).

Regions enriched in charged residues in general are prevalent in proteins associated with LLPS

We have investigated whether regions enriched in charged residues but not necessarily exhibiting regular repeating patterns are also associated with LLPS. To this end, besides the CRRs detected by the FT_CHARGE algorithm, a more general type of protein sub-sequences (CDRs) was similarly probed for relations to LLPS. We have formulated two different approaches for their detection. One of them defines CDRs as protein sub-sequences where the density of charged residues is significantly high. The other one defines them as protein sub-sequences with an overall charge that significantly differs from zero (see [Methods](#) section). We refer to the regions gained through the former approach as ‘unsigned CDRs’, while regions resulting from using the other definition were named ‘signed CDRs’ – as their charges may have a negative sign, too. The protocols for detecting such components are described in detail in the [Method](#) section.

Unsigned and signed CDRs

If we consider the top ~ 1% of all hits to be unsigned CDRs, then approximately 47.10% of human proteins contain some kind of CDR. When these CDRs are compared to CRRs, it turns out that 90.99% of CRRs display an above 90% overlap with CDRs, with 87.80% of all CRRs being entirely encompassed by CDRs, while 2.53% of them are free from any overlap. Sequence-wise, 15.92% of the human proteome is categorized as CDRs with the above-mentioned parameters.

Considering the top 1% of all hits as signed CDRs results in 22.76% of CRRs having at least 90% overlap with the top 1% of CDRs, and 20.25% having total coverage, while 37.14% having no coverage at all.

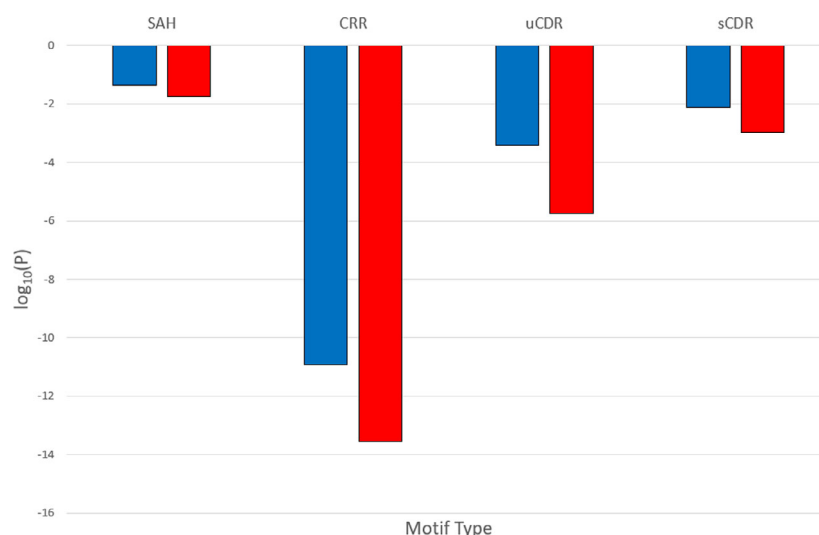


Fig. 3. The decimal logarithm of the P -values that belong to Fisher's tests where the functional attribute is the protein's relation to LLPS, and the structural attribute is the presence of different sequence motifs. Tests involving the entire proteome are showcased in red while tests that excluded transmembrane proteins are in blue.

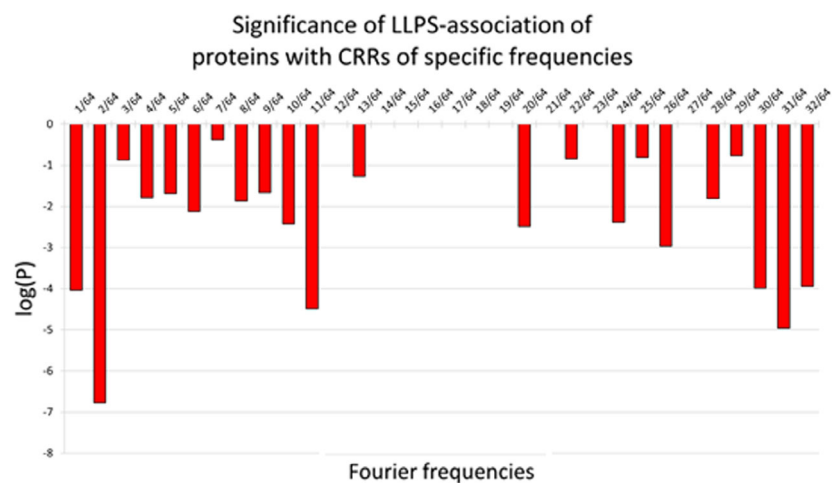


Fig. 4. The decimal logarithm of the P -values that belong to Fisher's tests where the functional attribute is the protein's relation to LLPS, and the structural attribute is the presence of CRRs with specific Fourier frequencies.

ROC analysis of motif presence as a predictor of LLPS

We have performed receiver operating characteristic (ROC) analysis to assess whether the presence of such motifs has predictive value for LLPS-association for a given protein. We found that while the presence of these motifs is not a strong predictor of LLPS, the absence of such regions largely precludes the participation of the protein to form condensates (Fig. 6).

Robustness of the motif-LLPS association in random sequence sets

The robustness of the association was also investigated in random sequence sets generated to match the size distribution of proteins in PhaSepDB (see [Methods](#) section). Our observations confirm that these motifs

are specifically enriched in LLPS-associated proteins (Fig. S2 and Table S4).

Functional analysis of charged motif-containing proteins in relation to LLPS

The Overrepresentation Test (OT) of Panther's Gene List Analysis tools was also utilized to compare the various special motifs in sequences that are related to LLPS versus the regions of proteins that are unrelated to the phenomenon. A total of eight tests were carried out, each comparing either LLPS-related or LLPS-unrelated SAHs, CRRs and signed as well as unsigned CDRs. From the OT-given Gene Ontology (GO) terms we manually selected the ones that are at the lowest level within their respective branch of the GO term hierarchy, meaning that only the most specific – and

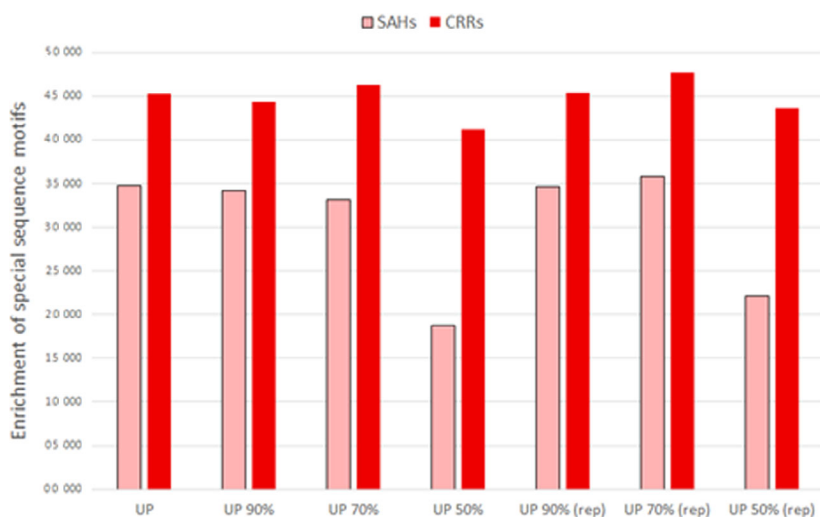


Fig. 5. CRR-enrichment (a proportional increase of the ratio of CRR-containing sequences) in case of the full proteome as well as its variants with reduced redundancy (with clusters and with representative sequences only, see [Methods](#) for details).

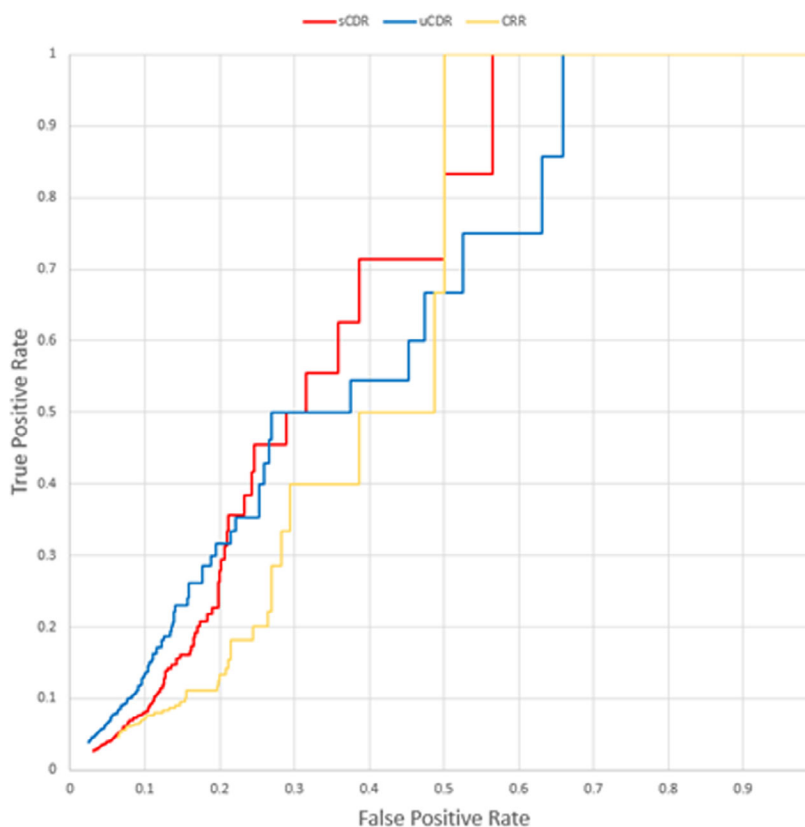


Fig. 6. Receiver operating characteristic (ROC) tests on CRRs, unsigned CDRs and signed CDRs in non-transmembrane proteins, evaluating the presence of these motifs as an indicator of affinity towards LLPS. For CRRs the test shows the highest value at 0.5 FPR (TPR = 1) when the threshold is set to 0.38, meaning that sequences, where at least 0.38 of all residues are assigned to CRRs, exhibit a high probability of participating in LLPS. These numbers are TPR = 0.667 and threshold = 0.80 for uCDRs, and TPR = 0.714 and threshold = 0.83 for sCDRs. The ROC curves suggest that the presence of CDRs/CRRs is not a strong predictor for LLPS, but the absence of such motifs is an indicator that the protein is not associated with LLPS.

thus least redundant – terms were analysed further. These terms were compiled into a table where each term's enrichment – or lack thereof – was noted in the case of the four different kinds of LLPS-related and LLPS-unrelated motifs. From there it is possible to categorize each term, based on the number of LLPS-related categories they were enriched in (0–4) and the number of LLPS-unrelated categories they were enriched in (0–4). This method of categorization

created a gradient table that shows the distribution of GO terms from highly related to LLPS to highly unrelated to LLPS. The table then could be reduced to contain GO terms with specific keywords in them, such as ‘regulation’, ‘RNA’ or ‘signal’ (Table 2, Fig. 7, Table S3).

A simple analysis of the relationship between postsynaptic localization, RNA-binding function, the presence of charged motifs and LLPS-association (Fig. 7 and

Table 2. Gradient tables depicting the distribution of all GO terms (a), terms related to RNAs (b), terms related to responses (c), and terms related to metabolism (d). In the latter case, the actual keyword was 'metabol' to include GO terms containing 'metabolism' and 'metabolic' as well, much like the keyword 'RNA' encompasses 'mRNA', 'rRNA', 'snoRNA', 'RNA-binding', etc.

		Total number of entries				307
Enrichment		0/4 Unrelated	1/4 Unrelated	2/4 Unrelated	3/4 Unrelated	4/4 Unrelated
4/4 Related	1	0	0	0	0	0
3/4 Related	10	0	0	0	0	0
2/4 Related	53	7	1	0	0	0
1/4 Related	90	5	3	0	0	0
0/4 Related	9	112	16	0	0	0
b)		Total number of entries related to 'RNA'				55
Enrichment		0/4 Unrelated	1/4 Unrelated	2/4 Unrelated	3/4 Unrelated	4/4 Unrelated
4/4 Related	0	0	0	0	0	0
3/4 Related	5	0	0	0	0	0
2/4 Related	13	1	0	0	0	0
1/4 Related	26	0	3	0	0	0
0/4 Related	0	7	0	0	0	0
c)		Total number of entries related to 'response'				28
Enrichment		0/4 Unrelated	1/4 Unrelated	2/4 Unrelated	3/4 Unrelated	4/4 Unrelated
4/4 Related	0	0	0	0	0	0
3/4 Related	0	0	0	0	0	0
2/4 Related	10	0	0	0	0	0
1/4 Related	9	0	0	0	0	0
0/4 Related	1	6	2	0	0	0
d)		Total number of entries related to 'metabol'				22
Enrichment		0/4 Unrelated	1/4 Unrelated	2/4 Unrelated	3/4 Unrelated	4/4 Unrelated
4/4 Related	0	0	0	0	0	0
3/4 Related	0	0	0	0	0	0
2/4 Related	1	0	0	0	0	0
1/4 Related	2	1	0	0	0	0
0/4 Related	4	14	0	0	0	0

Fig. S3–S9) reveal that there are barely any proteins related to either LLPS, PSDs or the molecular function of RNA-binding that are devoid of any motifs with either charged patterns or high charge density. Furthermore, the distributions implicate that RNA-binding proteins are more likely to participate in phase separation than PSD components as the ratio of entries in the middle (yellow) section doubles between the two figures. This enrichment can also be observed between PSD proteins and RNA-binding sequences when we only consider CRRs as a structural factor. Although the number of cases is small, it also seems that for RNA-binding proteins, the presence of charged motifs/CRRs is more prevalent when considering proteins participating in LLPS.

Discussion

Case studies

Our results clearly indicate that proteins with regions containing a high fraction of charged residues are preferentially associated with LLPS. Our approach allowed

us to investigate multiple aspects of this association, consideration of the presence or absence of specific repetitive patterns. Below we discuss several proteins, the possible relationship between LLPS and the presence of SAHs/CRRs/CDRs. Examples, wherever it is possible, are based on the information available in the richly annotated PhaSePro database, where information about the region responsible for LLPS is provided if available.

Table 1 contained 1054 protein sequences that encompass CRRs but out of those only 44 sequences are considered to be LLPS-related. These sequences were investigated further to explore their connection to phase separation. The investigation revealed that a large variety of proteins may participate in the phenomenon, including ribonucleases [21], splicing factors [22,23], transcriptional repressors [24,25], translational initiation factors [26], transport- [27,28], Zinc finger proteins [29,30] and so on. However, there happens to be a common feature in most of the investigated proteins, as 31 of them carry out the molecular function of RNA-binding [21–39]. The fact that 70.45% of

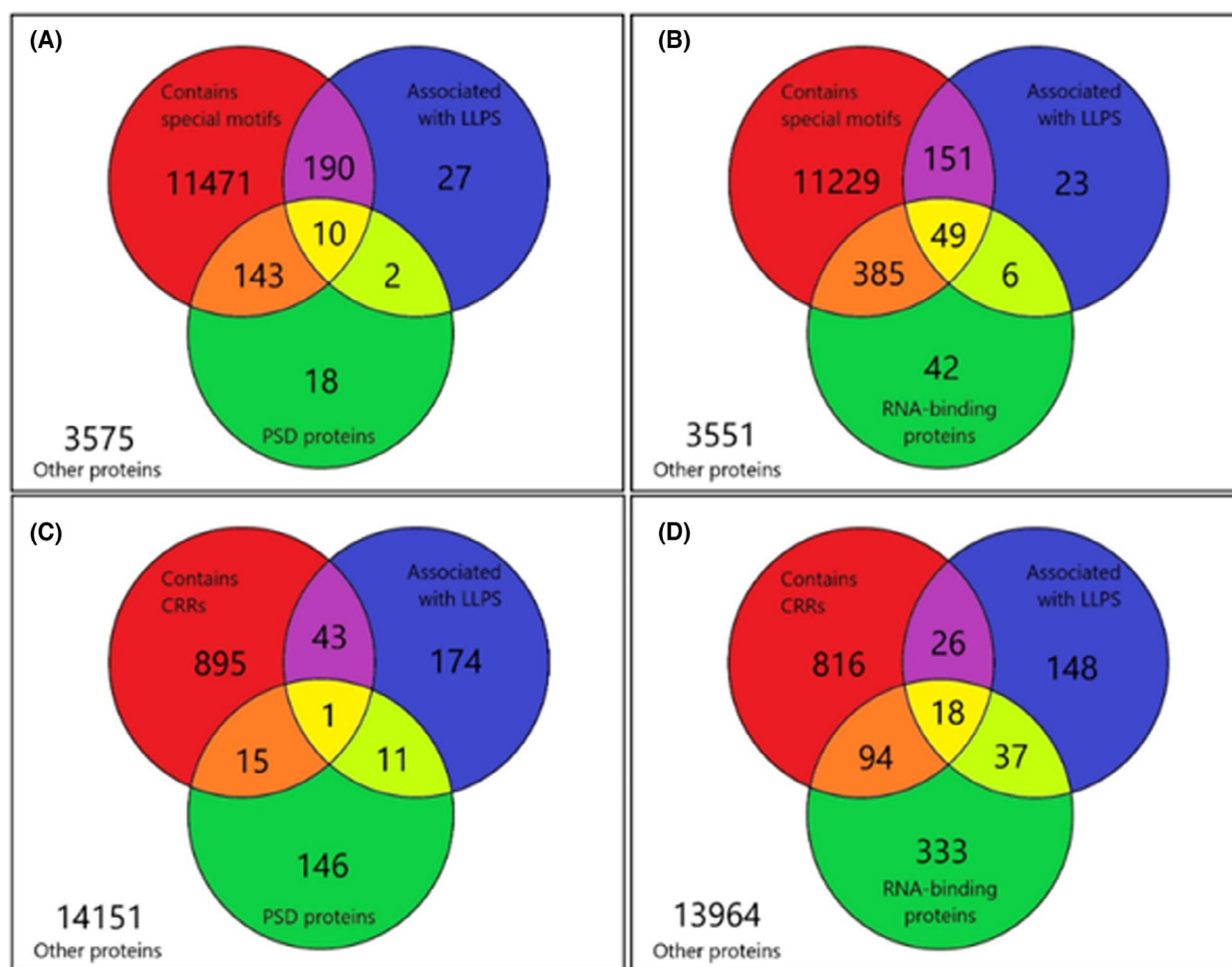


Fig. 7. Distribution of human, non-transmembrane proteins according to their special charged motif (A–B) or CRR (C–D) content, LLPS association and their functional role in the PSD (A, C) or RNA binding (B, D).

these proteins can bind RNA molecules is significant and reinforces that RNA-binding is a process that is promoted by charged protein regions, such as CRRs. Another plausible explanation may be that many MLOs – that are formed by LLPS – contain RNAs. According to the frequencies of the CRRs predicted by FT_CHARGE, 5/44 sequences possibly contain single α -helices, all of which constitute at least 10% of the respective sequence.

It is important to note that while RNA-binding connects most of these proteins, their sequences, as well as their biological function, vary greatly. There were only two cases where multiple sequences got clustered together. In the first case, the two proteins were the probable global transcription activator SNF2L2 and the transcription activator BRG1. Both are involved in transcriptional activation and repression of select genes by chromatin remodelling [40]. They are also

components of SWI/SNF chromatin remodelling complexes that carry out key enzymatic activities and change chromatin structure. While these two proteins proved to be participating in the same processes, they still only showed 59.75% sequence identity.

In the second case, there were three proteins in the same cluster, proline- and glutamine-rich splicing factor SFPQ being the representative sequence, non-POU domain-containing octamer-binding protein NONO showing 58.39% identity, and paraspeckle component 1 PSCP1 exhibiting 54.49% sequence identity. The three proteins cooperatively regulate androgen receptor-mediated gene transcription activity in the sertoli cell line [41,42] and, besides the noncoding RNA NEAT1, are major components of the paraspeckle. These proteins contain a conserved SAH located at the C-terminal end of a right-handed coiled-coil segment [43]. The region responsible for

dimerization, and thus LLPS, is the so-called NOPS region on the N-terminal side of the coiled-coil. The same goes for the trinucleotide repeat-containing gene 6B protein (TNRC6B) that participates in RNA-mediated gene silencing and acts as a scaffolding protein that can simultaneously interact with argonaute proteins regulating miRNA and siRNA maturation and recruit deadenylase complexes carrying out poly (A) tail-shortening processes [44]. Finally, the E3 ubiquitin-protein ligase RNF168 accumulates repair proteins to sites of DNA damage *via* binding to ubiquitinated histone H2A and H2AX [45]. Apart from SNF2L2 and BRG1, all the above proteins do encompass at least one alpha-helical structure, and all of them seem to exhibit some kind of RNA-binding trait, except RNF168.

SynGAP1 is part of the complexes clustered to NMDA receptors in excitatory synapses. The mouse ortholog has been explicitly shown to participate in LLPS, with the responsible region being its C-terminal segment, which forms a trimeric coiled-coil and has a PDZ-domain-binding motif that interacts with PSD-95. The role of SynGAP in LLPS formation in the PSD has been demonstrated in multiple experiments. The key feature for LLPS in SynGAP is its multivalent nature provided by the trimerization *via* its coiled-coil motif. SynGAP does not contain any SAHs or CRRs detected by our methods, but it bears CDRs and so does its binding partner PSD-95 (Disks large homolog 4). This interaction is part of a larger scaffolding protein network in the PSD involving DLGAP1 (the guanylate kinase-associated protein GKAP — a.k.a. Disks large-associated protein 1), Shank3 (SH3 and multiple ankyrin repeat domains proteins), Homer3 and the C-terminal region of the NR2B subunit of NMDAR. Mixed all together, they demonstrated phase separation *in vitro*, driven by a complex set of multivalent interactions [46]. All of the aforementioned PSD proteins contain CDRs, but neither SAHs nor other CRRs, and only SynGAP and PSD-95 are found in PhaSePro, where in both cases, the LLPS-driving region is overlapping with at least one CDR (Fig. 8). It should be noted, that unlike in the previously mentioned networks, none of these proteins are associated with nucleic acid binding, despite the presence of localized mRNA and miRNA that modulate local translation [47]. Instead, this group is rich in protein–protein interaction domains, such as SH3, PDZ, SAM domains and their binding motifs.

Synapsin-1 participates in neurotransmitter release in the presynaptic region. The region responsible for phase separation is its Pro/Gln-rich C-terminal half and the protein can form condensates on its own or

with binding partners containing SH3 domains. The condensates can also contain synaptic vesicles (SVs), providing the basis of the self-organization of SV clusters. Phosphorylation of Synapsin-1 by CamKII causes dissociation of the condensates [48]. Synapsin-1 does not seem to contain any SAHs/CRRs/CDRs (Fig. 8).

In the case of the 44 sequences that are related to LLPS and contain CRRs, it has also been observed that only one of them (U3 small nucleolar RNA-associated protein 6 homologs, UniProtKB ID: Q9NYH9) is devoid of predicted disordered segments according to their UniProtKB annotation. Furthermore, from the remaining 43 sequences, there are only two (Paraspeckle component 1, UniProtKB ID: Q8WXF1 and non-POU domain-containing octamer-binding protein, UniProtKB ID: Q15233) where none of the disordered segments overlap with any of the charged regions. It must be noted here that there are disordered segments within the other 41 sequences that do not overlap with any of the charged regions but in each of those 41 sequences, there is at least one disordered segment that does.

While these observations may indicate that disordered regions played a role in at least some of the cases where CRRs were associated with LLPS it must be urged that there are several proteins with experimental evidence where phase separation is directly driven by electrostatic interactions (Table S5). There are 14 such sequences within the PhaSePro database, out of which nine are human entries. Either CRRs, unsigned charge-dense region (uCDR)s and signed charge-dense region (sCDR)s are assigned to all but one of these human entries, the probable ATP-dependent RNA helicase DDX4 (UniProtKB ID: Q9NQI0) that contains at least one region with a charge density just below the threshold to be included in the top 1% of CDRs.

Role of charged regions in LLPS

While our results clearly demonstrate an association between charged segments and LLPS, the rationale behind this is not straightforward. The association between LLPS and charged motifs weakens but still remains significant when we exclude transmembrane proteins. It should be noted that while this decision is based on the general expectation that primarily soluble proteins contain charged motifs and participate in phase separation, there are examples of transmembrane proteins involved in LLPS. In our reference proteome two transmembrane proteins are associated with LLPS: Linker for activation of T-cells family member 1/LAT (UniProtKB ID: O43561) and nephrin (UniProtKB ID: O60500). We have also identified several

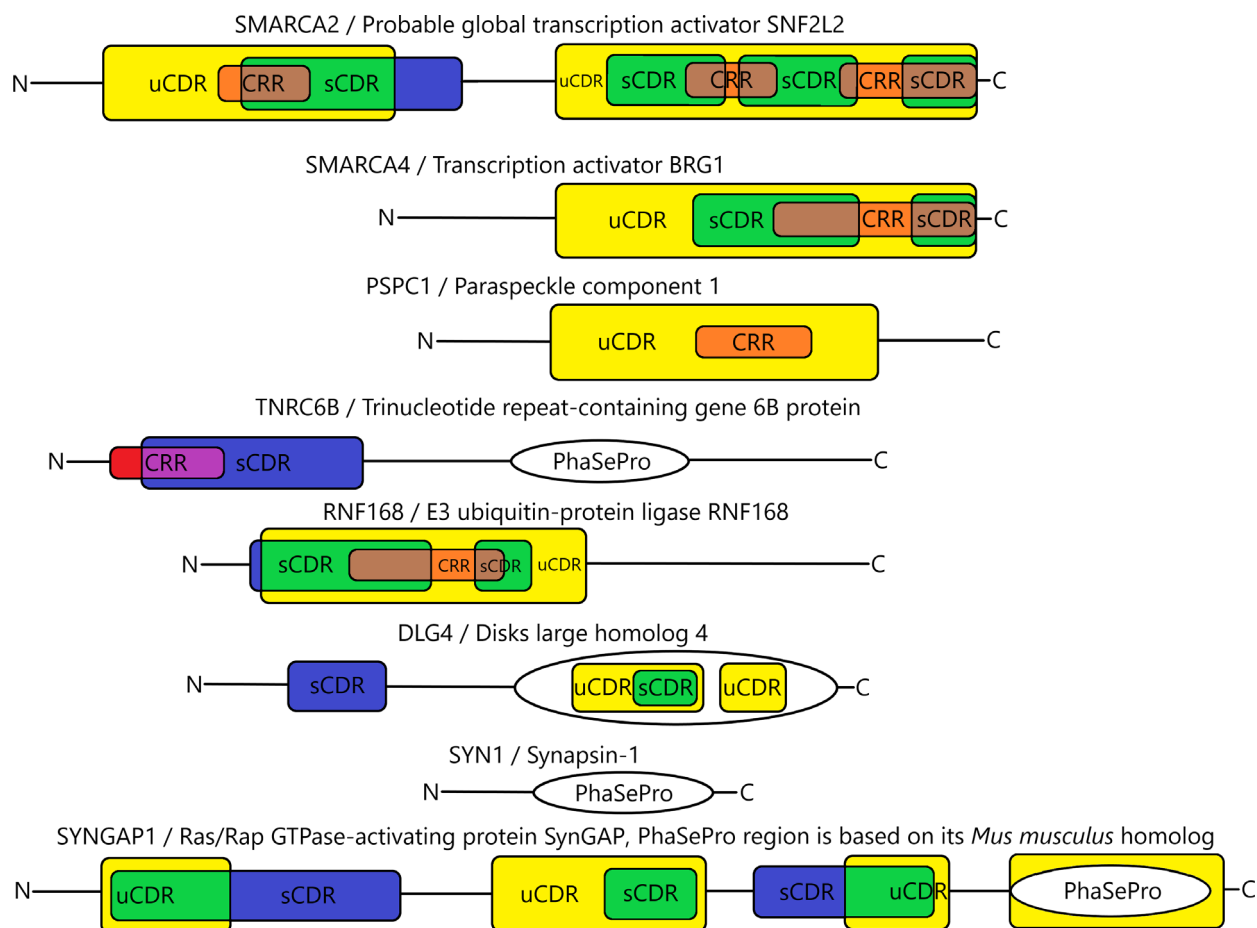


Fig. 8. Sequence motifs as well as PhaSePro-based sub-sequences that drive LLPS. The colour-coding of individual motifs as well as their overlaps are in accordance with Fig. 1. While currently there isn't any experimental evidence that the human protein SynGAP participates in the phenomenon, its *Mus musculus* homolog has, thus a PhaSePro region has been annotated according to that.

transmembrane proteins, both single- and multi-pass ones (e.g., vesicle-associated membrane protein-associated proteins, sodium bicarbonate transporter 3, etc.), that contain at least one SAH, the most specific charged motif investigated here. Thus, the inclusion of transmembrane proteins is also not necessarily a biologically unfeasible approach.

We have also investigated whether proteins of similar length to those in PhaSepDB exhibit a distribution of charged elements characteristic of proteins involved in LLPS. Our results suggest that all of the investigated motifs, but especially CRRs and uCDRs are more abundant in LLPS-associated proteins than in our random datasets. These results suggest that the observed association between charged motifs and LLPS is unlikely to reflect some trivial biophysical constraint (Table S4 and Fig. S2).

Importantly, our detailed investigation of the association between LLPS and charged motifs reveals that

the absence of such motifs largely precludes the participation of a protein in LLPS and not that the presence of a CDR, CRR or SAH would be a strong indicator of LLPS formation. Here, it should be noted that the accuracy of listings of proteins associated with LLPS is expected to increase with novel proteins recognized and curation efforts getting more intensive, probably resulting in an overall increase of proteins reliably proven to form condensates. That said, it is highly unlikely that charged regions identified here in this study might be directly responsible for LLPS in general. It is not the case for all proteins detailed above, let alone phase separating proteins without a CRR or CDR. Thus, it is more likely that the enrichment of CRRs and CDRs in LLPS-prone proteins has a more complex explanation. It should be noted that little is known about the structural features of CRRs and CDRs. Not surprisingly, many regions rich in charged residues, including SAHs, are predicted to be

mkaqgeteeseeklskmsslllerlhakfnqnrpwsetiklvrrqvmekrvvmssgghqhlvs
 cletlqkalkvtslpamtdrlesiarqnglgshlsasgtecyitsdmfyvevqldpagql
 cdvkvahhgenpvscepelvqqlreknfdefskhllkglvnllylpgdnklktkmylalqsl
 eqdlskmaimywkatnagpldkilhgsvgltprrsgghlmnlkyyvpsdlddktaspi
 ilhennvsrslgmnasvtiegtssavyklpiaplimgshpvdnkwtpsfssitsansvdlp
 acfflkfpqpipvsrafvqklqncctgiplfetqptyaplyelitqfelskdpdpiplnhn
 mrfyaalpqqhcyflnkdaplpdgrslqgtlvskitfqhpgrrvplilnlirhqvayntl
 igscvkrtilkedspgllqfevcplsesrfsvsfqhpvndslvcvmdvdqdsthvsckly
 kglstdalictddfiakvvqrcmsipvtmrairrkaetiqadtpalsliaetvedmvkknl
 ppasspygmttggnpmsgtttptntfpggpittlfnmsmsikdrhesvghgedfskvsq
 npiltslqitgnggstigssptpphhtpppvssmagntknhpmlmnlkdnpaqdfstl
 ygssplerqnsssgsprmeicsg**SNKTKKKKSSRLPPEKPKHQTEDDFQRELF SMDVD**sq
 npifdvnmtadtldtphitpapsqcstppptypqpvpvhpqpsiqrmlrslssdsigpdvt
 dilsdiaeeasklpstsdcpaigtplrdsessghsqstlfdsdvfqtnn**NENPYTDPAD**
LIADAAGSPSSDPTNHFFHDGVDFNPDLLNSQSQSGFGEEYFDESSQSGDNDdfkgfas
 qalntlgvpmllggdngetkfkgnnqadtvdfsiisvagkalapadlm**EHHSQSQGPLLTT**
GD LGKEKTQKR VKEGNGTSNSTLSGPG L DSKPGKRSRTPSNDGSKDKPKRKKADTEGK
SPSHSSSNRPFTPPTSTGGSKSPGSAGRSQTPPGVATPPIPKITIQIPKGTVMVGKPSH
SQYTSSGSVSSSGSKSHSHSSSSSSASTSGKMKSSKSEGSSSSKLSSSMYSSQGSSGS
SQSKNSSQSGGKPGSSPITKHGLSSGSSSTKMKPQGKPSLMNPSSLKPNISPSHSRPPG
GSDKLASPMKPVPGTPPSSKAKSPISSGSGGSHMSGTSSSSGMKSSSGLGSSGSLSQKTP
PSSNSCTASSSSSFSSSGSSMSSSQNHGSSK GKSPSRNKKPSLTAVIDKLLKHGVVTS GP
GEDPLDGQMGVSTNSSSHPMSSKHNMSSGGEFQ GKREKSDKDKSKVSTSGSSVDSSKKTSE
SKNVGSTGVAKIIISKHDGGSPIKAKVTLQKPGESSGGLRPMAS SKNYGSPLISGST
PKHERGSPSHSKSPAYTPQNL DSESESGSSIAEKSYQNSPSSDDGIRPLPEYSTEKHKKH
KKEKKKVKDKDRDRDRDKDRDKKSHSIKPESWSKSPISSDQSLSMTSNTILSADRPSRL
SPDFMIGEEDDLMDValign

Fig. 9. Sequence of MED1 coloured according to the presence of charged motifs (color-coded according to Fig. 1).

intrinsically disordered by several prediction methods. However, SAHs are known to possess a stable well-characterized structure and can be recognized by specialized algorithms. Thus, the presence of specific structural preferences for other types of CRRs cannot be completely excluded, and the formation of more or less well-defined monomeric or oligomeric structures might occur for specific sequences. Regularly alternating blocks of positively and negatively charged residues might offer a platform for intermolecular

associations in multiple mutual orientations. The mediator of RNA polymerase II transcription subunit 1 (MED1, UniProtKB ID: [Q15648](#)) is a great example of this, as it has a relatively large region (Fig. 9) experimentally connected to phase separation that contains acidic and basic segments [49]. This intrinsically disordered region also overlaps with CRR, uCDR and sCDR predictions.

Our current hypothesis on the role of charged regions in LLPS is that these can provide

structural/dynamical features that can be robustly maintained both in the solution and the condensed phase and help the relative positioning of the regions responsible for specific interactions. The possible more direct role of the charged motifs in actual condensates can also not be excluded but would require complex experimental investigations.

Methods

Surveying protein sequences for CRRs

The version of the human reference proteome from UniProtKB (<https://www.uniprot.org/proteomes/UP000005640>) used for this study encompasses 20659 human genes – with one isoform per gene – and represents the 2020 September 3 update. Some investigations involved the exclusion of transmembrane proteins that was carried out by removing entries from the above-mentioned fasta file based on another UniProtKB query that encompassed 41818 sequences – that are not necessarily part of the reference proteome – annotated as human transmembrane proteins (the exact query was `annotation:(type:transmem) AND organism: 'Homo sapiens (Human) [9606]'`). Furthermore, additional small sequence sets matching the length distribution in PhaSepDB have been generated by randomly selecting one – or ten – from the above-mentioned 20659 sequences for each (human) PhaSepDB entry where the two must have had a similar ($\pm 5\%$) length.

To identify segments with CRRs, the method FT_CHARGE was used. FT_CHARGE was developed to identify SAH regions in proteins that contain a repeating charge pattern with a specific frequency between 1/9 and 1/6 but it is also capable of detecting other motifs with regularly alternating positively and negatively charged residues [50]. In this work, the largest window size for FT_CHARGE was 64 residues, corresponding to a minimum frequency of 1/64. The frequency of 1/64 practically corresponds to a long, repeated segment of identically charged amino acids (e.g., a polylysine run), whereas the frequency of 1/2 corresponds to a region of residues with alternating charges (e.g., the sequence 'KEKEKEKEKE'). For more details on the implementation of FT_CHARGE refer to [51,52]. Case studies and statistics on CRRs are shown in Tables S5 and S7, as well as in Fig. S10.

Clustering human proteins based on structural predictions

To remove redundancy from the sequence set, the web-server CD-HIT was employed to cluster the sequences on their similarity. The cut-off values for this were 0.9, 0.7 and 0.5, meaning 90%, 70% and 50% sequence identity, respectively. Clustering was carried out with global sequence

identity, a 20-residue bandwidth of alignment, a minimal sequence length of 10 residues, and default alignment coverage parameters. Sequences were assigned to the best cluster that met the threshold. The output of CD-hit was further processed using MATLAB to ensure compatibility with our analysis pipeline. The three redundancy-filtered arrays obtained have contained clusters of UniProt IDs with a pre-set cut-off value (90%, 70% and 50% sequence identity).

Surveying protein sequences for CDRs

While SAHs – and CRRs in general – may contain high amounts of charged residues, their features mainly result from the specific patterns these residues repeat within the sequence. Thus, a separate set of investigations must be carried out if we are to expand the analysis of the correlation between charged protein regions and phase separation. We applied two alternative approaches for the identification of such regions, the first one considers a region highly charged if its ratio of charged residues reaches a given threshold, while the other identifies sub-sequences whose overall charge significantly differs from zero. Based on this difference the recognized CDRs are denoted as either 'signed' and 'unsigned', resulting in sCDR and uCDR motifs, respectively.

The scoring scheme for sCDR assigns a score of +1 to arginines, histidines and lysines, and –1 to aspartic and glutamic acids, while the one for uCDRs assigns a score of 1 to all five of the aforementioned residues. In both cases, the investigated sequence window is considered to contain a CDR if the absolute value of its overall score divided by its length reaches a pre-set cut-off value.

To determine adequate cut-off values for all the used window sizes, we run the algorithm on a randomized version of the human proteome with a cut-off value of zero. The randomized sequences had the same amino acid composition and length distribution as the natural ones. Based on the obtained scores, two cut-off values were determined for each window size, one that yields roughly 5% of all hits and one that yields only about 1% – note that the discrete nature of scoring (i.e., steps of 1/16 between 0 and 1 for a window length of 16 residues), selection of a threshold is somewhat approximated. In the case of 5% thresholds, stricter criteria were applied. Out of the two values closest to 5% the one below it was selected for each window size. In the case of 1% thresholds, out of the two values closest to 1% the one above it was selected for each window size. These choices were based on the fact that this way the thresholds for the four window sizes were as close to 5% (and 1%) as possible, while the way they were selected remained uniform. Using these two cut-offs, the full proteome was scanned to identify signed and unsigned CDRs in the wild-type sequences (Table 3).

Table 3. List of threshold values used to select a fraction of windows that were compiled into CDRs. In the case of unsigned CDRs, the values mark the fraction of residues within the given window that have to be either negatively or positively charged. In the case of signed CDRs, the ratio given by the absolute value of the residues' overall charge and the window size must reach the threshold.

	Thresholds that approximate 5% of all hits		Thresholds that approximate 1% of all hits	
	Cut-off values	Fraction of hits	Cut-off values	Fraction of hits
sCDRs	5/16	3.44%	6/16	1.09%
	7/32	3.76%	8/32	1.79%
	10/64	4.34%	13/64	1.08%
	16/128	4.74%	25/128	1.03%
uCDRs	8/16	4.86%	9/16	1.77%
	14/32	4.62%	16/32	1.23%
	26/64	3.77%	29/64	1.11%
	48/128	4.59%	54/128	1.19%

The segments identified using the same window sizes were merged for each sequence. During the process, care was taken to ensure that the merged segments are still above the predefined threshold, which needed special treatment for signed CDRs where the overall charge should also be maintained. In the final step, all data were unified for all window sizes used, 16, 32, 64 and 128.

Conducting Fisher's exact test of independence on members – and clusters – of the human proteome

To assess the association of specific sequence sets with LLPS, 2×2 contingency tables were generated and then Fisher's exact test as implemented in R [R statistics package reference] was applied to determine the P-value.

In the case of clusters, two different methods were applied for categorization. In the first one, evaluation is applied cluster-wise, meaning that association with LLPS for a given cluster was considered positive if any member (sequence) in the cluster was annotated as such in PhaSepDB. In the second one, a cluster was only considered to be participating in LLPS if the representative sequence kept by CD-HIT was annotated in PhaSepDB. Thus, in most of the investigations described in this article any protein's association with LLPS was based on its presence within the PhaSepDB2.0 database that has seen an overhaul in the way it includes entries. Prior to this 2021 update, admission had only required a source that confirmed a protein's localization within an MLO, regardless of its function. In the update, this was changed so that admission requires experimental evidence of the protein's participation in phase transitions, either as a driver or a recruited macromolecule.

Additionally, PhaSePro was utilized in the case studies as its detailed description of LLPS driver segments could be compared to the charged sequence motifs resulting from the surveys detailed above. The presence or absence of CRRs was established in an analogous manner.

Preparing ROC tests

The purpose of ROC tests is to evaluate an attribute as a viable indicator for classification. In the case of this study, they were used to investigate the presence of different sequence motifs as indicators for a protein's affinity towards LLPS. We have sorted the proteins in the human proteome according to their score obtained in the CDR/CRR detection and calculated the true positive, true negative, false positive and false negative rates.

Acknowledgements

This work was supported by a grant from the National Research, Development and Innovation Office under grant OTKA 124363 to Z.G. and FK-128133 to R.P., and a fellowship (grant number 158534) from the Tempus Public Foundation of the Hungarian government to R.P. A.S. acknowledges the support by an ÚNKP Fellowship.

Author contributions

ZG and PR designed the study, ALS developed software, ALS and AS analysed the data, all authors contributed to data interpretation, drafted the article and approved the final version.

Data accessibility

Tables summarizing the predictions and calculations supporting the findings of our study are available in the [Supplementary Material](#) of this article. Detailed prediction outputs are available from the corresponding author [gaspari.zoltan@itk.ppke.hu] upon reasonable request.

References

- Liu X, Liu X, Wang H, Dou Z, Ruan K, Hill DL, et al. Phase separation drives decision making in cell division. *J Biol Chem.* 2020;**295**:13419–31.
- Mitrea DM, Kriwacki RW. Phase separation in biology; functional organization of a higher order. *Cell Commun Signal.* 2016;**14**(1):1.
- Feng Z, Chen X, Zeng M, Zhang M. Phase separation as a mechanism for assembling dynamic postsynaptic density signalling complexes. *Curr Opin Neurobiol.* 2019;**57**:1–8.

- 4 Boeynaems S, Alberti S, Fawzi NL, Mittag T, Polymenidou M, Rousseau F, et al. Protein phase separation: a new phase in cell biology. *Trends Cell Biol.* 2018;**28**:420–35.
- 5 Shin Y, Brangwynne CP. Liquid phase condensation in cell physiology and disease. *Science.* 2017;**357**:eaaf4382.
- 6 Banani SF, Rice AM, Peeples WB, Lin Y, Jain S, Parker R, et al. Compositional control of phase-separated cellular bodies. *Cell.* 2016;**166**:651–63.
- 7 Molliex A, Temirov J, Lee J, Coughlin M, Kanagaraj AP, Kim HJ, et al. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell.* 2015;**163**:123–33.
- 8 Pancsa R, Vranken W, Mészáros B. Computational resources for identifying and describing proteins driving liquid-liquid phase separation. *Brief Bioinform.* 2021;**22**:1–20.
- 9 Lin Y, Currie SL, Rosen MK. Intrinsically disordered sequences enable modulation of protein phase separation through distributed tyrosine motifs. *J Biol Chem.* 2017;**292**:19110–20.
- 10 Li HR, Chiang WC, Chou PC, Wang WJ, Huang JR. TAR DNA-binding protein 43 (TDP-43) liquid-liquid phase separation is mediated by just a few aromatic residues. *J Biol Chem.* 2018;**293**:6090–8.
- 11 Pak CW, Kosno M, Holehouse AS, Padrick SB, Mittal A, Ali R, et al. Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein. *Mol Cell.* 2016;**63**:72–85.
- 12 Greig JA, Nguyen TA, Lee M, Holehouse AS, Posey AE, Pappu RV, et al. Arginine-enriched mixed-charged domains provide cohesion for nuclear speckle condensation. *Mol Cell.* 2020;**77**:1237–50.
- 13 You K, Huang Q, Chunyu Y, Shen B, Sevilla C, Shi M, et al. PhaSepDB: a database of liquid-liquid phase separation related proteins. *Nucleic Acids Res.* 2020;**48**:D354–9.
- 14 Nunes C, Mestre I, Marcelo A, Koppenol R, Matos CA, Nóbrega C. MSGP: the first database of the protein components of the mammalian stress granules. *Database.* 2019;**2019**:baz031.
- 15 Mészáros B, Erdős G, Szabó B, Schád É, Tantos Á, Abukhairan R, et al. PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res.* 2020;**48**(D1):D360–7.
- 16 Süveges D, Gáspári Z, Tóth G, Nyitray L. Charged single alpha-helix: a versatile protein structural motif. *Proteins.* 2009;**74**:905–16.
- 17 Barnes CA, Shen Y, Ying J, Takagi Y, Torchia DA, Sellers JR, et al. Remarkable rigidity of the single α -helical domain of myosin-VI As revealed by NMR spectroscopy. *J Am Chem Soc.* 2019;**141**:9004–17.
- 18 Wolny M, Batchelor M, Bartlett GJ, Baker EG, Kurzawa M, Knight PJ, et al. Characterization of long and stable *de novo* single alpha-helix domains provides novel insight into their stability. *Sci Rep.* 2017;**7**:44341.
- 19 Peckham M, Knight PJ. When a predicted coiled coil is really a single α -helix, in myosins and other proteins. *Soft Matter.* 2009;**5**:2493–503.
- 20 Luo H, Nijveen H. Understanding and identifying amino acid repeats. *Brief Bioinform.* 2014;**15**:582–91.
- 21 Martínez J, Ren YG, Thuresson AC, Hellman U, Astrom J, Virtanen A. A 54-kDa fragment of the Poly(A)-specific ribonuclease is an oligomeric, processive, and cap-interacting Poly(A)-specific 3' exonuclease. *RNA.* 2000;**275**:P24222–30.
- 22 Kim JH, Shinde DN, Reijnders MRF, Hauser NS, Belmonte RL, Wilson GR, et al. De novo mutations in SON disrupt RNA splicing of genes essential for brain development and metabolism, causing an intellectual-disability syndrome. *Am J Hum Genet.* 2016;**99**:711–9.
- 23 Heyd F, Lynch KW. Phosphorylation-dependent regulation of PSF by GSK3 controls CD45 alternative splicing. *Mol Cell.* 2010;**40**:126–37.
- 24 Klar M, Bode J. Enhanceosome formation over the beta interferon promoter underlies a remote-control mechanism mediated by YY1 and YY2. *Mol Cell Biol.* 2005;**25**:10159–70.
- 25 Wang S, Zhang B, Faller DV. Prohibitin requires Brg-1 and Brm for the repression of E2F and cell growth. *EMBO J.* 2002;**21**:3019–28.
- 26 Lee ASY, Kranzusch PJ, Cate JHD. eIF3 targets cell-proliferation messenger RNAs for translational activation or repression. *Nature.* 2015;**522**:111–4.
- 27 Gaudet P, Livstone MS, Lewis SE, Thomas PD. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief Bioinform.* 2011;**12**:449–62.
- 28 Pardo R, Molina-Calavita M, Poizat G, Keryer G, Humbert S, Saudou F. pARIS-htt: an optimised expression platform to study huntingtin reveals functional domains required for vesicular trafficking. *Mol Brain.* 2010;**3**:17.
- 29 Liang J, Wang J, Azfer A, Song W, Tromp G, Kolattukudy PE, et al. A novel CCCH-zinc finger protein family regulates proinflammatory activation of macrophages. *J Biol Chem.* 2008;**283**:6337–46.
- 30 Yamada K, Kawata H, Matsuura K, Shou Z, Hirano S, Mizutani T, et al. Functional analysis and the molecular dissection of zinc-fingers and homeoboxes 1 (ZHX1). *Biochem Biophys Res Commun.* 2002;**297**:368–74.
- 31 Egan ED, Collins K. An enhanced H/ACA RNP assembly mechanism for human telomerase RNA. *Mol Cell Biol.* 2012;**32**:2428–39.
- 32 Okuwaki M, Tsujimoto M, Nagata K. The RNA binding activity of a ribosome biogenesis factor, nucleophosmin/B23, is modulated by phosphorylation

- with a cell cycle-dependent kinase and by association with its subtype. *Mol Biol Cell*. 2002;**13**:2016–30.
- 33 Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, et al. Insight into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*. 2012;**149**:1393–406.
- 34 Baltz AG, Munschauer M, Schwanhäusser B, Vasile A, Murakawa Y, Schueler M, et al. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell*. 2012;**46**:674–90.
- 35 Ahn EY, DeKelver RC, Lo MC, Nguyen TA, Matsuura S, Boyapati A, et al. SON controls cell-cycle progression by coordinated regulation of RNA splicing. *Mol Cell*. 2011;**42**:185–98.
- 36 Jeon Y, Lee JT. YY1 tethers Xist RNA to the inactive X nucleation center. *Cell*. 2011;**146**:119–33.
- 37 Pereira B, Sousa S, Barros R, Carreto L, Oliveira P, Oliveira C, et al. CDX2 regulation by the RNA-binding protein MEX3A: impact on intestinal differentiation and stemness. *Nucleic Acids Res*. 2013;**41**:3986–99.
- 38 Fritz J, Strehblow A, Taschner A, Schopoff S, Pasierbek P, Jantsch MF. RNA-regulated interaction of transportin-1 and exportin-5 with the double-stranded RNA-binding domain regulates nucleocytoplasmic shuttling of ADAR1. *Mol Cell Biol*. 2009;**29**:1487–97.
- 39 Zhang Z, Theler D, Kaminska KH, Hiller M, de la Grange P, Pudimat R, et al. The YTH domain is a novel RNA binding domain. *J Biol Chem*. 2010;**285**:14701–10.
- 40 Bochar DA, Wang L, Beniya H, Inev A, Xue Y, Lane WS, et al. BRCA1 is associated with a human SWI/SNF-related complex. *Cell*. 2000;**102**:257–65.
- 41 Sewer MB, Nguyen VQ, Huang CJ, Tucker PW, Kagawa N, Waterman MR. Transcriptional activation of human CYP17 in H295R adrenocortical cells depends on complex formation among p54^{nrb}/NonO, protein-associated splicing factor, and SF-1, a complex that also participates in repression of transcription. *Endocrinology*. 2002;**143**:1280–90.
- 42 Passon DM, Lee M, Rackham O, Stanley WA, Sadowska A, Filipovska A, et al. Structure of heterodimer of human NONO and paraspeckle protein component 1 and analysis of its role of subnuclear body formation. *PNAS*. 2012;**109**:4846–50.
- 43 Dobson L, Nyitray L, Gáspári Z. A conserved charged single α -helix with a putative steric role in paraspeckle formation. *RNA*. 2015;**21**:2023–9.
- 44 Zipprich JT, Bhattacharyya S, Mathys H, Filipowicz W. Importance of the C-terminal domain of the human GW182 protein TNRC6C for translational repression. *RNA*. 2009;**15**:781–93.
- 45 Mattioli F, Vissers JHA, van Dijk WJ, Ikpa P, Citterio E, Vermeulen W, et al. RNF168 ubiquitinates K13–15 on H2A/H2AX to drive DNA damage signaling. *Cell*. 2012;**150**:1182–95.
- 46 Zeng M, Chen X, Guan D, Xu J, Wu H, Tong P, et al. Reconstituted postsynaptic density as a molecular platform for understanding synapse formation and plasticity. *Cell*. 2018;**174**:1172–87.
- 47 Muddashetty RS, Nalavadi VC, Bassel GJ, Yao X, Xing L, Laur O, et al. Reversible inhibition of PSD-95 mRNA translation by miR-125a, FMRP phosphorylation and mGluR signaling. *Mol Cell*. 2011;**42**:673–88.
- 48 Milovanovic D, Wu Y, Bian X, De Camilli P. A liquid phase of synapsin and lipid vesicles. *Science*. 2018;**361**:604–7.
- 49 Sabari BR, Dall'Agnese A, Boija A, Klein IA, Coffey EL, Shrinivas K, et al. Coactivator condensation at super-enhancers links phase separation and gene control. *Science*. 2018;**361**:eaar3958.
- 50 Gáspári Z, Süveges D, Perczel A, Nyitray L, Tóth G. Charged single alpha-helices in proteomes revealed by a consensus prediction approach. *BBA*. 2012;**1824**:637–46.
- 51 Kovács Á, Dudola D, Nyitray L, Tóth G, Nagy Z, Gáspári Z. Detection of single alpha-helices in large protein sequence sets using hardware acceleration. *J Struct Biol*. 2018;**204**:109–16.
- 52 Dudola D, Tóth G, Nyitray L, Gáspári Z. Consensus prediction of charged single alpha-helices with CSAHserver. *Methods Mol Biol*. 2017;**1484**:25–34.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Association of CRR/SAH/CDR-containing sequences with LLPS.

Table S2. Statistics with and without transmembrane proteins and additional data.

Table S3. Gene ontology analysis of CRR/SAH/CDR-containing sequences related and unrelated to LLPS.

Table S4. Statistics for random sequence sets selected to match the length distribution of proteins in Pha-SepDB.

Table S5. Charged regions and LLPS-associated segments in selected proteins.

Table S6. Charged motifs, PSD localization and RNA binding in all proteins of the human reference proteome.

Table S7. Sequence characteristics of CRRs and CDRs.

Fig. S1. Decimal logarithm of P-values obtained from Fisher's exact test of independence, carried out respectively on pairs of attributes where one attribute was the association to LLPS (whether the sequence was present in Pha-SepDB) and the other attribute was the presence of a given type of charged sequence motif.

Fig. S2. Number (top row) and fraction (bottom row) of sequences with charged motifs within the human reference proteome (red), PhaSepDB (green), ten smaller sequence sets where one sequence with a similar length (5%) has been randomly selected for each human PhaSepDB entry (blue), and one sequence set with ten randomly selected, similarly long sequences per PhaSepDB entry (light blue).

Fig. S3. Categorization of sequences based on their charged motif-content, association to LLPS and relation to the postsynaptic density (PSD).

Fig. S4. Number of entries within each category shown on Figure S3.

Fig. S5. Decimal logarithm of the number of entries within each category shown on Figure S3.

Fig. S6. Categorization of sequences based on their charged motif-content, association to LLPS and ability to bind RNAs.

Fig. S7. Number of entries within each category of Figure S6.

Fig. S8. Decimal logarithm of the number of entries within each category of Figure S6.

Fig. S9. Categorization of all human proteins (including transmembrane proteins) according to their special charged motif- (a-b) or CRR-content (c-d), LLPS-association and their functional role in the PSD (a, c) or RNA-binding (b, d).

Fig. S10. Number of FT_CHARGE hits organized by their Fourier frequency (1/64 - 32/64).