Multiple polyvalency provided by intrinsically disordered segments is a key feature of postsynaptic scaffold proteins

Annamária Kiss-Tóth, Bálint Péterfia, Annamária F. Ángyán, Balázs Ligeti, Gergely Lukács, Zoltán Gáspári

Abstract

The postsynaptic density, a key regulator of the molecular events of learning and memory is composed of an elaborate network of interacting proteins capable of dynamic reorganization. Despite our growing knowledge on specific proteins and their interactions, atomic-level details of its full three-dimensional structure and its rearrangements are still largely elusive. In this work we addressed the extent and possible role of intrinsic disorder in postsynaptic proteins in a detailed in silico analysis. Using a strict consensus of predicted intrinsic disorder and a number of other protein sets as controls, we show that postsynaptic proteins are particularly enriched in disordered segments. Although the number of interacting partner proteins is not exceptionally large, the estimated diversity of the combinations of putative complexes is high in postsynaptic proteins.

Introduction

The postsynaptic density (PSD) is a characteristic part of excitatory chemical synapses. It forms a disk-shaped electrodense entity about 200-800 nm wide and 30-50 nm thick beneath the postsynaptic membrane. It is composed of a dense network of proteins and provides a complex link between the intracellular parts of membrane receptors and adhesion molecules and the cytoskeleton (1). In-cell studies strongly suggest that the PSD is a highly organized molecular network, possibly with spatially distinct 'nanodomains' with functional relevance such as AMPA receptor positioning and anchoring (2, 3, 4). The emerging view is that the PSD is constantly remodeled and has a characteristic dynamics that is manifested not only by the addition and elimination of components over time but also dynamic restructuring. This strongly suggests the presence of a number of intermolecular interacting sites with an elaborate and thoroughly regulated distribution of occupied, unavailable and available partner binding sites with a high level of redundancy. The PSD can most likely be imagined as a supramolecular machine capable of integrating and transmitting signals *via* dynamic reorganization. The underlying mechanisms of this most likely include competitive binding events, allostery and cooperativity at the level of proteins and the network might act as an amplifier and integrator of these. Recent experimental observations suggest the possible role of phase transitions in both post- and presynaptic processes (5, 6).

Protein intrinsic disorder, defined by the lack of adopting a stable three-dimensional structure of a particular protein or a segment of it, is considered as a key factor in mediating weak yet specific protein:protein interactions. The principal working model for intrinsically disordered regions (IDRs) in this respect is folding upon binding, i.e. the full or partial - both in terms of locality and residual flexibility - ordering of a disordered region when forming a complex with a binding partner (7). Besides other roles, such segments can act as binding sites for structural domains, recognition sites for posttranslational modifications such as phosphorylation, and flexible linkers between globular domains contributing the structural versatility of the complexes formed (8). Disordered segments have been shown to play important roles in scaffold proteins (9), signaling pathways (7, 10) and be at key positions within interaction networks (8). In particular, the synaptosome has been listed as an intracellular component associated with the presence of intrinsic protein disorder (11). Multivalent proteins containing a number of similar interaction sites along with multiple domains have been suggested to be key for phase separation in protein complexes (12).

In this study we investigated the idea whether intrinsic disorder might play an important role in PSD proteins with respect to the organization, versatility and the ability for dynamic rearrangement of the PSD. Whereas there are domain types, such as the PDZ domain, that have been described to be abundant in postsynaptic proteins, we are not aware of any detailed analysis on the role of disorder in the synaptome. Using the human proteome available in UniProt, we identified IDRs using a consensus of different prediction methods and filtering out oligomeric fibrillar regions (13) to extract segments likely to be intrinsically disordered under cellular conditions.

Results and Discussion

Proteins with the highest number of IDRs and protein:protein interaction regions as identified with ANCHOR (14) segments were found to be associated with Gene Ontology terms related to synaptic transmission and neural development. Protein length and overall percentage of residues in IDRs results in a lower number of related associated terms. We note that the most prominent GO categories are related to histone methylation, chromatin remodeling and other nuclear processes, as well as cytoskeletal functions. As a characteristic and relatively concise example, results of the GO-SLIM cellular component analysis are shown in Table 1.

Table 1. Summary of PANTHER GO-SLIM cellular component analysis of the proteins with highest values of selected simple descriptors. The most specific terms that are enriched in the sets are listed. (+) denotes a term for which enrichment was observed but was not the most specific reported one for the given set in the term hierarchy.

GO term	Protein length	Overall disorder percentage	No. of. IDRs	No. of ANCHOR regions
Descriptor value	>=1830 residues	>=54 %	>= 6	>= 23
No. of proteins in the list	593	1221	698	640
Cell junction	+			
Extracellular matrix	+			+
Cytoskeleton	+	(+)	+	(+)
Tubulin complex		+		
Nucleoplasm		+	+	+
Nuclear chromosome		+		+
Actin cytoskeleton		+		
Apical part of cell			+	+
Postsynaptic membrane			+	+

Cell junction			+	+
Chromosome			+	(+)
Nuclear envelope				+
Microtubule				+

To get further insight into the relevance of these observations, we have investigated protein sets derived from the human proteome reference set in UniProt. Four sets defined in SynaptomeDB (15) along with sets defined based on the Gene Ontology (GO) terms *Chemical synaptic transmission’* and ‘*postsynaptic density’*, as well as a small set obtained from the UniProt website with a simple search for ‘postsynaptic scaffold’. Reference sets include the full human proteome, the immunome as a set of primarily globular multidomain proteins (16), proteins involved in signal transduction pathways (17), and cytoskeletal proteins (18). The latter two sets were chosen because of their connection with the function and organization of the synaptome. We have also included histone methylases, known to have high IDR content (19) and a wider list of proteins in the nucleus (20), where phase separation has also been observed (21). All calculated descriptors for each protein along with the functional sets are available in Table S1. Basic properties of the sets are summarized in Tables S2 and S3.

We calculated various descriptors related to the presence and functional role of IDRs, including the length, percentage of residues in IDRs, number of IDRs and number of binding segments identified with ANCHOR. In addition, the number of interacting partner proteins in the ComPPI database (22) and in the BioPlex 2.0 dataset (23) were also considered. To assess the potential of the proteins to be engaged in multivalent interactions, we introduced a descriptor called ‘*diversity of potential interactions*’, *DPI*, defined as:

$$DPI = DPI_{disorder} + DPI_{domain} = \sum_{e=1}^{Ne} ln(ELM_e) + \sum_{d=1}^{Nd} ln(DOM_d)$$

Where *Ne* and *Nd* denote the number of linear motifs based on the Eukaryotic Linear Motif (ELM) resource (24) and domain types, whereas ELM_e and DOM_d correspond to the number of ELM sites of type *e* and domains of type *d*, respectively. Only predicted ELM sites located in the identified disordered regions and of classes LIG/DOC were considered. We are aware that ELM prediction and also simple domain counts clearly overestimate the number of actual binding sites but we use DPI as a comparative rather than an absolute measure related to the upper limit of potential interaction sites in the protein sets examined.

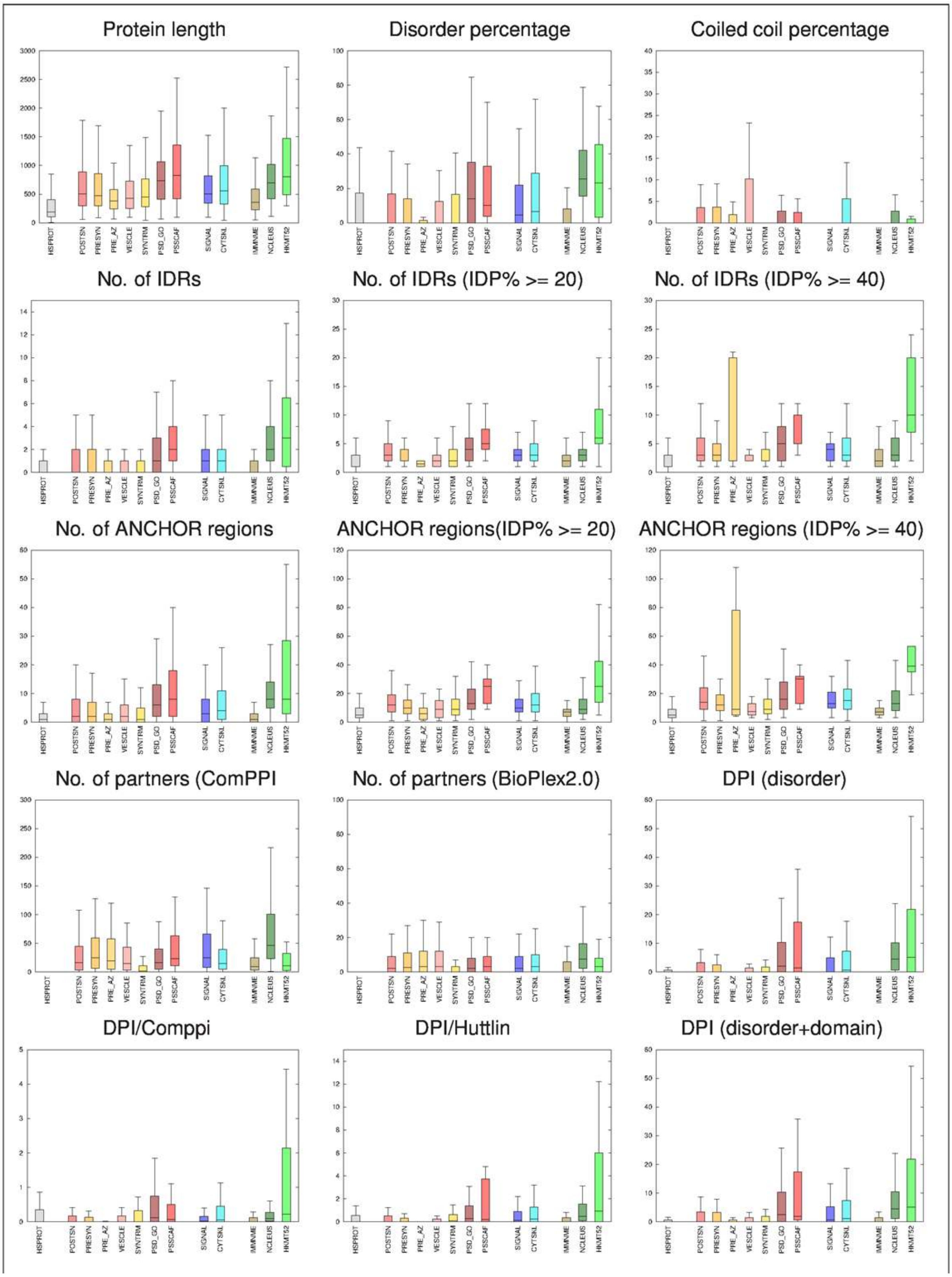

Figure 1. Box plots of selected descriptors in the protein sets investigated. Sets are non-exlcusive (a protein can be present in multiple sets).

Postsynaptic proteins tend to be longer on average than other proteins in the human proteome except presynaptic ones and histone methylases although the standard deviation is large in all categories (Table S4). This length difference is also maintained in proteins with high disorder content. The extent of overall disorder and the number of disordered segments is also large in postsynaptic proteins, and also in nuclear and cyoteskeletal proteins and the histone methylase set. Only the among histone methylases is the percentage of proteins with at least 5 intrinsically disordered segments larger than in postsynaptic scaffold proteins.

PSD scaffold proteins also exhibit high percentage of residues in ANCHOR-defined segments, and again only histone methylases contain more sequences with high number of ANCHOR segments (Table S4). The descriptor that distinguishes postsynaptic scaffold proteins and histone methylases from other proteins and sets is the DPI measure. This trend is robust as it is present in all proteins and when considering sequences with high disorder content only (Table S4). We note that the differences mainly stem from $DPI_{disorder}$, with DPI_{domain} showing an observable contribution to the *DPI* of postsynaptic scaffold proteins.

It should be noted that these trends are also observable if we omit proteins assigned to multiple functional sets (Figure S2), or consider a nonredundant part of the human proteome (Figure S4) or only proteins in SwissProt (Figure S5). For the latter two analyses, results are shown for all proteins allowing overlapping functional sets as otherwise the number of proteins in some of the sets would be extremely low (Table S2).

We have also investigated the binding regions in orthologous proteins. Orthologs of synaptic proteins were identified in 4 species: Western lowland gorilla (Gorilla gorilla gorilla), Sumatran orangutan (Pongo abelii), common chimpanzee (Pan troglodytes) and house mouse (Mus musculus). We found that the predicted binding regions are largely conserved among these species.

Correlation between descriptors

Table S5 contains correlations between selected descriptors, focusing on the overall disorder content, number of disordered regions and diversity of potential interactions (DPI).

Somewhat surprisingly, the number of disordered regions does not or only moderately correlate with the overall disorder percentage. There is, not unexpectedly, a relatively high correlation when all sequences, i.e. those with negligible disorder content are also considered, but analysis of proteins with 20 or 40% of overall disorder content reveals that this stems from the trivial connection that zero disorder content is trivially associated with the absence of any disordered regions. In largely disordered proteins, i.e. those with over 40% of residues in disordered regions, there is practically no correlation between these two descriptors.

The DPI descriptor is, in most datasets, also uncorrelated with overall disorder content, again with the exception of the trivial cases where proteins with no or very low disorder content are also considered. Only proteins associated with presynaptic vesicles (VESCLE) and histone methylases (HKMT52) show meaningful correlation (above 0.7) between overall disorder content and DPI.

In contrast, the number of disordered regions shows remarkable correlation with DPI in most groups. With the exception of the VESCLE and immunome (IMMNME) groups, this correlation is maintained in the proteins with high disorder content.

These observations suggest that the number of disordered regions, rather than overall disorder content, is key in determining the versatility of interactions in most proteins investigated.

Conclusions

We propose that the high number of IDRs and the associated large DPI is an important aspect of providing the versatility of PSD complexes both across cell types and in different cellular states. We denote this feature multiple polyvalency, which is also a likely key feature of proteins and protein complexes capable of phase separation, as observed for PSD proteins, synapsin and a number of complexes involved in transcription.

Methods

Data sets

The UniProt database was used as the sole source of all protein sequences investigated in the study. In the case of specific protein sets, the UniProt IDs of the individual proteins were extracted and used thereafter. As the main data source, the human proteome set of UniProt (organism: *Homo sapiens* [9606], proteome ID:up 000005640, consisting of 71607 sequences) was downloaded and all specific protein sets were defined as a subset of this (25). A nonredundant version of the full proteome was compiled as follows. First, residues in the non-globular parts of the sequences, as defined by the disorder, coiled coil, collagen and CSAH predictions as described above, were masked with ‘x’, then CD-HIT (26) was invoked using on these with a similarity cutoff of 70%. This resulted in 25682 proteins that have nonredundant globular regions.

Specific protein sets were obtained from various sources, all of them defined as a subset of the full proteome (i.e. no additional proteins were added even if some external sets/searches contained such). For synaptic proteins, the SynaptomeDB (15) database and its classification of protein localization was used, defining four sets: postsynaptic, presynaptic, presynaptic active zone and vesicle-associated proteins. Two additional PSDS-related sets were compiled based on Gene Ontology annotations, one using the localization term ‘postsynaptic density’ and the other the functional assignment ‘chemical synaptic transmission’. The hierarchical nature of the GO annotation was taken into account by selecting all proteins for which these terms or any of their descendants were found in the UniProt annotation. A third PSD-related set was defined based on a simple search on the UniProt website with the term ‘postsynaptic scaffold human’. Proteins of the immunome were extracted from the Immunome Knowledge Base (16), signal transduction proteins from the SignaLink database (17), and cytoskeletal proteins (18). The list of nuclear proteins were taken from the supplementary material of Frege et al. 2015 (20), whereas the set of histone methylases were extracted from the accompanying data of Lazar et al. 2016 (19). All sets are assigned a 6-character identifier (Table S1). It should be noted that many proteins belong to more than one sets (Table S2). For some analyses proteins assigned to multiple sets were excluded, this is always indicated in the text.

Lists of interacting protein pairs were taken from either of two sources: the Comppi database and the comprehensive BioPlex 2.0 experiment.

Orthologs were collected from 2 different databases, eggNOG (27) and OMA (28). Comparing the 2 databases, eggNOG yields a slightly higher number of related proteins. Since there are no contradictions between the 2 databases, the union of the orthologs identified in the two databases were used. OMA uses a special identifier, eggNOG uses ENSEMBL IDs for the sequences. They were both converted into Uniprot IDs.

Disorder prediction & analysis

Disorder prediction was done in a way that aims to minimize false positive hits. First, the consensus of two disorder prediction methods, IUPred (29) and VSL2B (30) were determined. In the second step, all residues predicted to be in oligomeric fibrillar motifs were eliminated from the set of disordered residues (31). Oligomeric fibrillar motifs were determined using a permissive prediction, namely, residues to form coiled coils as predicted either by COILS (32) or Paircoil2 (33), single alpha helices as identified using FT_CHARGE (34) or collagen triple helix as obtained by HMMER (35) with the Pfam HMM for collagen (ID: PF01391.13) segments in Pfam (36). The remaining set of residues is considered to be a good representation of segments being disordered under cellular conditions.

Binding regions within disordered segments were identified with ANCHOR (14). It should be noted that the ANCHOR-predicted set of regions is much more restricted than the set of IDRs identified with the methodology described above, and can not be considered as a subset of it. Thus, the two sets rather complement each other in the sense that they arise from different concepts.

The number of linear motifs participating in actual binding events was estimated using the regular expressions listed in the ELM database. It should be noted that we are aware that this treatment of ELM data results in massive overprediction of binding motifs, but we only use the obtained results in a comparative context. Besides using all ELM classes, a sub-section of LIG and DOC motif classes was also used. We calculated the overall number of motifs as well as the number of different motif types per protein, as well as a diversity measure termed '*diversity of potential interactions*', *DPI*, defined as:

$$DPI = DPI_{disorder} + DPI_{domain} = \sum_{e=1}^{Ne} ln(ELM_e) + \sum_{d=1}^{Nd} ln(DOM_d)$$

Where *Ne* and *Nd* denote the number of ELM classes and domain types, whereas ELM_e and DOM_d correspond to the number of ELM sites of type *e* and domains of type *d*, respectively. Only predicted ELM sites located in the identified disordered regions and of classes LIG/DOC were considered. We are aware that ELM prediction and also simple domain counts clearly overestimate the number of actual binding sites but we use DPI as a comparative rather than an absolute measure related to the upper limit of potential interaction sites in the protein sets examined.

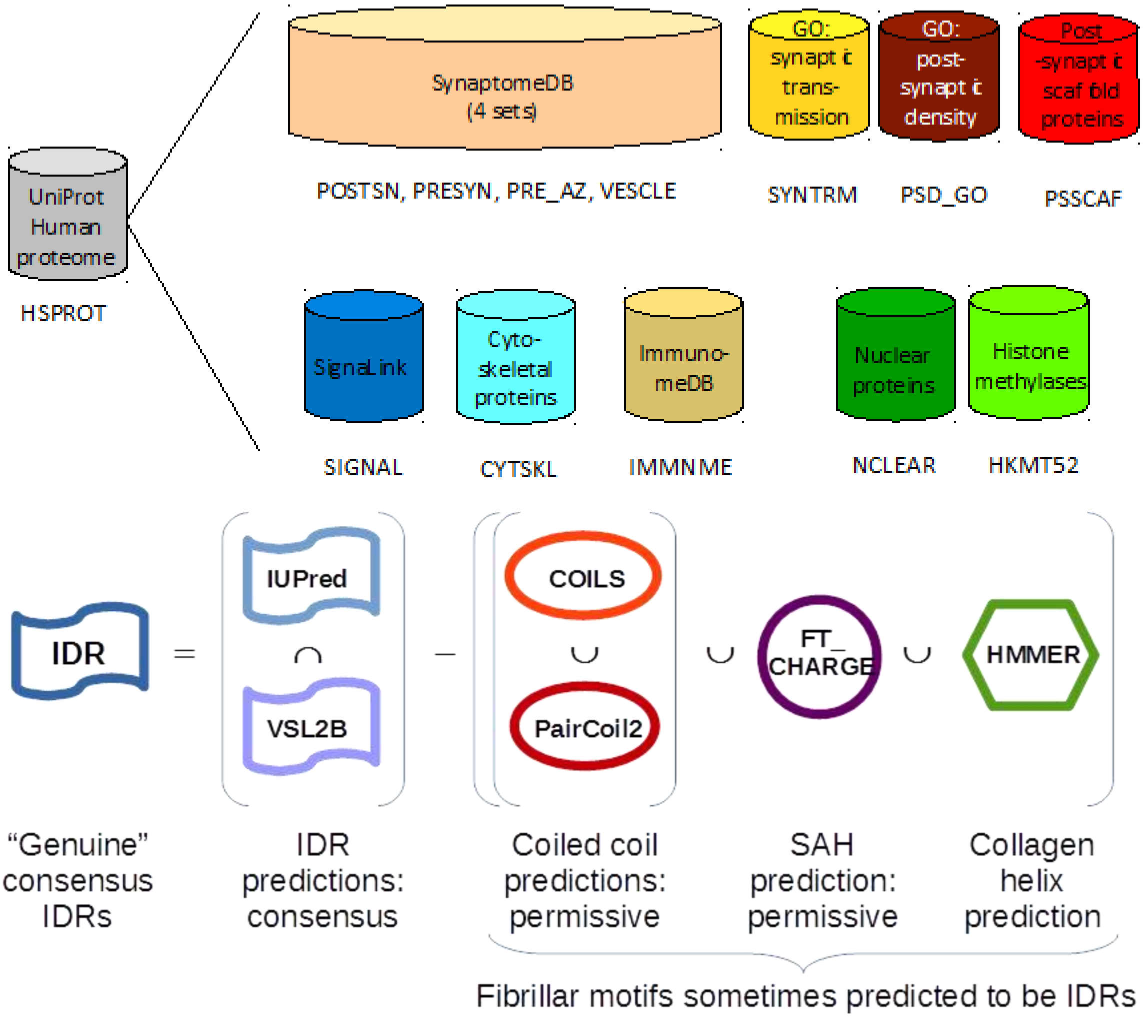
UniProt Human proteome
HSPROT
SynaptomeDB (4 sets)
POSTSN, PRESYN, PRE_AZ, VESCLE
GO: synaptic trans-mission
SYNTRM
GO: post-synaptic density
PSD_GO
Post-synaptic scaffold proteins
PSSCAF
SignaLink
SIGNAL
Cyto-skeletal proteins
CYTSKL
Immuno-meDB
IMMNME
Nuclear proteins
NCLEAR
Histone methylases
HKMT52
IDR
=
IUPred
∩
VSL2B
−
COILS
∪
PairCoil2
∪
FT_CHARGE
∪
HMMER
"Genuine" consensus IDRs
IDR predictions: consensus
Coiled coil predictions: permissive
SAH prediction: permissive
Collagen helix prediction
Fibrillar motifs sometimes predicted to be IDRs

Figure S1. Top: Overview of the data sets used in the study. Bottom: schematic workflow for identifying IDRs excluding those in predicted fibrillar regions

Gene ontology analysis

The human proteome dataset was filtered according to the UniProt IDs listed in the human reference protein set if the PANTHER database (37), obtained from the PATHERDB web site (ftp://ftp.pantherdb.org/sequence_classifications/current_release/PANTHER_Sequence_Classification_files/PTHR13.1_human, last modified 02/01/2018), resulting in 19781 proteins. In this case, to avoid any bias originating from multiple consideration of UniProt entries that were merged after the release of the PANTHER list, only the primary accessions were considered. Average and standard deviation of specific descriptors (length, percentage of residues in IDRs, number of IDRS, number of ANCHOR sites) of the proteins in this list was calculated and proteins with higher descriptor value than the average plus two standard deviations were analyzed with PANTHER accessed through the AmiGO2 website (amigo.geneontology.org/amigo) (38).

Acknowledgements

The authors acknowledge the support of the National Research, Development and Innovation Office – NKFIH through grants no. NN124363 (to ZG) and the 'János Bolyai Research Scholarship' Program of the Hungarian Academy of Sciences (to ZG). The research has been carried out within the project Thematic Research Cooperation Establishing Innovative Informatic and Info-communication Solutions, which has been supported by the European Union and co-financed by the European Social Fund under grant number EFOP-3.6.2-16-2017-00013.

References

1. Feng, W. & Zhang, M. Organization and dynamics of PDZ-domain-related supramodules in the postsynaptic density. *Nature Reviews Neuroscience* **10,** 87–99 (2009).
2. Macgillavry, H. D., Song, Y., Raghavachari, S. & Blanpied, T. A. Nanoscale scaffolding domains within the postsynaptic density concentrate synaptic AMPA receptors. *Neuron* **78,** 615–622 (2013).
3. Nair, D. *et al.* Super-resolution imaging reveals that AMPA receptors inside synapses are dynamically organized in nanodomains regulated by PSD95. *Journal of Neuroscience* **33,** 13204–13224 (2013).
4. Smith, K. R. et al. Psychiatric Risk Factor ANK3/Ankyrin-G Nanodomains Regulate the Structure and Function of Glutamatergic Synapses. *Neuron* **84**, 399–415 (2014).
5. Zeng, M. *et al.* Phase transition in postsynaptic densities underlies formation of synaptic complexes and synaptic plasticity. *Cell* **166,** (2016).
6. Milovanovic, D., Wu, Y., Bian, X. & Camilli, P. D. A liquid phase of synapsin and lipid vesicles. *Science* **361,** 604–607 (2018).
7. Csizmok, V., Follis, A. V., Kriwacki, R. W. & Forman-Kay, J. D. Dynamic protein interaction networks and new structural paradigms in signaling. *Chemical Reviews* **116,** 6424–6462 (2016).
8. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS Journal* **272,** 5129–5148 (2005).
9. Cortese, M. S., Uversky, V. N. & Dunker, A. K. Intrinsic disorder in scaffold proteins: Getting more from less. *Progress in Biophysics and Molecular Biology* **98,** 85–106 (2008).
10. Tantos, A., Han, K.-H. & Tompa, P. Intrinsic disorder in cell signaling and gene transcription. *Molecular and Cellular Endocrinology* **348,** 457–465 (2012).
11. Vucetic, S. *et al.* Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *Journal of Proteome Research* **6,** 1899–1916 (2007).
12. Boeynaems, S. *et al.* Protein Phase Separation: A new phase in cell biology. *Trends in Cell Biology* **28,** 420–435 (2018).
13. Szappanos, B., Süveges, D., Nyitray, L., Perczel, A. & Gáspári, Z. Folded-unfolded cross-predictions and protein evolution: The case study of coiled-coils. *FEBS Letters* **584,** 1623–1627 (2010).
14. Dosztanyi, Z., Meszaros, B. & Simon, I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25,** 2745–2746 (2009).
15. Pirooznia, M. *et al.* SynaptomeDB: an ontology-based knowledgebase for synaptic genes. *Bioinformatics* **28,** 897–899 (2012).

16. Ortutay, C. & Vihinen, M. Immunome Knowledge Base (IKB): An integrated service for immunome research. *BMC Immunology* **10,** 3 (2009).
17. Fazekas, D. *et al.* SignaLink 2 – a signaling pathway resource with multi-layered regulatory networks. *BMC Systems Biology* **7,** 7 (2013).
18. Guharoy, M., Szabo, B., Martos, S. C., Kosol, S. & Tompa, P. Intrinsic structural disorder in cytoskeletal proteins. *Cytoskeleton* **70,** 550–571 (2013).
19. Lazar, T. *et al.* Intrinsic protein disorder in histone lysine methylation. *Biology Direct* **11,** (2016).
20. Frege, T. & Uversky, V. N. Intrinsically disordered proteins in the nucleus of human cells. *Biochemistry and Biophysics Reports* **1,** 33–51 (2015).
21. Plys, A. J. & Kingston, R. E. Dynamic condensates activate transcription. *Science* **361,** 329–330 (2018).
22. Veres, D. V. et al. ComPPI: a cellular compartment-specific database for protein–protein interaction network analysis. *Nucleic Acids Research* **43,** 485–493 (2014).
23. Huttlin, E. L. et al. Architecture of the human interactome defines protein communities and disease networks. *Nature* **545,** 505–509 (2017).
24. Dinkel, H. et al. ELM 2016—data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Research* **44,** (2015).
25. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45,** (2016).
26. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26,** 680–682 (2010).
27. Jensen, L. J. *et al.* eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research* **36,** (2007).
28. Schneider, A., Dessimoz, C. & Gonnet, G. H. OMA Browser Exploring orthologous relations across 352 complete genomes. *Bioinformatics* **23,** 2180–2182 (2007).
29. Dosztányi, Z., Csizmók, V., Tompa, P. & Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of Molecular Biology* **347,** 827–839 (2005).
30. Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K. & Obradovic, Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7,** 208 (2006).
31. Gáspári, Z. Is Five Percent Too Small? Analysis of the Overlaps between Disorder, Coiled Coil and Collagen Predictions in Complete Proteomes. *Proteomes* **2,** 72–83 (2014).
32. Lupas, A., Dyke, M. V. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).
33. Berger, B. et al. Predicting coiled coils by use of pairwise residue correlations. *Proceedings of the National Academy of Sciences* **92,** 8259–8263 (1995).
34. Kovács, Á. et al. Detection of single alpha-helices in large protein sequence sets using hardware acceleration. *Journal of Structural Biology* **204,** 109–116 (2018).
35. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39,** (2011).
36. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* **44,** (2015).
37. Mi, H. et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research* **45**, (2016).
38. Carbon, S. et al. AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**, 288–289 (2008).