Pázmány Péter Catholic University

Roska Tamás Doctoral School of Sciences and Technology

Bioinformatics analysis of interactions involved in the formation of postsynaptic protein complexes

Theses of the PhD dissertation

Zsófia Etelka Dobson-Kálmán

Supervisor: Zoltán Gáspári, PhD

2022

1 Introduction

1.1 The postsynapse

The brain is our most complex organ, where neuronal cells connecting to each other are responsible for a plethora of functions. Synapses are functional units, and their elaborate connection provides a presynaptic and postsynaptic side, where neurons on the postsynaptic side receive the information that are transmitted from the presynaptic side. The postsynaptic side forms a sophisticated network of thousands of proteins, where the most components do not function independently, but interacts with each other. Many interactions form a higher order organization and produce the so-called supercomplexes. Despite the growing number of identified supercomplexes up to 220 were identified so far [1], most of our data is still stuck at the binary level of protein-protein interactions. Within the postsynaptic neuron lies the postsynaptic density, a structure connected to the membrane bilayer, that is also visible by electron-microscopy. The PSD consists of around 1 000 distinct proteins [2], with the mass of 1gDa, which means most of the proteins are present in multiple copies [3]. The concentration of different components highly varies, with one of the most abundant components being PSD-95, presenting ∼300 copies in each synapse.

1.2 The coiled-coil motif

Coiled-coils are one of the earliest discovered supersecondary structural elements, where two or more alpha helix forms a spiral in a way to maximize the number of hydrophobic contacts between chains [4]. They are frequent motifs in different organisms, and they usually provide 3-10% of different proteomes [5, 6]. Coiled-coils consist of special repeating seven residue length patterns: HPPHPPP, where H are hydrophobic residues, P are polar ones [4]. Different positions within the heptad are marked ('abcdefg') [7, 8], position 'a' and 'd' are usually hydrophobic, 'e' and 'g' are electrostatic, the rest of the amino acids are solvent accessible. There is a scientific agreement that a folding initialization segment, a so-called trigger sequence, is present in many coiled-coils, forming a unit capable of autonomous folding to reduce the conformations space and aid the folding of the full length coiled-coil [5]. Most coiled-coils consist of two or three helices, but higher order organizations are also possible [9], providing different oligomerization states. The situation is more complicated, as the alpha-helices orientation relative to each other can be parallel or antiparallel [7], subunits can be identical (homooligomer) or different (heterooligomer) [6], and sometimes helices are from the same protein (monomer). Coiled-coils have highly diverse functions [10], they are present in motor proteins, receptors and transcription factors.

1.3 Protein-protein interactions

Protein-protein interactions are the direct physical associations of at least two proteins via molecular docking, occurring *in vivo* in living cells [11]. These interactions are evolved to serve a special function, therefore they are highly specific [11]. In a living cell, many interactions competing with each other, and some of them are transient and have a lower affinity. In principle, the most dominant effect driving the interaction is still not the affinity, but the spatiotemporal regulation of cells, resulting in a different abundance of various components at one localization. Interactions are mediated by domains and flexible segments. Most of our knowledge is limited to domain-domain interactions, where hundreds of thousands binding event and more than 10 000 complexes were already discovered [12, 13]. In contrast to domain-domain interactions, flexible (or disordered) segments have the functional advantage to use their partner domain as a structural template, therefore these segments are capable of interacting with multiple partners [14]. Within disordered segments short linear motifs are a special subclass, responsible for mediating transient interactions. Although it is estimated that there are around a million linear motif-mediated interactions present in eukaryotic cells, only a handful were experimentally verified [12]. Protein-protein interactions have different affinity and half life, where as the ones mediated by short linear motifs are primarily very low affinity and transient, often in micromolar and millisecond range [15]. There are domain-domain interactions on the other side of the spectra, where the binding event lasts for hours and the affinity range can be nano- or picomolar. There are dozens of techniques (and their combinations) to detect protein-protein interactions, many of them providing different (or sometimes complementary) information about the binding.

1.4 Disease-causing germline mutations

Replication is one of the most fundamental processes in living cells, occurring in a highly precise manner, where the mutations rate (i.e. frequency of errors) is around 10^{-10} with every cell division [16]. Errors occurring in germline cells are passed forward between generations, that means that each cell in the descendant will carry the error too [17, 18]. Each mutation has an effect, and the phenotype can be placed on a spectrum, where some of them will be beneficial and aid the positive selection, some of the are neutral (e.g. color of the eye) and some of them will have a negative effect and will result in the loss in the quality of life [17]. Negative effects cause diseases, such as lactose-intolerance and Huntington's disease, however, in the case of germline mutations these phenotypes are usually weaker - in contrast to somatic mutations, many consequences are incompatible with life.

2 Methods

During my work I used freely accessible softwares and data. To process the data I relied on scripts and programs written in Python3, mostly under the Linux operating system. Here I list the most important databases and programs that I used during my doctoral period:

Databases:

Uniprot: `https://www.uniprot.org/`

PDB: `https://www.rcsb.org/`

Pfam: `https://pfam.xfam.org/`

SynaptomeDB: `http://metamoodics.org/SynaptomeDB/index.php`

G2C: `https://www.genes2cognition.org/`

SynGO: `https://www.syngoportal.org/`

BioGRID: `https://thebiogrid.org/`

IntAct: `https://www.ebi.ac.uk/intact/home`

STRING: `https://string-db.org/`

Methods:

BLAST: `https://blast.ncbi.nlm.nih.gov/Blast.cgi`

CD-HIT: `https://www.bioinformatics.org/cd-hit/`

ClustalO: `https://www.ebi.ac.uk/Tools/msa/clustalo/`

GO: `http://geneontology.org/`

DeepCoil: `https://toolkit.tuebingen.mpg.de/tools/deepcoil`

Ncoils: `https://predictprotein.org/`

Paircoil: `http://cb.csail.mit.edu/cb/paircoil/`

Marcoil: `https://toolkit.tuebingen.mpg.de/tools/marcoil`

Logicoil: `http://coiledcoils.chm.bris.ac.uk/LOGICOIL/`

3 New scientific results

3.1 First thesis group: investigation of the coiled-coil structural motif

3.1.1 I proved that disease-causing germline mutations accumulate in postsynaptic proteins, compared to neutral polymorphisms, in contrast to the human proteome used as a background. Frequency of disease mutations in other structural compartments (transmembrane and disordered regions) are similar in the proteome and in postsynaptic proteins.

I determined the distribution of mutations in postsynaptic proteins ('PS_STRICT dataset') and in the human proteome, considering different structural and functional regions. I investigated transmembrane (CCTOP), coiled-coil (Deepcoil, Ncoils, Marcoil és Paircoil), disordered (IUPred) and domain (Pfam) segments (respectively, in case of contradiction), and post-translational modifications regulating the proteins (PhosphoSitePlus). I found that in disordered regions polymorphisms (PM) accumulate compared to disease-causing mutations (DM) (Figure 1). In contrast to this observation, DMs are more frequent in other structural regions. These observations are true for both the PS and the human proteome. Notably, trends are the same, but effects are more markedly present in the PS in each case. The only exception we found to be different is present in coiled-coil regions: in the PS DMs have a higher frequency in these segments compared to PMs (Figure 1). This results is somewhat surprising: I) this is the only case when the PS and the human proteome shows different characteristics II) we expected coiled-coils in the human proteome have higher DM frequency, similar to other structural parts. Disordered regions tolerate more DMs, as they are more flexible and they have more conformational freedom, in other words there is no structure that could be impaired. To have a deeper understanding of how coiled-coil structures are impared via DMs we investigated the role of mutations in coiled-coil regions.

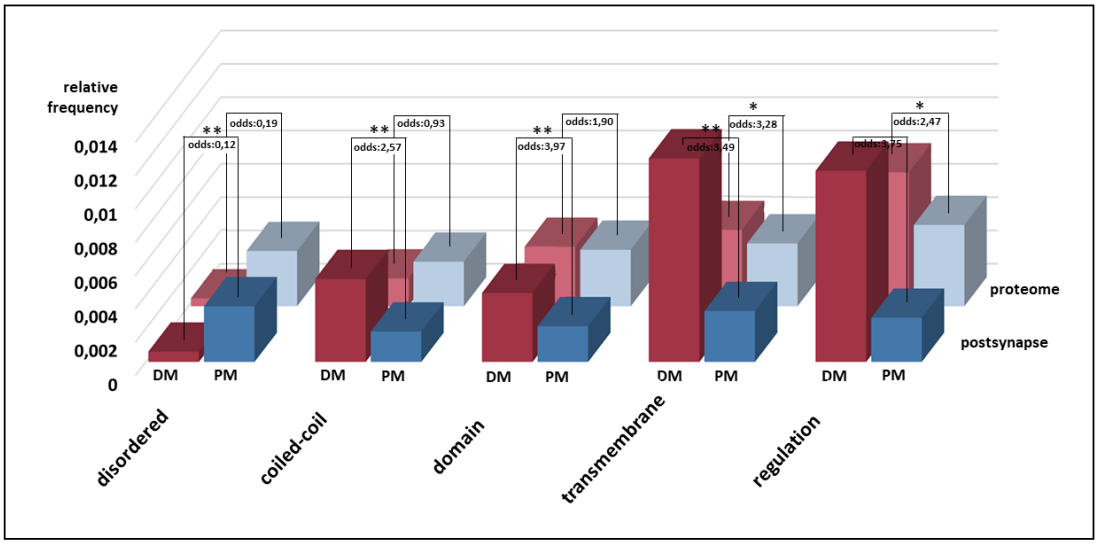


Figure 1: Relative frequency of disease-causing mutations (DM) and polymorphisms (PM) in postsynaptic proteins (PS) and in the human proteome, considering different structural regions (x-axis). DMs are red, PMs are blue. Darker shade indicates PS proteins, lighter shade is for the human proteome.

3.1.2 I showed that the relative frequency of DMs are significantly higher in the N-terminal region of coiled-coils, indicating the important role of these segments.

Despite the lower number of DMs in coiled-coil regions, thousands of harmful mutations still falling into these segments. First, we investigated the distribution of mutations along the sequence, dividing it into 5 segments with equal lengths (1-20%, 21-40%, 41-60% etc.) (Figure 2.A). This approach did not show a clear difference between regions. Next, we investigated the distribution of mutations based on the heptad repeat (1-7, 8-14, 15-21, 22-28) (Figure 2.B), where it became clear that DMs mostly fall into the first heptad. Comparing the N-terminal 7 residue to other parts of the coiled-coil the accumulation of DMs is significant according to χ^2 teszt ($p < 0.01$), with a log odds ration of 1,33 between DMs and PMs (Figure 2.C). This effect is strong enough to influence the number of mutations falling into coiled-coils with various lengths: DMs are more frequent in shorter coiled-coils, while PMs show a uniform distribution (Figure 2.D). We hypothesize that N-terminal regions have an essential role in the folding of CC-s, possibly by the presence of the trigger sequence.

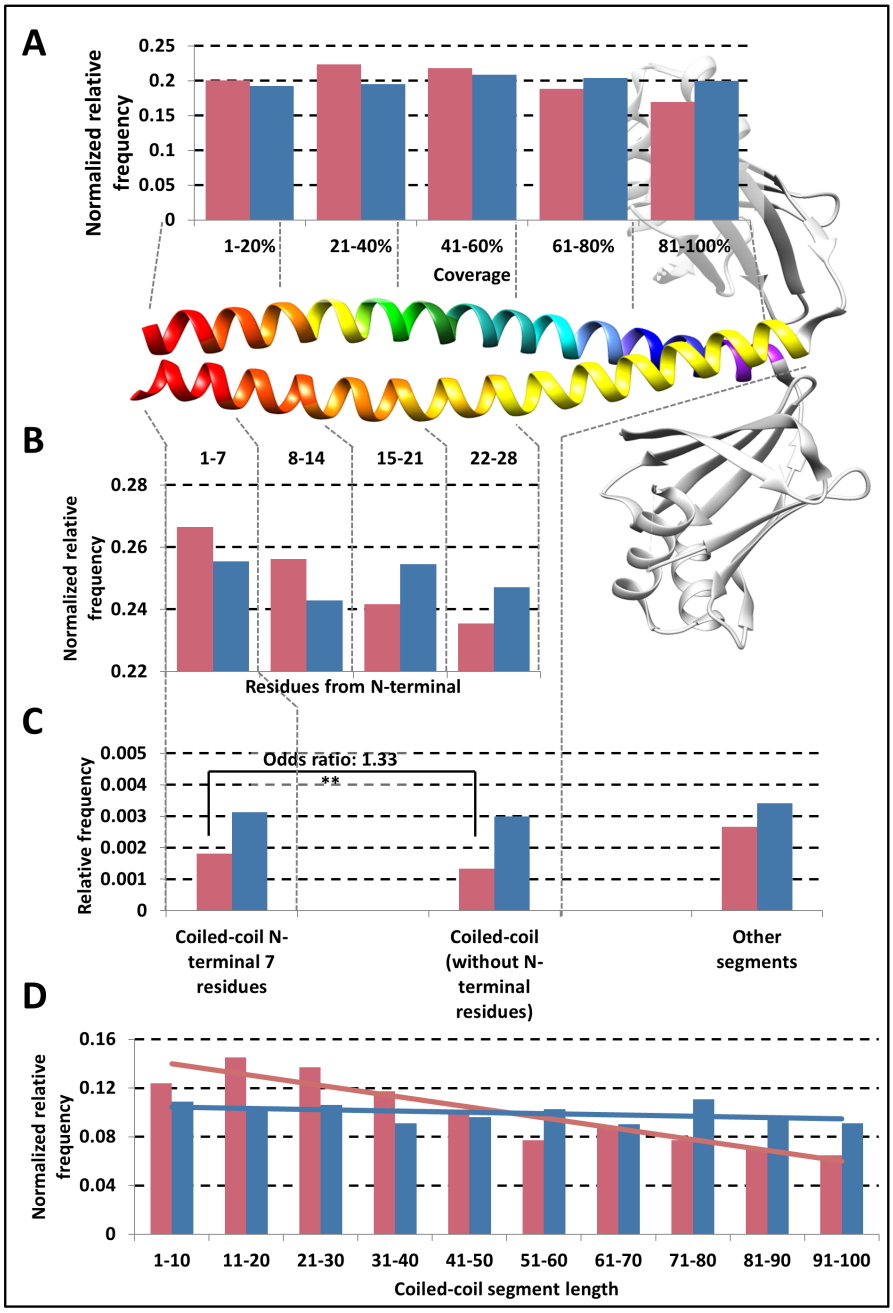


Figure 2: Distribution of DMs and PMs in coiled-coils. A) Relative frequency of mutations when slicing the coiled-coils into five parts with equal lengths. B) Relative frequency of mutations considering the heptad repeat C) Relative frequency of mutations in the first 7 residue of coiled-coils, coiled-coil without the N-terminal residues, and in other segments. D) Distribution of mutations based on the length of the coiled-coil segments. DMs: red, PMs: blue

3.1.3 Based on DiseaseOntology I showed that DMs in coiled-coils can be mostly associated with nervous system diseases.

Proteins having coiled-coil regions have a diverse range of functions. Diseases associated with DMs falling into coiled-coils can highlight what critical functions are maintained with coiled-coils segments. DiseaseOntology (DO) is similar to GeneOntology, where disease are structured hierarchically. By mapping DO, the CC_SEQ and the humsavar datasets, we can identify the disease group of each DM. This analysis not only investigates individual proteins, but also enables us to look at DMs in a more systematically and highlight disease groups that are more frequently associated with DMs.

Many diseases can be found in different groups at one time; for example, Alzheimer-disease can be grouped as a genetic disease and an anatomic entity. With this investigation, I showed that mutations in coiled coils are primarily associated with central nervous system diseases - notably, muscle and skin diseases are also frequent consequences (see Figure 3 below).

DO level1	Numb. of prot.	DO level2	Numb. of prot.	DO level3	Numb. of prot.
Disease of anatomical entity	734	Integumentary system disease	72	Skin disease	71
		Muskosceletal system disease	165	Muscle disease	161
		Nervous system disease	397	Central nervous system disease	294
				Sensory system disease	74
		Cardiovascular disease	60	Heart disease	60
Syndrome	52				
Genetic disease	56				
Disease of metabolism	103	Inherited metabolic disease	63	carbohydrate metabolic disorder	63

Figure 3: Disease groups considering DMs in coiled-coil regions and the number of cases involved

3.2 Second thesis group: investigation of postsynaptic protein-protein interactions

3.2.1 We established the PostSynapticInteractionDataBase, where my curation work was extended using other resources: protein-protein interaction collection of other curator groups, structural and functional features, disease categories.

Our fundamental goal was to establish a database that is specific the protein-protein interactions (PPIs) occurring in the postsynapse. We aimed to develop a database, where this information is extended with the binding regions, structural information that can help to understand how these binding processes are mediated and functional information to gain insight into the regulation of interactions. In addition, we aimed to enrich the database with other disease related information that may be relevant to understanding the function of the network. I manually curated more than 2 000 interaction data using a predefined curation system that was extended with 4 other interaction data resources: IntAct, BIOGRID, PDB and STRING databases. We included the following structural features: transmembrane topology (HTP database), coiled-coil regions (DeepCoil prediction and UniProt annotation), intrinsically disordered regions and disordered binding segments (IUPred2A), phase separation (PhasePro), domains (Pfam). We also included functional and disease related information: short linear motifs (Eukaryotic Linear Motif database), Phosphorylation (UniProt annotation), GeneOntology annotation, Disease-causing germline mutations (Humsavar) and disease groups (DiseaseOntology). The database is available at `https://psindb.itk.ppke.hu` (Figure 4).

3.2.2 Based on standards of the curation community I developed a curation system that can be used to describe postsynaptic protein-protein interactions. Using this system I annotated more than 2 000 protein-protein interactions and revealed unknown protein-protein interaction subnetworks.

We constructed the PSINDB database with other scientists, that have experience with the IntAct curation system (the gold standard curation system of the scientific field). We concluded that to establish PSINDB, the most appropriate level of curation is an extended MIMIx format - this enables relatively fast, yet still exhaustive curation work. In our case, this meant that we extended MIMIx format with the binding regions of proteins. We followed HUPO-PSI guidelines and used controlled vocabularies (Ontology Lookup Service) to define the features and parameters of the interactions. We also used PSI-MI format to publish our results to help the data integration effort of other curator groups. PSINDB can be used for various analyses and investigations. Here I show how my curation work revealed local subnetworks: SHSA6 and SHSA7 are bitopic transmembrane receptors located in the synapse membrane and are known to be involved in regulating transmission in CA3-CA1 synapses, possibly via regulating AMPA-type glutamate receptors. The existing interactions mapped for SHSA6

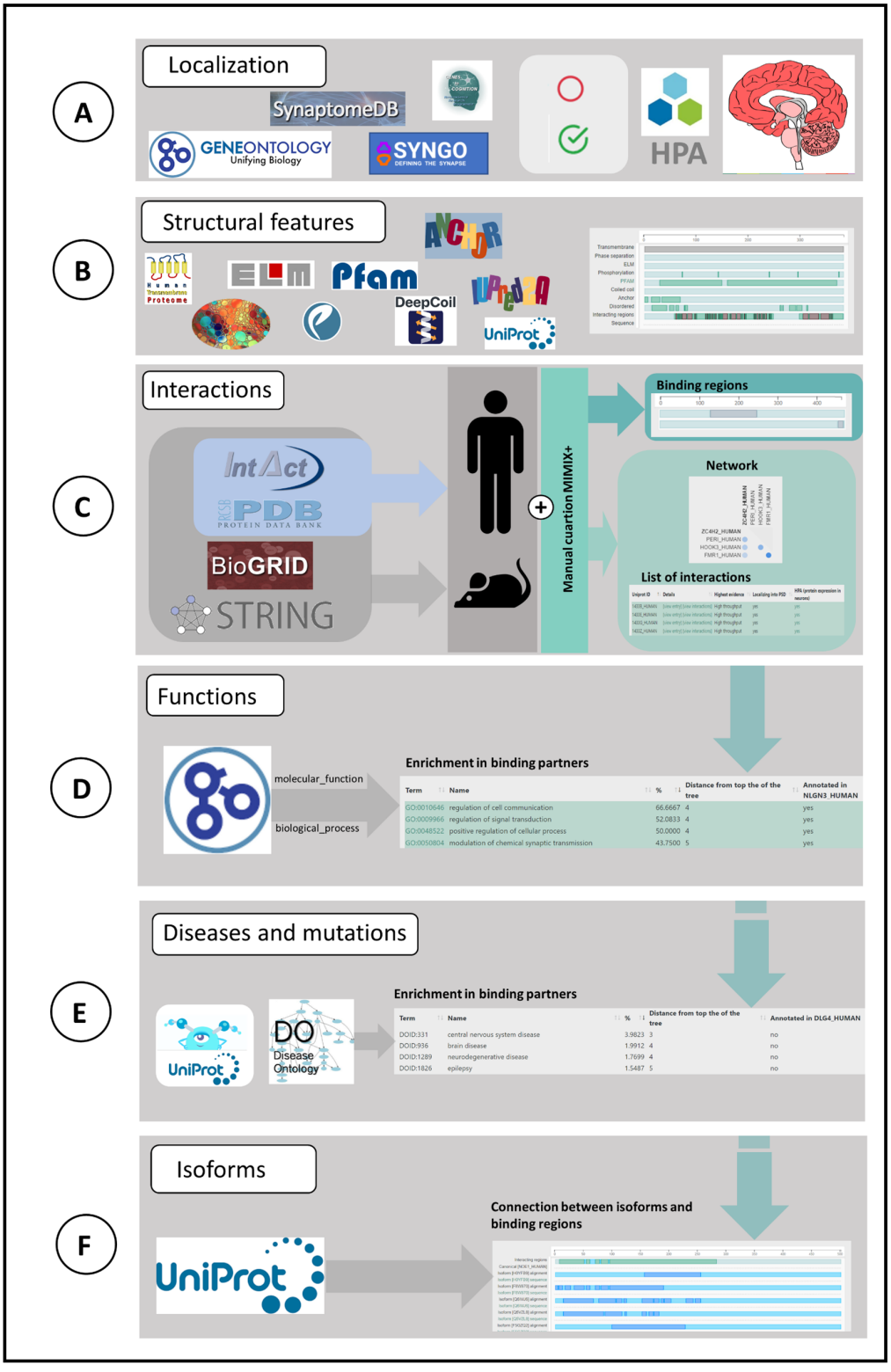


Figure 4: Data layers in PSINDB A: postsynaptic localization evidence (SynaptomeDB, G2C, SynGO, GO) and expression data (HumanProteinAtlas), B: Structural and functional features of proteins. C: Interactions, D: Functions, E: Disease related information, F: Protein isoforms

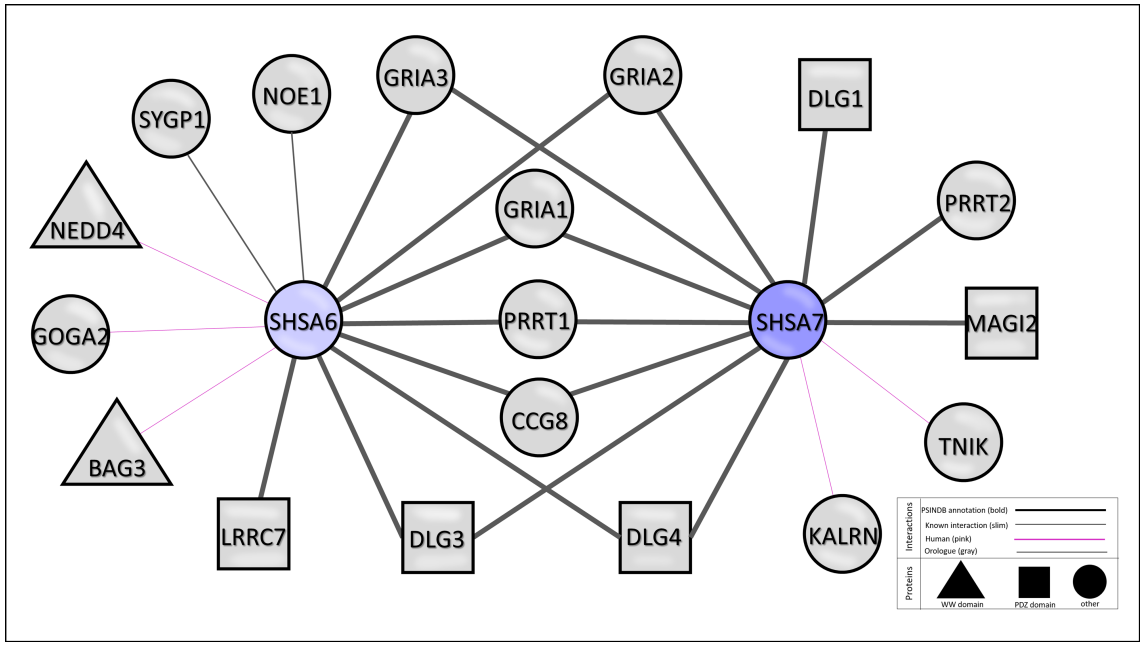
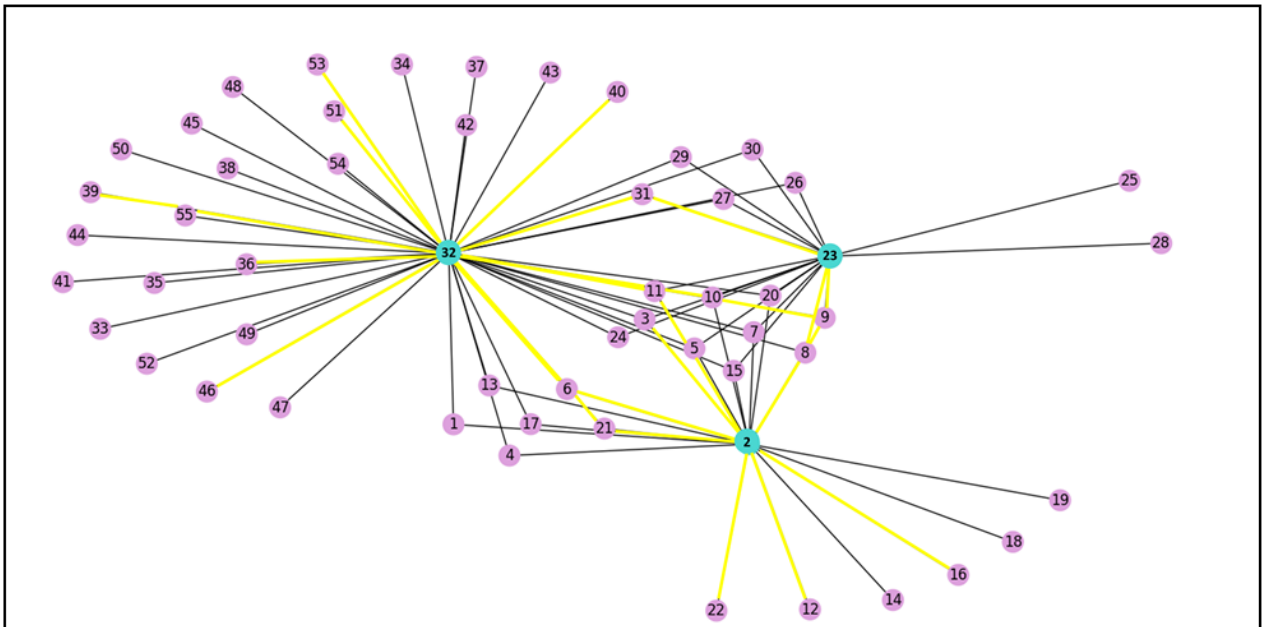


Figure 5: Interaction network of SHSA6 and SHSA7 proteins extended by PSINDB Interactions are represented by lines; notation: thin: already known, bold: new annotation, pink: human taxon, grey: orthologue Protein annotation: square: contains PDZ domain, triangle: contains WW domain, circle: contains other domain

and SHSA7 were low in numbers, limiting the assessment of the full functional repertoire of these receptors based on easily accessible interaction data. Our manual curation efforts significantly extended the available interaction data for both receptors, adding 8 and 11 interactions for SHSA6 and SHSA7 on top of the already known 6 and 2 interactions. Figure 5 shows the current interaction network of these two receptors with newly annotated interactions in bold. PSINDB allows for the assessment of interactions along different criteria, such as the reliability of the data encoded in the weight of the edges. This example also highlights how mapping interactions through close homologs can enrich interaction networks. Several high-confidence interactions in this sub-network were derived from mouse studies, which can be reliably mapped to their human counterparts as the mouse and human SHSA proteins share a very high degree of sequence identity.

3.2.3 Using the information available in the PSINDB I suggeted new binding regions from domain-domain and domain-motif interactions.

Although there are more than 100 000 interactions the PSINDB, interaction regions was only defined in a small fraction of experiments. These regions are only available in our curation system, or in case the data is derived from the PDB or IntAct databases. Furthermore, in the case of interactions of orthologous proteins (i.e. non-human) binding regions are not mirrored. However, to better understand the synapse we need to have a more profound knowledge of what regions drive the binding. Domain-domain interaction is a highly studied field, while information about motif mediated interactions are somewhat limited. Complex assembly and competitive binding are also fundamental underlying processes in the forming of networks that can be only investigated once we have information about the exact binding regions. I made suggestions for binding regions using two different approaches:
I) using PSINDB, PDB and Pfam databases, I extended the possible domain-domain interaction data of PSINDB.
II) using the PSINDB, the ELM and the Pfam databases, I created a ranked list of candidate domain-motif interactions that implicate the binding regions as well.
Using the Dlg family (Dlg1, Dlg3, Dlg4), domain mediated interactions suggest new binding regions for 11% of Dlg1 partners, 8% of Dlg3 partners and 6% of Dlg4 partners. For a quick comparison, 18%, 2% and 7% of binding regions were known before the analysis, respectively. Dlg family has SH3 and PDZ domains, that are famous for binding short linear motifs: I detected 152 possible linear motif in their partners, from which 50 cases the binding regions was unknown before, and in 18 cases the data is contradicting (Figure 6).



1	CTND2_HUMAN	11	CCG2_HUMAN	21	DUS10_HUMAN	31	NLGN3_HUMAN	41	RIMS1_HUMAN	51	SHAN1_HUMAN				
2	***DLG1_HUMAN***	12	GPC5B_HUMAN	22	EZRI_HUMAN	*32*	***DLG4_HUMAN***	42	DYN1_HUMAN	52	NRX2A_HUMAN				
3	DLGP4_HUMAN	13	IQEC1_HUMAN	*23*	***DLG3_HUMAN***	33	CDKL5_HUMAN	43	FBX2_HUMAN	53	SHAN2_HUMAN				
4	KCNJ2_HUMAN	14	LCK_HUMAN	24	CCG4_HUMAN	34	ABL2_HUMAN	44	HSP7C_HUMAN	54	HNRPK_HUMAN				
5	TANC2_HUMAN	15	SHSA9_HUMAN	25	NEDD4_HUMAN	35	AT1A1_HUMAN	45	SI1L1_HUMAN	55	ABI1_HUMAN				
6	EFNB3_HUMAN	16	EPB41_HUMAN	26	NETO1_HUMAN	36	HGS_HUMAN	46	NLGNX_HUMAN						
7	SHSA7_HUMAN	17	ASAP1_HUMAN	27	SHSA6_HUMAN	37	CADH4_HUMAN	47	SOS1_HUMAN						
8	DLGP1_HUMAN	18	MYO6_HUMAN	28	DC1I1_HUMAN	38	SCN2A_HUMAN	48	ABLM1_HUMAN						
9	GRIK2_HUMAN	19	HCN1_HUMAN	29	ZDHC5_HUMAN	39	KCND2_HUMAN	49	PP1B_HUMAN						
10	CNKR2_HUMAN	20	DLG2_HUMAN	30	KCNA1_HUMAN	40	SYNPO_HUMAN	50	KPCE_HUMAN						

Figure 6: Domain-motif interactions of Dlg1, Dlg3 and Dlg4. Top panel: revealed domain-motif network (blue: protein with domain, pink: protein with motif, yellow: suggested binding regions for existing connections. Bottom panel: partner protein names: light gray: Dlg1 and Dlg4 partners, blue: Dlg3 and Dlg4 partners, dark gray: Dlg1, Dlg3 and Dlg4 partners)

4 Summary

Postsynapses are an important unit of neuronal function whose molecular mechanisms are elements of basic brain functions such as memory and learning. Although our knowledge has increased greatly over the past decades, much of the information needed to understand this organelle is still lacking. Postsynaptic proteins are the basis for the functioning of the postsynapse, and they carry out their biological function by creating networks of almost unimaginable complexity.

My dissertation, in which I have used *in silico* methods to investigate the postsynaptic system from several aspects, with a focus on the binary interactions of proteins, the functional elements playing a part in connections and their involvement by mutations. This work will hopefully contribute to the future understanding of this complex network. I first investigated the effects of germline mutations on the structural elements of postsynaptic proteins. An interesting finding - the prevalence of disease-causing mutations in proteins - and its low literature coverage prompted us to investigate the effects of germline mutations in coiled-coil structures in more detail. However, I have extended this analysis to the entire human proteome. Interestingly, the involvement of central nervous system was detected aswell. In the second half of my dissertation, I present the construction of a postsynaptic interaction database (PSINDB), which contained individual and biological data relevant to the interactions. Our database can also be considered a primary database, as it contains ∼2000 interaction data that are not included in definitive interaction databases. Manual curation was performed using best practices in the field. In addition, we used observations from the analysis of PSINDB data to predict domain-domain and domain-motif interactions. Both of these interactions may be of paramount importance in the function of the postsynapse, and some of these are certainly still unexplored.

Understanding the function of the postsynapse would be important not only to understand the molecular mechanisms of our basic brain functions, but also to understand the more than one hundred neuronal diseases that have been associated with the proteins of this organelle.

5 Publication list

Included in disseration

Kalman, Zs. E., Mészáros, B., Gáspári, Z., Dobson, L. (2020). Distribution of disease-causing germline mutations in coiled-coils implies an important role of their N-terminal region. Scientific reports, 10(1), 1-12.

Kalman, Zs. E., Dudola, D., Mészáros, B., Gáspári, Z., Dobson, L. (2022). PSINDB: the postsynaptic protein–protein interaction database. Database, 2022, baac007.

Other publications

Quaglia, F., Mészáros, B., Salladini, E., Hatos, A., Pancsa, R., Chemes, L. B., ... Piovesan, D. (2022). DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. Nucleic Acids Research, 50(D1), D480-D487.

Frank, K., Bana, N. Á., Bleier, N., Sugar, L., Nagy, J., Wilhelm, J., ... Steger, V. (2020). Mining the red deer genome (CerEla1. 0) to develop X-and Y-chromosome-linked STR markers. PLoS One, 15(11), e0242506.

Not peer-reviewed journal

Kalman, Zs. E., Gáspári, Z. (2021). A preliminary study on the cistrome of human postsynaptic density from an evolutionary and network-based perspective. bioRxiv.

Bibliography

[1] Seth GN Grant. Synapse molecular complexity and the plasticity behaviour problem. *Brain and Neuroscience Advances*, 2:2398212818810685, 2018.

[2] Takeshi Kaizuka and Toru Takumi. Postsynaptic density proteins and their involvement in neurodevelopmental disorders. *The Journal of Biochemistry*, 163(6):447–455, 2018.

[3] Morgan Sheng and Eunjoon Kim. The postsynaptic organization of synapses. *Cold Spring Harbor perspectives in biology*, 3(12):a005678, 2011.

[4] Derek N Woolfson. The design of coiled-coil structures and assemblies. *Advances in protein chemistry*, 70:79–112, 2005.

[5] Jody M Mason and Katja M Arndt. Coiled coil domains: stability, specificity, and biological implications. *ChemBioChem*, 5(2):170–176, 2004.

[6] Linda Truebestein and Thomas A Leonard. Coiled-coils: The long and short of it. *Bioessays*, 38(9):903–916, 2016.

[7] Andrei N Lupas, Jens Bassler, and Stanislaw Dunin-Horkawicz. The structure and topology of α-helical coiled coils. *Fibrous Proteins: Structures and Mechanisms*, pages 95–129, 2017.

[8] Gevorg Grigoryan and Amy E Keating. Structural specificity in coiled-coil interactions. *Current opinion in structural biology*, 18(4):477–483, 2008.

[9] Oliver D Testa, Efrosini Moutevelis, and Derek N Woolfson. Cc+: a relational database of coiled-coil structures. *Nucleic acids research*, 37(suppl_1):D315–D322, 2009.

[10] Andrei N Lupas and Markus Gruber. The structure of α-helical coiled coils. *Advances in protein chemistry*, 70:37–38, 2005.

[11] Zeeshan Shaukat, Sara Aiman, Chun-Hua Li, et al. Protein-protein interactions: Methods, databases, and applications in virus-host study. *World Journal of Virology*, 10(6):288, 2021.

[12] Peter Tompa, Norman E Davey, Toby J Gibson, and M Madan Babu. A million peptide motifs for the molecular biologist. *Molecular cell*, 55(2):161–169, 2014.

[13] Erica A Golemis, Erica Golemis, and Peter David Adams. *Protein-protein interactions: a molecular cloning manual*. CSHL Press, 2005.

[14] Alexander Cumberworth, Guillaume Lamour, M Madan Babu, and Jörg Gsponer. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochemical Journal*, 454(3):361–369, 2013.

[15] Xuan Yang and Andrey A Ivanov. Computational structural modeling to discover ppi modulators. In *Protein–Protein Interaction Regulators*, pages 87–108. 2020.

[16] Bradley D Preston, Tina M Albertson, and Alan J Herr. Dna replication fidelity and cancer. In *Seminars in cancer biology*, volume 20, pages 281–293. Elsevier, 2010.

[17] Catarina D Campbell and Evan E Eichler. Properties and rates of germline mutations in humans. *Trends in Genetics*, 29(10):575–584, 2013.

[18] Yanmei Dou, Heather D Gold, Lovelace J Luquette, and Peter J Park. Detecting somatic mutations in normal cells. *Trends in Genetics*, 34(7):545–557, 2018.