Occurrence of ordered and disordered structural elements in postsynaptic proteins supports optimization for interaction diversity

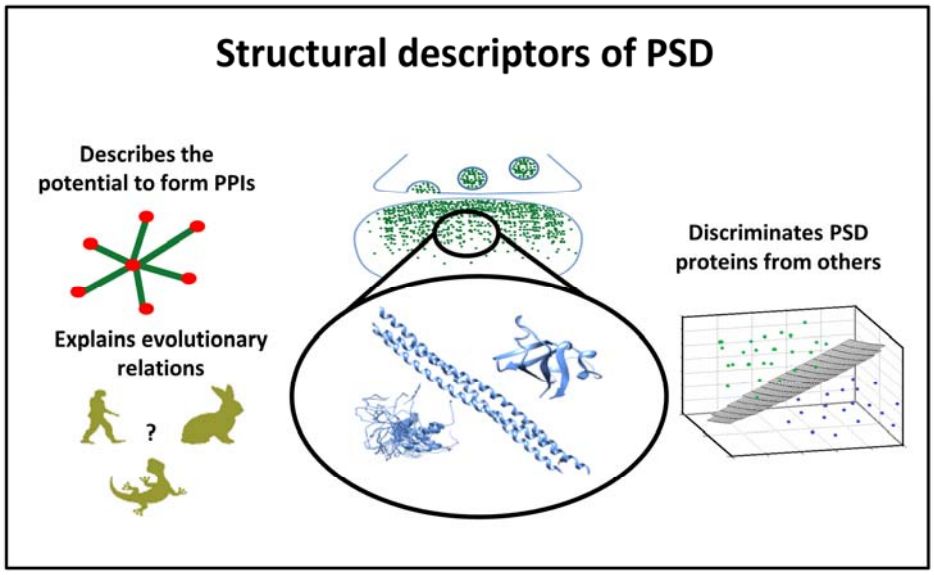
Annamária Kiss-Tóth[1,2], Laszlo Dobson[1], Bálint Péterfia[1], Annamária F. Ángyán[1], Balázs Ligeti[1], Gergely Lukács[1], Zoltán Gáspári[1]

[1]Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Práter u. 50A, 1083, Budapest, Hungary
[2] 3in-PPCU Research Group, Esztergom, 2500-Hungary

Abstract

The human postsynaptic density is an elaborate network comprising thousands of proteins, playing a vital role in the molecular events of learning and the formation of memory. Despite our growing knowledge of specific proteins and their interactions, atomic-level details of their full three-dimensional structure and their rearrangements are mostly elusive. Advancements in structural bioinformatics enabled us to depict the characteristic features of proteins involved in different processes aiding neurotransmission. We show that postsynaptic protein-protein interactions are mediated through the delicate balance of intrinsically disordered regions and folded domains, contributing to a high level of protein modularity. We introduce structure and regulation based descriptors to assess the diversity of interactions required to maintain the synaptic plasticity. Our approach reveals that the postsynaptic proteome has markedly different features compared to those of other proteins included in the human proteome.

Graphical abstract



Key terms

Postsynaptic density, Protein-protein interaction, Intrinsically disordered proteins, Protein modularity

Introduction

Synaptic signal transduction is an elaborate process that not only provides the means of transmitting the excited state of one neuron to another but also plays an important role in basic phenomena underlying neural development, learning and memory [1,2]. The basic cellular phenomenon of memory, long-term potentiation (LTP) results in the strengthening of synaptic connections. In the early phase of LTP, (approximately within 1 hour of its initiation) there are no changes in synaptic size and composition, whereas in the late phase novel proteins are synthesized and the size of the synapse is increasing [3,4].
The postsynaptic density (PSD) is an essential part of excitatory chemical synapses. It is composed of a dense network of proteins and provides an intricate link between the intracellular parts of membrane receptors and adhesion molecules and the cytoskeleton [5]. PSD composition and organization changes during development and the individual history of the particular neuron [6,7], leading to morphological differences between PSDs in different brain regions [8]. Recent results point to a reorganization of the PSD during sleep by the exchange of Homer protein isoforms and contributing to synaptic scaling [9,10]. The emerging view is that the PSD is continuously remodeled and has a characteristic dynamics that is manifested not only by the addition and elimination of components over time but also by dynamic restructuring while retaining its composition [11–13].

From a structural point of view, the formation of interactions established through binding sites, displaying a diverse range of structural entities. Protein intrinsic disorder is defined by the lack of ability of a particular protein/segment to adopt a stable three-dimensional structure and is considered as a critical factor in mediating protein interactions. Disordered segments have been shown to play important roles in the postsynaptic scaffold (PSC) proteins [14], signaling pathways [15,16] and be at key positions within interaction networks [17]. The organization of synaptic proteins also exhibits a high amount of ordered domain interactions, resulting in macromolecular assemblies occupying the PSD. The highly organized assembly of PSD strongly suggests the presence of several intermolecular interacting sites with an elaborate and thoroughly regulated distribution of occupied and unavailable, or 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 include competitive binding events, allostery, and cooperativity tightly regulated through post-translational modifications (PTMs) [7,18], short linear motifs [19], and alternative splicing [20].

Most current descriptions of human postsynaptic proteome characterize PSD proteins using experimental procedures [21] or manual collection extended with gene annotation services [22]. To this date, there is no comprehensive computational analysis focusing on the structural organization of PSD proteins. In this paper we use a wide range of bioinformatic methods to describe the structural characteristics of PSD proteins, focusing on features that may directly contribute to the synaptic plasticity through the formation of Protein-Protein Interactions (PPIs).

Methods

Datasets

The human proteome was downloaded from UniProt [23] (2018_June release). For synaptic proteins, the SynaptomeDB [24] database and its classification of protein localization were used, defining four sets: postsynaptic, presynaptic, presynaptic active zone and vesicle-associated proteins. An additional 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 [25]. The list of nuclear proteins was taken from the supplementary material of Frege et al. [26], whereas the set of histone methylases were extracted from the accompanying data of Lazar et al. [27]. Lists of interacting protein pairs were taken from BioPlex 2.0 database [28]. Orthologs of postsynaptic scaffolding (PSC) proteins were downloaded from EggNOG database [29]. Only species with at least 90% of scaffold proteins were kept. Finally, only proteins present in all species were selected for further analysis.

Annotation and prediction of protein properties

Amino acids were classified into different groups, based on their physicochemical properties: hydrophobic (A, I, L, M, V), aromatic (F, W, Y), polar (N, Q, S, T), positively charged (H, K, R,), negatively charged (D, E), rigid (P), flexible (G), and covalently interacting (C). AAindex was used to obtain various amino acid scales [30]. To reduce the number of features, the ward.d2 function of R was used for K-means clustering [31], then based on the elbow method 5 clusters were selected to represent the features (Supplementary Figure 1). The following amino acid scales were selected randomly, each from one of the clusters: NAKH900104, KHAG800101, JUNJ780101, HUTJ700103, FAUJ880103 (for explanation see Supplementary Table 1). We later discuss that solvation energy is a major descriptor (EISD860101). Although the five randomly selected representative scales does not contain solvation free energy, it has high correlation to NAKH900104, thus it can be expected they share similar values on the proteome too.

Disorder prediction was made in a way that aims to minimize false positive hits arising from the inclusion of oligomeric fibrillar motifs [32]. First, the consensus of two disorder prediction methods, IUPred [33] and VSL2B [34] were determined. In the second step, all residues predicted to be in oligomeric fibrillar motifs were eliminated from the set of disordered residues [32]. Oligomeric fibrillar motifs were determined using a permissive prediction, namely, residues to form coiled coils as predicted either by COILS [35] or Paircoil2 [36], single alpha helices as identified using FT_CHARGE [37] or collagen triple helix as obtained by HMMER [38] with the Pfam HMM for collagen (ID: PF01391.13) segments in Pfam [39]. 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 [33]. Low complexity regions were predicted using SEG [40]. CCTOP [41] predicted TM regions. Signal peptides were predicted using SignalP [42]. Linear motifs participating in binding events were retrieved from the ELM database [43], considering only "LIG/DOC" classes. The ELM prediction

was also utilized, considering only those hits that passed the various filters as described by *Gouw et al.* [43]. Protein domains were taken from UniProt (2018_June release). Since domain annotation for orthologs is not available in many cases, PFAM and HMMER were used to define domains for these. PTMs were downloaded from PhosphoSitePlus [44] and were also predicted by NetPhos [45].
Venn diagrams were created using Venny [46].
All calculated values are available in Supplementary Table 1. Means and variances for different groups are available in Supplementary Table 2. Results for orthologous proteins are available in Supplementary Table 3.

Machine learning
For redundancy filtering CD-HIT [47] was used in an incremental manner, filtering identical proteins to 90, 70, 50 and finally to 40% identity. The remaining sequences were used to train the predictor. Labels were assigned based on the SynaptomeDB annotation.
For machine learning a feedforward neural network was developed, using one hidden layer with 40 neurons. Due to the non-proportional data for training and testing, bootstrap aggregating (BAGGING) was used. In each step, ten down-sampled sets were created and used to calibrate the Artificial Neural Networks (ANNs). For the final prediction, the results of individual ANNs were aggregated and weighted based on their reliability (defined as the output neuron probability). Benchmarking was done using ten-fold cross-validation and independent dataset too
Two different predictors were built: the first uses all annotations presented in Supplementary Table 1. The second predictor only uses features that can be derived from the amino acid sequence and do not depend on the annotation of different databases (shown with a grey background in Supplementary Table 1).

Results

Datasets
In this paper, we investigated three ever smaller subsets of proteins to the human proteome (21766 proteins): all proteins from the synaptome (1891 proteins), proteins localizing into the postsynaptic density (1761 proteins), and postsynaptic scaffold proteins (51 proteins). Three additional control dataset was defined: histone methylases (52 proteins), proteins from the nucleus (180 proteins) and the immunome (834 proteins).
According to our definition, there is considerable overlap between synaptome proteome and proteins of the PSD, and their characteristics are on par in every case. Therefore, conclusions drawn for PSD is valid for the synaptome too, even if it is not declared.

Sequences feature a mix of disorder and order promoting properties
General sequence properties often help to gain insight about structural features of proteins. As a very first step we compared the length distribution of proteins in the PSD to the human proteome: on average proteins in the PSD are longer (the average length of proteins is 583 and 705 residue in the proteome and in the PSD, respectively) (Figure 1, Supplementary Figure 2). We also calculated the average amino acid content of proteins. For this purpose, we classified

residues into different groups, based on their physicochemical properties (see Methods). Although values seem similar and their variances are high, proteins in the PSD have some preference to include more charged residues (~28% and ~25%, in the PSD and in the proteome, respectively) and to avoid cysteines (~2% and ~3%, in the PSD and in the proteome, respectively). These properties may promote disorder content; on some level, the lower proline content (5% and 6%, in the PSD and in the proteome, respectively) compensates these characteristics. Low complexity is also often related as an indicator of protein disorder: interestingly proteins in PSD are not enriched in such regions. Last, but not least, we analyzed selected amino acid property scales: the most distinct feature is related to the solvation energy, which is much lower in PSD proteins (Supplementary Table 2). In overall, these results hint a balanced presence of intrinsically disordered regions (IDRs) and ordered globular domains in the PSD.

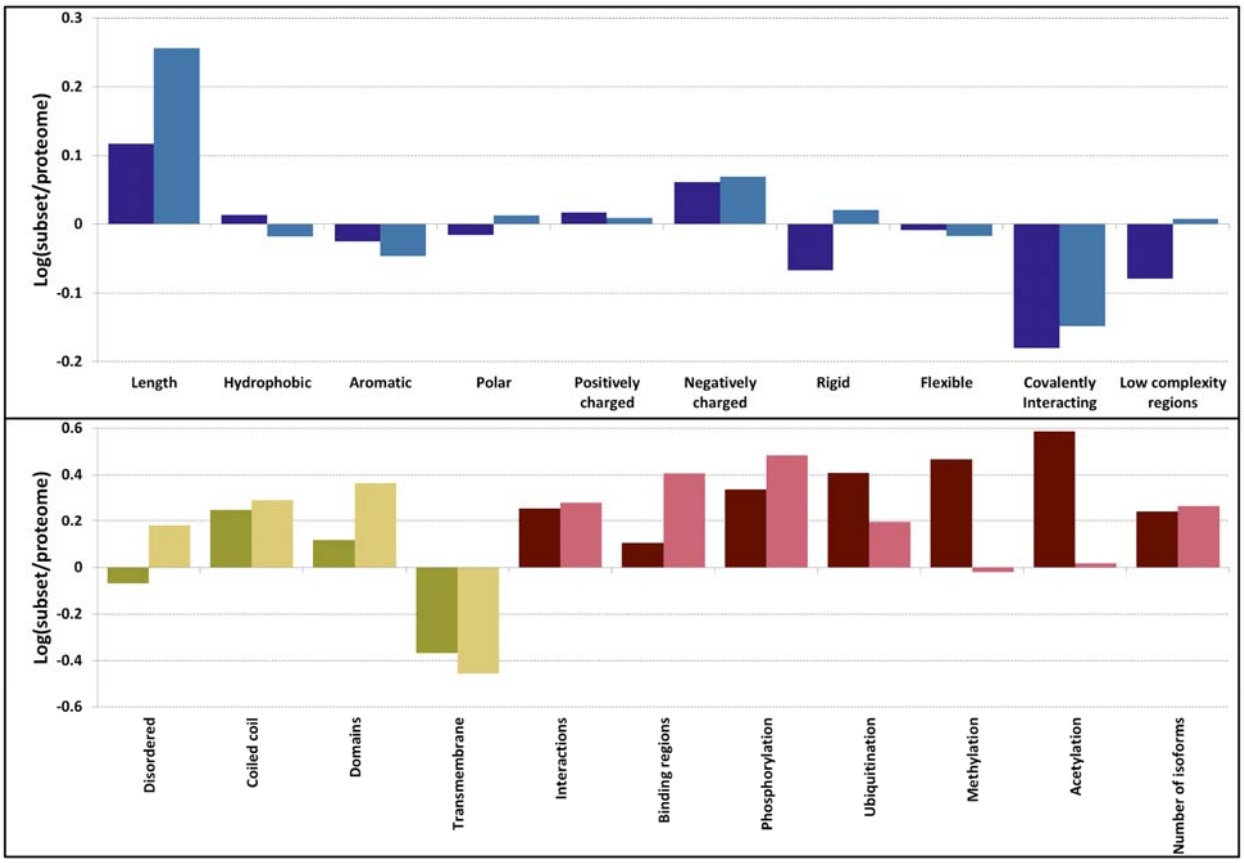


Figure 1: Change of different sequence (blue), structure (yellow) and PPI/regulation (red) related properties in PSD (darker shade) and PSC (lighter shade) compared to the human proteome. Amino acids were grouped as hydrophobic (A, I, L, M, V), aromatic (F, W, Y), polar (N, Q, S, T), positively charged (H, K, R), negatively charged (D, E), rigid (P), flexible (G), and covalently interacting (C).

PSC proteins are a subset of PSD proteins with somewhat distinct features. We observed their average length is even higher compared to proteins from PSD, and almost the double of proteins of the proteome (on average 970 residue length). They also contain more prolines (~13%) and low complexity regions (~29%) than PSD proteins or the proteome (Figure 1).

Other characteristics of PSC proteins are on par with PSD proteins. Thus it is expected they contain more IDRs.

PSD proteins exhibit a diverse range of structured elements

As sequence data hint differences compared to the proteome, various prediction methods were utilized to reveal different types of structural elements. We classified segments as disordered regions and ordered globular domains; then we extended this classification to groups often falling outside the classical "globular-disordered" partition: coiled-coils, a structural element already described in various PSD proteins and transmembrane segments assumed to play important role transmitting information outside the cell. On the one hand, PSD proteins operate with a similar amount of disorder content as proteins in the whole proteome (Figure 1). On the other hand, PSD proteins contain a large number of coiled coils and structured domains. We observed that not only the number of structured domains is higher, but flexible linker regions are also generally shorter in proteins of the PSD, allowing more dense placements of domains along the polypeptide chain (Supplementary Figure 3). Note that our stringent structure prediction pipeline (see Methods) has an emphasis on discriminating disordered and coiled-coil regions. Therefore the commonly occurring cross predictions between them are expected to only moderately bias these results. Another distinctive feature is the number of transmembrane proteins: proteins composed of seven transmembrane helices are greatly missing from the PSD (Supplementary Figure 4).

PSC proteins are on par with PSD proteins according to most of these features, however, on a more extreme way, utilizing even more coiled coils and domains. As expected from sequence characteristics, they also utilize a large amount of IDRs.

The overlap between protein-protein interactions and post-translational modifications hint a tightly regulated protein network in PSD

The synaptic plasticity of the PSD maintained through a high amount of PPIs (Figure 1, Supplementary Figure 5). Interestingly, proteins from the PSD are not particularly enriched in IDRs, a commonly used element to mediate PPIs. This is also reflected in the relatively low number of predicted disordered binding regions. The extremely high amount of phosphorylation and ubiquitination sites suggest that post-translational regulation may play an essential role in PSD proteins.

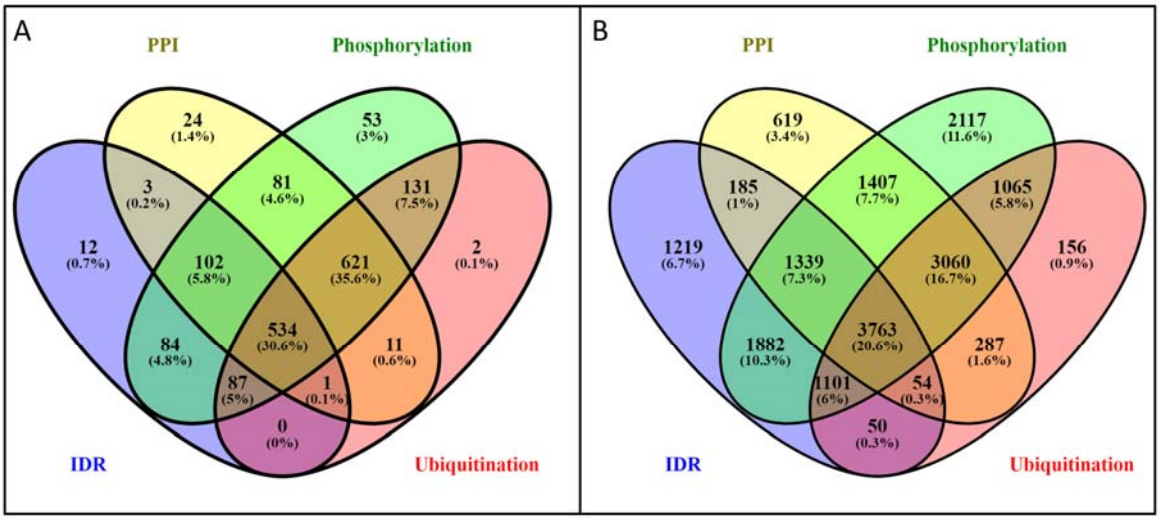


Figure 2: Venn diagram of proteins utilizing IDRs and PTMs to establish PPIs. A) PSD; B) human proteome.

As the increased number of IDRs and binding regions suggest, PSC proteins may use different mechanisms to promote PPIs. In the case of phosphorylation and ubiquitination, the trends are similar to PSD proteins. However, methylation and acetylation values are more close to the proteome used as a reference.

Since we do not have information about the specific binding regions and the interacting partners, we tried to characterize the nature of these interactions by observing their co-occurrence with proteins containing IDRs and PTMs. We assessed the mutual presence of PPIs, PTMs and protein disorder in the proteins of the PSD. Protein disorder for PPI formation may be a slightly neglected mechanism in the PSD compared to the human proteome (i.e., ~29% and ~42% of interacting proteins do not contain any IDR at all in the proteome and the PSD, respectively). However, PTMs may regulate more tightly the emerging PPIs (i.e., only ~90% compared to ~82% of proteins participating in interactions have phosphorylation or ubiquitination site in the PSD and the proteome, respectively) (Figure 2).

Alternative splicing can also contribute to regulation, by splicing binding regions and rewiring PPI networks. PSD and PSC proteins both seem to utilize this mechanism with an increased amount of isoforms compared to the human proteome (Figure 1).

Potential of the proteins to be engaged in multivalent interactions

Proteins in the PSD capable of establishing a high amount of interactions, which heavily depends on the number of basic building blocks: coiled coils can multimerize to build protein complexes, while IDRs, transmembrane proteins, and domains also can form inter- or intramolecular interactions. To assess the ability of a protein to contribute to interactions, we defined modularity as

$$Modularity = ln(DOM) + ln(CC10) + ln(IDR10) + ln\ )$$

where DOM is the number of domains, CC10 is the number of coiled-coil regions longer than ten residues, IDR10 is the number of IDRs longer than ten residues, and TMH is the number of transmembrane segments. Proteins in the PSD (and in the synaptome) are slightly more modular compared to the proteome, while PSC proteins utilize an extremely high amount of modules (Figure 3/a). However, we found that modularity alone did not describe well enough the potential to form an interaction: PSD and PSC proteins are on par regarding the number of interactions (Figure 1), which is not really reflected on modularity values (Figure 3). Therefore we defined Diversity of Protein Interaction (DPI) as a second descriptor and introduced additional elements in the equation:

$$DPI = ln(DOM) + ln(CC10) + ln(IDR10) + ln(TMH) + ln(Ph + Ub) + ln(ELM)$$

where Ph and Ub are the number of phosphorylations and ubiquitination sites along the protein sequence and ELM is the number of short linear motifs aiding binding processes (LIG/DOC

classes of the ELM database). Using DPI as measure PSD proteins are more on par with PSC proteins, as expected from the number of interactions (see Figure 1 and Figure 3/b).

To have a better picture how these descriptors behave, we also included three control sets with distinct features: immunome as a set of primarily globular multidomain proteins, histone methylases, known to have high IDR content and a more comprehensive list of proteins from the nucleus, shown to operate with a high amount of PPIs. We calculated the correlation between the descriptors and the average number of interactions in the selected protein sets. Modularity shows a slight increment from random (correlation: 0.23), while DPI fairly correctly estimates the potential to the formation of complexes (correlation: 0.63).

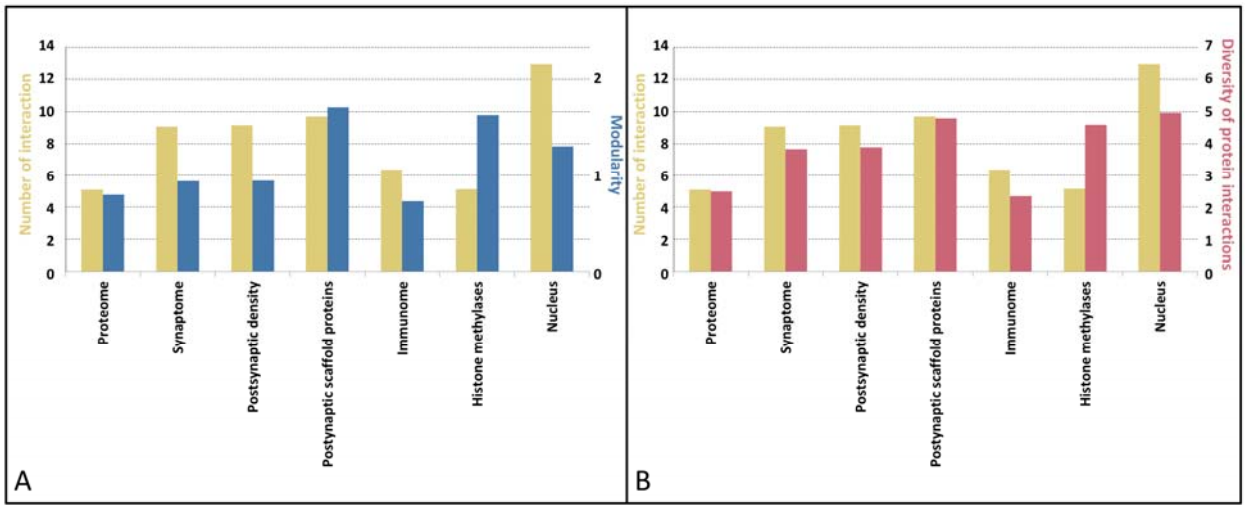


Figure 3: A) Average modularity and number protein interaction values in different protein sets. B) Average DPI and number protein interaction values in different protein sets.

Modularity serves as a base for the recruitment of proteins to the PSC during evolution

To further demonstrate the usability of such descriptors, we analyzed evolutionary relations between PSC proteins. The extreme high modularity of these proteins supporting the idea that domain organization may provide a filter to be included in synaptic transmission. To further test this hypothesis we collected 23 proteins localized in the PSC from 46 phylogenetically closely related orthologs, and calculated the average sequence length and modularity in them (Supplementary Figure 6). To better visualize how these values differ in different level organisms, we classified species into two groups, whether they are mammals or other vertebrates, and calculated how average sequence length and modularity vary within the groups. Both the sequence length and the modularity show a similar pattern: longer protein sequences and higher modularity are more dominant higher level organisms (Figure 4). Note, modularity seems to better emphasize the evolutionary relationship by displaying a monotonous increment/decline in contrast to the sequence length. For this analysis DPI calculations were discarded as PTM-s can only be collected through annotations, a largely missing tool in case of orthologs.

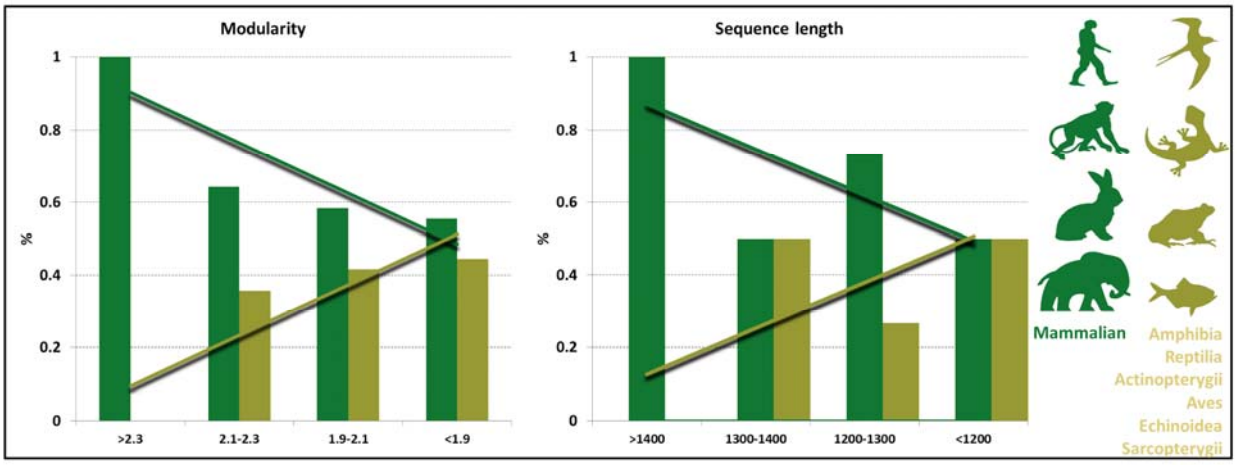


Figure 4: Proportion of classes in different sequence length (left) and modularity (right) intervals. Species were grouped by class: Mammalians (green) and other (yellow): Actinopterygii, Amphibia, Aves, Echinoidea, Reptilia, Sarcopterygii. Linear trends were added based on the values in bins.

Sequential and structural features discriminate PSD proteins from other proteins

Although some of the above-discussed trends look promising, to demonstrate the prediction power of these features we established an artificial neural network (ANN). The input of the ANN consists of the calculated features discussed above. The output is the binary classification defining the localization of the protein (i.e., localizing in the PSD/not). The distribution of positive and negative labels was not proportional; therefore we used bootstrap aggregation on the data. Briefly, this means that random samples of the data are used to train several predictors, and the final result is provided by aggregating the various outputs. The approach also helps to deal with the relatively noisy data, as in some cases it is expected that the same proteins localize inside and outside the PSD too. The algorithm achieved moderately high accuracy by incorporating all of the features presented in previous chapters (Area Under Curve (AUC): 0.84) (Figure 5 and Table 1).

Table 1: Prediction accuracy of the Neural Network. (MCC: Matthew Correlation Coefficient, BAC: Balanced Accuracy, AUC: Area Under Curve)

	Cross-validation			**Independent dataset**		
	MCC	**BAC**	**AUC**	**MCC**	**BAC**	**AUC**
All features	0.54+-0.09	0.77+-0.04	0.85+-0.03	0.52	0.76	0.84
Intrinsic features	0.32+-0.08	0.66+-0.04	0.75+-0.03	0.38	0.68	0.76

Some of the characteristics used as input are database dependent; therefore it is hard to assess how the ANN works on proteins with more/less annotation. To overcome this problem, a second predictor was built: the same bootstrapping approach was used. However, the feature space was reduced to contain only those characteristics that can be estimated based on the amino

acid sequence of the proteins (intrinsic features). In some cases, annotations were replaced by prediction (e.g., phosphorylation). Although the performance of the prediction decreased, these features alone still reliably discriminate PSD proteins from others (AUC: 0.76).
Besides AUC we also measured Matthew Correlation Coefficient and Balanced Accuracy. We observed similar values during the cross-validation and on the independent dataset too. The results suggest the selected structural features are descriptive for PSD proteins and the prediction method is quite robust regarding both features and sample sets.

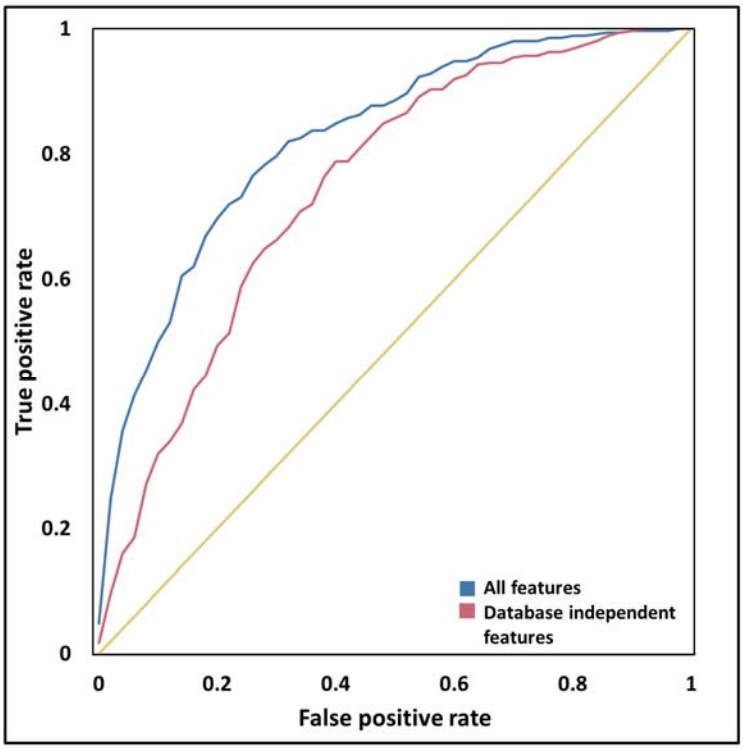


Figure 5: Receiver operating characteristics of predictors.

Discussion

Synaptic plasticity is facilitated by proteins of the PSD, most plausibly by using their functional repertoire provided by the ability to reorganize PPIs dynamically. Our results indicate that PSD and especially PSC proteins have an increased potential to form diverse interactions and this feature can even be used to discriminate PSD proteins from the rest of the proteome.

General sequence properties are often used to characterize particular protein subsets, as they open prospects to structural features. Grouping amino acids based on physicochemical properties highlights exciting trends. Proteins from PSD seem to be devoid of cysteins, an indicator of the abundance of well-ordered domains. On the one hand, most PSD proteins localize to the cytoplasm, where the reducing environment is not favorable to form disulfide bridges. On the other hand, this also means stabilization by clustering metal is a not much-utilized feature of PSD proteins (at least not with cysteins). The increased number of charged residues would also hint a higher extent of intrinsic disorder; however, the low amount of prolines seems to counterbalance this effect. Amino acid scales can also be descriptive: in this

case, the very low solvation free energy suggest protein collapse and folding is more dominant than protein disorder [48]. PSC proteins have similar characteristics, however with higher proline content that may promote intrinsic disorder.

In a recent study protein interactions were analyzed based on the structural state of participating partners, revealing major differences between “classical” protein disorder, and complexes formed by various partners. Three basic types of complexes were distinguished: autonomous folding and independent binding (i.e., the binding of two or more ordered proteins), coupled folding and binding (where an ordered protein stabilize an IDP partner) and mutual synergistic folding (interactions formed exclusively by disordered proteins) [49]. The overall amino acid content of PSC proteins shows high correlation with the group composed of proteins going through coupled binding and folding. In contrast, proteins from PSD are more close to the “mutual synergistic folding” (MSF) class (Supplementary Figure 7). We also noted that the predicted IDR content of PSD proteins lags behind that of the proteome (Figure 1). These observations raise the question whether PSD proteins utilize a different flavor of intrinsic disorder to establish interactions. A possible scenario might be that such flexible regions are overlooked by disorder prediction methods, as they lacked such protein sets for training. We note that coiled coils can be regarded as a subclass of MSF complexes, however our pipeline explicitly detects them and they do not show enrichment compared to PSC proteins (Figure 1). Thus, it is plausible that MSF mechanisms other than coiled-coil formation also contribute to complex formation in the PSD.

The structural organization of the PSD, utilizes a balanced distribution of different types of elements promoting PPIs. Coiled-coils, domain-domain interactions, intrinsically disordered regions, and transmembrane proteins all contribute to the formation of protein complexes in the PSD. Besides statistics presented in the results, detailed structure-function studies provide experimental evidence on how PPIs are formed and maintained in the PSD. Below we discuss different aspects of PSD organization by using specific well-characterized proteins as demonstrative examples.

The Shaker channel is a voltage-dependent potassium channel responsible for conducting depolarizing potassium currents when the membrane potential increases [50]. The C-terminal tail of the channel is in random coil state and contains a PDZ domain recognition motif. The motif assists the interaction with the PSD-95 scaffolding protein, helping the channel to cluster at unique membrane sites, which is important for the proper assembly and functioning of the synapse [51]. Anchor [52] can relatively accurately predict such interactions, also described as coupled binding and folding [53]. Both amino acid content (Supplementary Figure 7) and Anchor predictions indicate that PSC proteins are enriched in disordered binding regions relative to the proteome, while the number of unique partners lags behind this enrichment (Figure 1/b). The stacking of WW domain and MAPK binding motifs was described earlier in the case of TANC protein [54], also supporting the idea that PSC proteins “allow” their partners to pick from multiple unoccupied sites to fine-tune their function. In contrast, this enrichment is less pronounced in PSD and does not follow the enrichment in the number of interacting partners. These results theorize PSD proteins either more heavily utilize competitive binding with alternative partners binding to the same region, or promote different interaction modes. Since

our former recommendation is rather hard to asses using computational methods, we collected other possible interaction modes between proteins.

Activity-regulated cytoskeleton-associated protein (Arc) is a crucial regulator of long-term synaptic plasticity. The protein is highly modular and contains a flexible C-terminal tail [55]. Arc oligomerization is aided by the terminal IDRs, leading to the assembly of a structure reminiscent to an HIV capsid. This way Arc can encapsulate RNA and can mediate their transfer, which was shown to play a role in cell to cell communication in the nervous system [56]. Coiled-coils are also important oligomerization motifs, linking two or more partners with high specificity and a wide range of stability [57]. The PSD protein Homer can form high-order complexes with a diverse range of partners. A tetrameric form resulting in a coiled-coil develops in a two-step process: first, the C-terminal ~70 residues of the proteins form a parallel dimer, then two Homer dimers serve as a base for the antiparallel tetramer structure [58]. This arrangement is long enough to connect elements through the thickness of the PSD, serving direct connection between the plasma membrane and intracellular proteins through EVH1 domains binding to various scaffolding partners [59]. The mechanisms mentioned above share the characteristic feature of forming complexes by natively unfolded regions.

One can argue whether computational tools cross predict IDRs and coiled-coils, leading to a bias in our observations. To overcome this problem, our pipeline uses coiled coils as a filter when predicting IDRs. However, the lack of IDRs in PSD proteins could be explained by the fact, that we classified them as coiled coils. Calculating the overlap between IDR and coiled-coil containing proteins, and proteins forming interactions confirmed that this is not the case (Supplementary Figure 8), IDR and coiled-coil containing proteins do not dominate PPIs in PSD proteins.

PPIs can also occur between ordered segments. The amyloid beta precursor protein is a receptor helping neurite growth and neuronal adhesion. The structure of the E2 domain of the protein was solved multiple times, both in dimeric [60] and monomeric [61] forms, suggesting the domain is stable in both forms. It was proposed that dimerization aids the masking of several sites with biological activity [62], and thus the oligomerization state may regulate fibroblast growth by hiding a short sequence motif (RERMS), which was shown to promote growth processes [63].

A particular case of domain-domain interaction is the complex formation by transmembrane (TM) proteins. Membrane proteins often assemble into higher-order complexes to mediate transport or signaling. Nicotinic acetylcholine receptors are ion channels assisting fast chemical neurotransmission with high functional diversity, assembling into a pentameric form. Monomers consist of four TM helices, and the complex has an inner ring of helices, shaping the pore, while the outer three helices of each monomer shield the inner rings from lipids. In the closed state, the inner helices are in proximity and gate the channel. Upon acetylcholine binding, a conformational change occurs: the individual TM subunits remain rigid, however, the arrangement of the channel changes to become permeable for sodium ions [64].

An interesting, however somewhat logical aspect of the TM proteome of PSD is the underrepresentation of different GPCR proteins, otherwise constituting the most abundant protein family [65]. Over half of GPCRs encode olfactory receptors, responsible for perceiving a high variety of chemicals. Cell-cell communication relies on receptors, however, the highly

specialized ways of synaptic transmission may not require to distinguish such a diverse range of ligands as olfactory perception.

PPIs depend on structural elements providing binding sites, however, they are also intensively regulated through the spatiotemporal control in the cell, often using post-translational modifications to precisely modulate the properties of the protein.
PSD-95 is a scaffolding protein in the synaptome, and the Serine at 561 position was shown to be subject to phosphorylation and work as a molecular switch. Phosphomimetic (Ser -> Asp) mutation promotes an intramolecular interaction between the guanylate kinase (GK) and the SH3 domains, inducing a highly dynamic, yet closed conformation with buried binding sites. In contrast, mutating Ser561 to Alanine leads to a stable open conformation, facilitating the interaction between PSD-95 and its partners [66]. Besides such simple mechanism phosphorylation is often used cumulatively to produce an electrostatic effect [67] to fine-tune PPIs [68] or in a combinatorial way, where specific patterns can be responsible for regulating synaptic activity [69].
A more direct way to modulate PPIs is the control of the presence of its components at specific locations. Although it is hard to asses large scale protein synthesis and expression level data, ubiquitination sites hint the possibility of proteasome activity and degradation. It was shown that blocking either the synthesis or the degradation of synaptic proteins impairs the maintenance of late-phase long-term potentiation (L-LTP), however simultaneous blockade of protein degradation and protein translation leads to the rescue of L-LTP [70], suggesting an important role of ubiquitination in the PSD.
Additional regulation phenomena may provide further functional diversity. Using alternative splicing, different isoforms displaying characteristic sets and distributions of interaction sites may carry out distinct functions, as it was described in the case of PDZ-containing proteins [20]. Short Linear Motifs also play a role by mediating protein-protein interactions to contribute to a plethora of functions. Their short length and structural flexibility provide plasticity to emerge by convergent evolution and fine-tune interaction networks [71].

The multivalent nature of PSD proteins can also lead to additional structure-related phenomena: the SynGAP postsynaptic protein forms a coiled-coil trimer, shortly followed by a C-terminal binding motif interacting with the third PDZ domain of PSD-95. After complex formation, the assembly goes through liquid-liquid phase separation. The enrichment of SynGap and PSD-95 was shown to highly correlate with synaptic activity [72], advocating the critical role of the formation of membraneless organelles in synaptic processes [73].

We introduced descriptors to assess the presence and distribution of different elements in PSD proteins and their possible role in interactions. Since PPI formation may occur through many distinct structural elements, we included several of these to thoroughly catch many possible aspects of protein complex formation. Modularity is a structural descriptor, showing moderate correlation with the number of interaction partners. By extending modularity with post-translational regulation members, the potential of a protein to form distinct interactions can be adequately defined. Verification with different control sets confirmed that it can be used directly on proteins sets to estimate their tendency to establish elaborate protein networks.

Although modularity alone cannot determine the possibility of PPIs, it is robust enough to demonstrate evolutionary relationships, even between closely related species. Proteins in different species show some level of conservation. Even simpler molecules have conserved segments recognizable as part of modern structures, showing the evolutionary path of protein development [74]. Such conserved segments carry functional importance [75] and can be considered as ancestral modules of proteins [76]. The basic signaling mechanisms found in synapses are highly conserved across the different kingdoms of life and serve as a base of proteins selected in the PSD [77]. Although these observations connect distant evolutionary stages, our results show similar results in very closely related species. Organisms with more complex nervous system exhibit PSD proteins with higher modularity and longer sequences. These results are also in line with other studies focusing on the correlation between conservation and protein length [78] and the conservation of functional units in the PSD [79].

Our protein-level descriptions are all focused on PPIs, and thus it is not trivial whether they are a discriminative feature of PSD proteins are only relevant in the context of specific protein complexes. Machine learning is a commonly used tool to discriminate protein subsets when the number of features and their variance is too high to overlook. Our results demonstrate the prediction power of the presented sequential, structural and regulation features. Considering the noise present in the datasets (as some proteins may be localized in the PSD and other locations too, moreover the possible false classification of source databases also adds a bias), the prediction is remarkably accurate. Incorporating structure based machine learning algorithm may enhance synaptome database development by expediting data collection and reduce manual effort.

In conclusion, we suggest that postsynaptic proteins and in particular, postsynaptic scaffold proteins are capable of forming diverse kinds of interactions with their partners that we propose to play a key role in the functional organization of the postsynaptic density and its dynamic rearrangements upon stimuli.

Acknowledgments

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. The constructive remarks of Bálint Mészáros and the technical assistance of Zsófia Kálmán are gratefully acknowledged.

References

1. Rudy, J.. *The Neurobiology of Learning and Memory*; Sinauer Associates, Sunderland, MA, USA, 2008;
2. Ho, V.M.; Lee, J.-A.; Martin, K.C. The Cell Biology of Synaptic Plasticity. *Science (80-.).* **2011**, *334*, 623–628.
3. Bailey, C.H.; Kandel, E.R. Chapter 10 Synaptic remodeling, synaptic growth and the storage of long-term memory in Aplysia. In *Progress in brain research*; 2008; Vol. 169, pp. 179–198.
4. Bourne, J.N.; Harris, K.M. Coordination of size and number of excitatory and inhibitory synapses results in a balanced structural plasticity along mature hippocampal CA1 dendrites during LTP. *Hippocampus* **2011**, *21*, 354–373.
5. Feng, W.; Zhang, M. Organization and dynamics of PDZ-domain-related supramodules in the postsynaptic density. *Nat. Rev. Neurosci.* **2009**, *10*, 87–99.
6. Dosemeci, A.; Tao-Cheng, J.H.; Vinade, L.; Winters, C.A.; Pozzo-Miller, L.; Reese, T.S. Glutamate-induced transient modification of the postsynaptic density. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 10428–32.
7. Ehlers, M.D. Activity level controls postsynaptic composition and signaling via the ubiquitin-proteasome system. *Nat. Neurosci.* **2003**, *6*, 231–242.
8. Farley, M.M.; Swulius, M.T.; Waxham, M.N. Electron tomographic structure and protein composition of isolated rat cerebellar, hippocampal and cortical postsynaptic densities. *Neuroscience* **2015**, *304*, 286–301.
9. Diering, G.H.; Nirujogi, R.S.; Roth, R.H.; Worley, P.F.; Pandey, A.; Huganir, R.L. Homer1a drives homeostatic scaling-down of excitatory synapses during sleep. *Science (80-.).* **2017**, *355*, 511–515.
10. de Vivo, L.; Bellesi, M.; Marshall, W.; Bushong, E.A.; Ellisman, M.H.; Tononi, G.; Cirelli, C. Ultrastructural evidence for synaptic scaling across the wake/sleep cycle. *Science* **2017**, *355*, 507–510.
11. Blanpied, T.A.; Kerr, J.M.; Ehlers, M.D. Structural plasticity with preserved topology in the postsynaptic protein network. *Proc. Natl. Acad. Sci.* **2008**, *105*, 12587–12592.
12. MacGillavry, H.D.; Song, Y.; Raghavachari, S.; Blanpied, T.A. Nanoscale Scaffolding Domains within the Postsynaptic Density Concentrate Synaptic AMPA Receptors. *Neuron* **2013**, *78*, 615–622.
13. Meyer, D.; Bonhoeffer, T.; Scheuss, V. Balance and Stability of Synaptic Structures during Synaptic Plasticity. *Neuron* **2014**, *82*, 430–443.
14. Tantos, A.; Han, K.-H.; Tompa, P. Intrinsic disorder in cell signaling and gene transcription. *Mol. Cell. Endocrinol.* **2012**, *348*, 457–465.
15. 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 J.* **2005**, *272*, 5129–5148.
16. Vucetic, S.; Xie, H.; Iakoucheva, L.M.; Oldfield, C.J.; Dunker, A.K.; Obradovic, Z.; Uversky, V.N. Functional Anthology of Intrinsic Disorder. 2. Cellular Components, Domains, Technical Terms, Developmental Processes, and Coding Sequence Diversities Correlated with Long Disordered Regions. *J. Proteome Res.* **2007**, *6*, 1899–1916.
17. Cortese, M.S.; Uversky, V.N.; Dunker, A.K. Intrinsic disorder in scaffold proteins: getting more from less. *Prog. Biophys. Mol. Biol.* **2008**, *98*, 85–106.
18. Schuman, E.M.; Dynes, J.L.; Steward, O. Synaptic Regulation of Translation of Dendritic mRNAs. *J. Neurosci.* **2006**, *26*, 7143–7146.
19. Songyang, Z.; Fanning, A.S.; Fu, C.; Xu, J.; Marfatia, S.M.; Chishti, A.H.; Crompton, A.;

Chan, A.C.; Anderson, J.M.; Cantley, L.C. Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science* **1997**, *275*, 73–7.
20. Sierralta, J.; Mendoza, C. PDZ-containing proteins: alternative splicing as a source of functional diversity. *Brain Res. Rev.* **2004**, *47*, 105–115.
21. Bayés, À.; Collins, M.O.; Croning, M.D.R.; van de Lagemaat, L.N.; Choudhary, J.S.; Grant, S.G.N. Comparative Study of Human and Mouse Postsynaptic Proteomes Finds High Compositional Conservation and Abundance Differences for Key Synaptic Proteins. *PLoS One* **2012**, *7*, e46683.
22. Pirooznia, M.; Wang, T.; Avramopoulos, D.; Valle, D.; Thomas, G.; Huganir, R.L.; Goes, F.S.; Potash, J.B.; Zandi, P.P. SynaptomeDB: an ontology-based knowledgebase for synaptic genes. *Bioinformatics* **2012**, *28*, 897–9.
23. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **2018**, *46*, 2699–2699.
24. Pirooznia, M.; Wang, T.; Avramopoulos, D.; Valle, D.; Thomas, G.; Huganir, R.L.; Goes, F.S.; Potash, J.B.; Zandi, P.P. SynaptomeDB: an ontology-based knowledgebase for synaptic genes. *Bioinformatics* **2012**, *28*, 897–9.
25. Ortutay, C.; Vihinen, M. Immunome knowledge base (IKB): an integrated service for immunome research. *BMC Immunol.* **2009**, *10*, 3.
26. Frege, T.; Uversky, V.N. Intrinsically disordered proteins in the nucleus of human cells. *Biochem. Biophys. Reports* **2015**, *1*, 33–51.
27. Lazar, T.; Schad, E.; Szabo, B.; Horvath, T.; Meszaros, A.; Tompa, P.; Tantos, A. Intrinsic protein disorder in histone lysine methylation. *Biol. Direct* **2016**, *11*, 30.
28. Huttlin, E.L.; Bruckner, R.J.; Paulo, J.A.; Cannon, J.R.; Ting, L.; Baltier, K.; Colby, G.; Gebreab, F.; Gygi, M.P.; Parzen, H.; et al. Architecture of the human interactome defines protein communities and disease networks. *Nature* **2017**, *545*, 505–509.
29. Jensen, L.J.; Julien, P.; Kuhn, M.; von Mering, C.; Muller, J.; Doerks, T.; Bork, P. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* **2007**, *36*, D250–D254.
30. Kawashima, S.; Kanehisa, M. AAindex: amino acid index database. *Nucleic Acids Res.* **2000**, *28*, 374.
31. Ward, J.H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.
32. Gáspári, Z. Is Five Percent Too Small? Analysis of the Overlaps between Disorder, Coiled Coil and Collagen Predictions in Complete Proteomes. *Proteomes* **2014**, *2*, 72–83.
33. Mészáros, B.; Erdős, G.; Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **2018**, *46*, W329–W337.
34. Obradovic, Z.; Peng, K.; Vucetic, S.; Radivojac, P.; Dunker, A.K. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins Struct. Funct. Bioinforma.* **2005**, *61*, 176–182.
35. Lupas, A.; Van Dyke, M.; Stock, J. Predicting coiled coils from protein sequences. *Science (80-.).* **1991**, *252*, 1162–1164.
36. Berger, B.; Wilson, D.B.; Wolf, E.; Tonchev, T.; Milla, M.; Kim, P.S. Predicting coiled coils by use of pairwise residue correlations. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92*, 8259–63.
37. 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*, 109–116.
38. Finn, R.D.; Clements, J.; Eddy, S.R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **2011**.
39. Finn, R.D.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; et al. The Pfam protein families database:

towards a more sustainable future. *Nucleic Acids Res.* **2016**, *44*, D279–D285.
40. Wootton, J.C. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.* **1994**, *18*, 269–85.
41. Dobson, L.; Reményi, I.; Tusnády, G.E. CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res.* **2015**.
42. Almagro Armenteros, J.J.; Tsirigos, K.D.; Sønderby, C.K.; Petersen, T.N.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **2019**, *37*, 420–423.
43. Gouw, M.; Sámano-Sánchez, H.; Van Roey, K.; Diella, F.; Gibson, T.J.; Dinkel, H. Exploring Short Linear Motifs Using the ELM Database and Tools. In *Current Protocols in Bioinformatics*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2017; Vol. 58, p. 8.22.1-8.22.35.
44. Hornbeck, P. V.; Zhang, B.; Murray, B.; Kornhauser, J.M.; Latham, V.; Skrzypek, E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **2015**, *43*, D512–D520.
45. Blom, N.; Sicheritz-Pontén, T.; Gupta, R.; Gammeltoft, S.; Brunak, S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **2004**, *4*, 1633–1649.
46. Oliveros, J.C. VENNY. An interactive tool for comparing lists with Venn Diagrams.
47. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152.
48. Eisenberg, D.; McLachlan, A.D. Solvation energy in protein folding and binding. *Nature* **1986**, *319*, 199–203.
49. Mészáros, B.; Dobson, L.; Fichó, E.; Tusnády, G.E.; Dosztányi, Z.; Simon, I. How folding and binding intertwine during protein complex formation provides an additional layer of functional regulation. *J. Mol. Biol.* **2019**, *submitted*.
50. Hoshi, T.; Zagotta, W.N.; Aldrich, R.W. Biophysical and molecular mechanisms of Shaker potassium channel inactivation. *Science* **1990**, *250*, 533–8.
51. Magidovich, E.; Orr, I.; Fass, D.; Abdu, U.; Yifrach, O. Intrinsic disorder in the C-terminal domain of the Shaker voltage-activated K+ channel modulates its interaction with scaffold proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 13022–7.
52. Dosztányi, Z.; Mészáros, B.; Simon, I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **2009**, *25*, 2745–2746.
53. Sugase, K.; Dyson, H.J.; Wright, P.E. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* **2007**, *447*, 1021–1025.
54. Gasparini, A.; Tosatto, S.C.E.; Murgia, A.; Leonardi, E. Dynamic scaffolds for neuronal signaling: in silico analysis of the TANC protein family. *Sci. Rep.* **2017**, *7*, 6829.
55. Myrum, C.; Baumann, A.; Bustad, H.J.; Flydal, M.I.; Mariaule, V.; Alvira, S.; Cuéllar, J.; Haavik, J.; Soulé, J.; Valpuesta, J.M.; et al. Arc is a flexible modular protein capable of reversible self-oligomerization. *Biochem. J.* **2015**, *468*, 145–158.
56. Pastuzyn, E.D.; Day, C.E.; Kearns, R.B.; Kyrke-Smith, M.; Taibi, A. V; McCormick, J.; Yoder, N.; Belnap, D.M.; Erlendsson, S.; Morado, D.R.; et al. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer. *Cell* **2018**, *172*, 275–288.e18.
57. Gáspári, Z.; Nyitray, L. Coiled coils as possible models of protein structure evolution. *Biomol. Concepts* **2011**, *2*, 199–210.
58. Hayashi, M.K.; Ames, H.M.; Hayashi, Y. Tetrameric Hub Structure of Postsynaptic Scaffolding Protein Homer. *J. Neurosci.* **2006**, *26*, 8492–8501.
59. Hayashi, M.K.; Tang, C.; Verpelli, C.; Narayanan, R.; Stearns, M.H.; Xu, R.-M.; Li, H.; Sala, C.; Hayashi, Y. The postsynaptic density proteins Homer and Shank form a polymeric network structure. *Cell* **2009**, *137*, 159–71.

60. Ninomiya, H.; Roch, J.M.; Sundsmo, M.P.; Otero, D.A.; Saitoh, T. Amino acid sequence RERMS represents the active domain of amyloid beta/A4 protein precursor that promotes fibroblast growth. *J. Cell Biol.* **1993**, *121*, 879–86.
61. Hoopes, J.T.; Liu, X.; Xu, X.; Demeler, B.; Folta-Stogniew, E.; Li, C.; Ha, Y. Structural Characterization of the E2 Domain of APL-1, a Caenorhabditis elegans Homolog of Human Amyloid Precursor Protein, and Its Heparin Binding Site. *J. Biol. Chem.* **2010**, *285*, 2165–2173.
62. Lee, S.; Xue, Y.; Hu, J.; Wang, Y.; Liu, X.; Demeler, B.; Ha, Y. The E2 Domains of APP and APLP1 Share a Conserved Mode of Dimerization. *Biochemistry* **2011**, *50*, 5453–5464.
63. Ninomiya, H.; Roch, J.M.; Sundsmo, M.P.; Otero, D.A.; Saitoh, T. Amino acid sequence RERMS represents the active domain of amyloid beta/A4 protein precursor that promotes fibroblast growth. *J. Cell Biol.* **1993**, *121*, 879–86.
64. Morales-Perez, C.L.; Noviello, C.M.; Hibbs, R.E. X-ray structure of the human α4β2 nicotinic receptor. *Nature* **2016**, *538*, 411–415.
65. Dobson, L.; Reményi, I.; Tusnády, G.E. The human transmembrane proteome. *Biol. Direct* **2015**, *10*, 31.
66. Wu, Q.; Sun, M.; Bernard, L.P.; Zhang, H. Postsynaptic density 95 (PSD-95) serine 561 phosphorylation regulates a conformational switch and bidirectional dendritic spine structural plasticity. *J. Biol. Chem.* **2017**, *292*, 16150–16160.
67. Serber, Z.; Ferrell, J.E. Tuning Bulk Electrostatics to Regulate Protein Function. *Cell* **2007**, *128*, 441–444.
68. Sun, Q.; Jackson, R.A.; Ng, C.; Guy, G.R.; Sivaraman, J. Additional Serine/Threonine Phosphorylation Reduces Binding Affinity but Preserves Interface Topography of Substrate Proteins to the c-Cbl TKB Domain. *PLoS One* **2010**, *5*, e12819.
69. Coba, M.P. Regulatory mechanisms in postsynaptic phosphorylation networks. *Curr. Opin. Struct. Biol.* **2019**, *54*, 86–94.
70. Fonseca, R.; Vabulas, R.M.; Hartl, F.U.; Bonhoeffer, T.; Nägerl, U.V. A Balance of Protein Synthesis and Proteasome-Dependent Degradation Determines the Maintenance of LTP. *Neuron* **2006**, *52*, 239–245.
71. Davey, N.E.; Cyert, M.S.; Moses, A.M. Short linear motifs - ex nihilo evolution of protein regulation. *Cell Commun. Signal.* **2015**, *13*, 43.
72. Zeng, M.; Shang, Y.; Araki, Y.; Guo, T.; Huganir, R.L.; Zhang, M. Phase Transition in Postsynaptic Densities Underlies Formation of Synaptic Complexes and Synaptic Plasticity. *Cell* **2016**, *166*, 1163–1175.e12.
73. 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.
74. Trifonov, E.N.; Frenkel, Z.M. Evolution of protein modularity. *Curr. Opin. Struct. Biol.* **2009**, *19*, 335–340.
75. Lipman, D.J.; Souvorov, A.; Koonin, E. V; Panchenko, A.R.; Tatusova, T.A. The relationship of protein conservation and sequence length. *BMC Evol. Biol.* **2002**, *2*, 20.
76. Trifonov, E.N.; Frenkel, Z.M. Evolution of protein modularity. *Curr. Opin. Struct. Biol.* **2009**, *19*, 335–340.
77. Emes, R.D.; Grant, S.G.N. The human postsynaptic density shares conserved elements with proteomes of unicellular eukaryotes and prokaryotes. *Front. Neurosci.* **2011**, *5*, 44.
78. Zhang, J. Protein-length distributions for the three domains of life. *Trends Genet.* **2000**, *16*, 107–9.
79. Emes, R.D.; Grant, S.G.N. The human postsynaptic density shares conserved elements with proteomes of unicellular eukaryotes and prokaryotes. *Front. Neurosci.* **2011**, *5*, 44.