Bioinformatics analysis of two protein structural elements involved in neural signaling

Thesis excerpt

Pázmány Péter Catholic University
Faculty of Information Technology and Bionics

Written by
Dániel Dudola

Supervisor
dr. Zoltán Gáspári

Budapest, 2020

Chapter 1

Introduction

1.1 The single α-helices

The charged single α-helix isa rare structural motif, one of its main features is that although single α-helices are generally not stable on their own, in the presence of specific amino acid sequences (e.g., four glutamic acids (E) and four positively charged amino acids, alternating repeats of arginine (R) and lysine (K)) are stable in solution phase without the stabilizing effect of the tertiary-quaternary structure. This ER / K motif (which is found in many proteins of many organisms) is responsible for a lot of different functions [12]. The detection of single α-helices (SAH) is complicated by the fact that the charged regions searched by the algorithms are also found in coiled coils, although the proportion of charged amino acids in the SAH domains is higher. For this reason, some of the coiled-coils detected by the algorithms are presumably SAH domains, but further research is required to unambiguously identify them. [9] The formation of the SAH domain is likely, if both negatively and positively charged amino acids are present in large numbers, but lack the hydrophobic seam (the overlapping sites of the α-helices that form the coiled-coil) and there are many E-K, R-K interactions within the helix.

Although SAH domains are much more prevalent than thought at the time of their discovery, the exact role of the domains remains unclear, still, but their presence may have a highly probable physical significance in several cases. One of these functions is the extension of the arm of the myosin molecule, which increases the distance the myosin molecule travels on the actin fiber during one power stroke. Measurements have shown that the SAH domains found in some varieties of myosin are rigid enough to perform this arm-extending function. [1]

1.2 The PDZ domains

PDZ domains are involved in protein-protein recognition and interaction, contributing to – among other things – the formation of protein complexes essential for neuronal signaling pathways. The sequence that consists of approyimately 90 amino acids long is very common, it can be found in photoreceptors of the fruit fly (Drosophila) and mammalian synapses and is one of the most common domains in sequenced genomes [6]. The name itself is an acronym consisting of the names of the three proteins in which the PDZ domain was first detected, these are the postsynaptic PSD-95, the Discs in the tight junction of the fruit fly and the ZO-1 proteins.

Its role in protein-protein interactions can be traced back to the structural properties of the protein. The binding site of the peptide ligands is a binding pocket between the β2 fold and the α2 helix, in which the βfold of the bound ligand protein is joining to the βfold of the binding pocket [5]. During binding, the C-terminus of the ligand peptide is wedged into the small binding pocket, and due to the electrostatic properties of the pocket it favors binding sequences which end in hydrophobic amino acids (valine, isoleucine, leucine). In addition to ligand peptide C-terminals, similar internal motifs may be also be involved in binding to the PDZ domain.

Because the PDZ domain binds short, flexible C-termini and similar internal motifs, it is able to form bonds with virtually every proteins, especially membrane proteins – such as ion channels – which typically have short C-termini. Binding is also "gentle" with respect to the ligand, the structure of the bound ligand does not change compared to the free state [6]. In addition, the strength of the bond is not very high, due to its transient nature, the equilibrium rearranges quickly and easily.

In the bound state, the β-folds stabilize the position of the ligand, in which the side chain involved in the ligand binding process points directly towards the binding pocket of the PDZ domain. Thus, the binding specificity of the domain can be traced back to the position of the ligands β fold and its polarized side chain.

In addition to the binding affinity for short C-termini, the PDZ domain is also able to recognize internal protein motifs, if they are structurally similar to the chain terminus. This type of bonding typically occurs only when the C-terminus of the ligand peptide is difficult to access due to the structure of the ligand, and its spatial position makes the binding difficult or impossible. Due to the binding specificity of the domain, such binding with an internal motif of the ligand is believed to occur less frequently because the geometry and polarization of the binding site of the domain are highly specific.

1.3 Analysis of dynamic structural protein ensembles

In the structural and functional research of proteins, the dynamic properties of the structures are increasingly important as the functional role of these internal motions became well recognized. In the past, the experimental methods used for structure determination did not allow detailed observation of structural movements, from which it was erroneously concluded that the proteins have a well-defined, rigid structure. One of the first experimental methods used for structure determination is X-ray crystallography, which – as it appears in the name of the method – is a single-crystal diffraction of a protein. In the crystallized - rather unrealistic - state, the examined proteins are inherently incapable of motion, so there is no first hand information on the dynamic properties of the structures.

Nuclear magnetic resonance spectroscopy (NMR) is the first experimental method to obtain information on the dynamic properties of structures at the atomic level. The dynamic role of protein structures also appeared in the modeling of proteins in the form of dynamic protein structure sets, since a model cannot describe the conformational space samppled by the structure. In this conformational space a dynamic equilibrium is observed between the conformational states, the shift of which can also affect the biological properties of the protein.

The ensemble representation of protein structures is also supported by the fact that for some systems, models based on one conformer do not represent well the actual structure of a given protein. A good example of this is the group of intrinsically disordered proteins (IDPs), whose high flexibility relative to globular proteins soon shed light on the fact that such structures can only be characterized in an ensemble-based manner. The highly dynamic structure of these protein structures helped the discovery of many previously unknown biochemical mechanisms, although these are often difficult to characterize on the atomic level.

The approach considered most effective today is a combination of experiments and calculations, where the parameters derived from the measurements are used as constraints in molecular dynamics calculations using the appropriate spatial interpretation. This approach is very similar to traditional structure determination, the difference lies in the use of experimental results. In the generation of such dynamic ensembles, the interpretation of parameters is actually population-based, i.e., instead of expecting individual conformers to correspond to all experimental data, compliance is interpreted and expected on the ensemble.

Chapter 2

Methods

In the SAH detection algorithms SCAN4CSAH and FT_CHARGE, only the charged amino acids which play a role in SAH formation were taken into consideration. To reduce false positive hits, I implemented a helicity filter that removes the structures from the hits which are not expected to be helical.

To find an optimal parametrisation the filter, I used the DSSP annotation [2] of the sequences in the non-redundant PDB SELECT database [4]. For the alpha helices with the minimum lenght of 15 amino acids I calculated the average Chou-Fasman Pα value per amino acid. I fitted an EVD (extreme value distribution) to these values using an R program [10]. To determine the probability threshold obtained on this basis, I examined the SAH sequences predicted in the 2015 edition of the SwissProt database, and based on my findings, the P value was chosen to be 0.5, i.e. only sequences with a lower P value were considered.

To generate an extended PDZ set, I ran molecular dynamics simulations without constraining parameters with the PDZ sets found in the PDB database (2KPL, 2LC6, 2LC7, 2LOB, 1GM1, 1OZI, 2KPK, 1VJ6, 1UM7), to which experimentally determined chemical shift were also avaialble. I checked the compliance of the ensembles generated by the simulation with the experimentally determined parameters with the CoNSEnsX+ server, and with its new functionality, I created subenesmbles with selection based on chemical shift compliance. By extending these ensembles with other PDZ models from the PDB, I created a "PDZ core" ensemble, which was one of our main references during the comparative analysis. The different ensembles were concatenated into a larger ensemble by a MAMMOTH alignment applied to all models.

Chapter 3

New scientific results

3.1 First thesis group: Analysis of SAH domains

1. I implemented a helicity filter in the FT_CHARGE algorithm to filter out the detected structures, which are not expected to be helical.

In addition to the presence and order of charged amino acids, the co-composition of amino acids in the sequence also plays a role in the formation (or prevention) of the helix, e.g., high proline sequences generally do not form a helix. To address this problem, I even incorporated a helicity filter on the detected SAH domains into the detection workflow based on the Chou-Fasman method.

During the later development of the web server, I further refined the filter because I noticed that it does not work reliably for long, more complex SAH sequences. To refine the filter, I applied the filter not only to the entire SAH sections, but because FT_CHARGE analyzes segments in a given window size due to the FFT procedure, I appled it to each segment identified by FT_CHARGE, and I only used the sections that have 'passed' through the filter to assemble the entire SAH domain. This ensures that a long SAH domains won't get filtered out of the final filter, but that the relevant, expected truly helical section can be kept. Both filtering steps have been incorporated into the csahdetect.pl script, which concatenates the sections found by FT_CHARGE, establishes consensus with the SCAN4CSAH procedure, and thus determines the exact boundaries of the SAH sections given as final output [II, III].

2. I optimized the parametrisation of the FT_CHARGE algorithm using proteins proven for containing SAH domains.

The emergence of new, experimentally determined SAH domains has highlighted the weaknesses of existing algorithms, namely that the algorithms did not detect those SAH domains whose existence was experimentally proven, necessitating a

Name	UniProt ID	SAH start	SAH end	Suggested parameters A>=7 P<=0.05	High sensitivity A=0.5 P<=0.5	Previous values A>=10 P<=0.01
Caldesmon	A0A1L1RXH5	196	252	174-316 408-521	1-36 153-384 394-528	178-312 411-517
GCP60	Q9H3P7	183	238	158-274	158-283	158-268
INCENP	P53352	503	715	502-679	484-769	
MAP4K4	O95819	417	480	374-503	355-533	386-500
MFAP1	P55081	267	344	225-259	173-273	
Myosin 10	Q9HD67	813	909	773-870	765-880 903-955	
Myosin 7	P97479	866	935	848-929	829-939	
Myosin 6	Q9UM54	915	980	932-995	889-923 932-995	932-994
Snu23	G0S6R0	131	164		110-214	

Table 3.1: *The optimized parametrisation for the FT_CHARGE algorithm based on experimentally determined SAH domains*

review of the parameterization of the algorithms [III].

To create a better parameterization for the FT_CHARGE algorithm, I used 9 experimentally verified SAH domains and the proteins containing them, as well as negative controls. In the experimental setup, I ran the FT_CHARGE procedure with the multiple combinations of its parameters for the proteins in the setup and I examined the proportion of amino acids in the SAH sections that could be identified with each parameterisation by the consensus method (SCAN4CSAH + FT_CHARGE) and with only using the FT_CHARGE algorithm. Based on the examination of the obtained results, I proposed new parametrisation. It is important to note that the prediction/ascertainment of exact domain boundaries is not realistic, as from an experimental point of view it is not expected to produce and analyze a large number of constructs by adding amino acids one by one, and even in this case it would not be easy and clear e.g. in the case of the α-helical character obtained by CD spectroscopy, to draw a border between the valuesobtained for each construct. Thus, the determination of exact from-to positions for SAH segments is not necessarily viable, because the transition is presumably more or less continuous towards disordered and / or coild coil sections along the sequence.

3. I have identified previously unknown SAH-containing proteins with the optimized paramatrisation, some of which are also involved in neural signaling.

The consensus method predicted 8 SAH domain containing proteins that are also involved in neuronal signaling. Septin 7 protein belongs to a relatively newly identified family of cytoskeletal proteins, which form fibers and rings that act as scaffolds and diffusion barriers. They are involved in, among other things, the development of dendritic spikes and the process of cell division. Cytoskeletal (e.g., CLIP2, myosin 6, caldesmon) proteins and proteins found in the cell cortex (piccolo, drebrin) are involved in the development of cell appendages such as dendrites and the release of synaptic vesicles. During the search for SAH domains, we found 17 proteins that can also be found in SynaptomeDB, and the UniProt annotation of 13 proteins suggests that their function can also be related to neural development. The PCLO (protein piccolo) protein is located in presynaptic active zones and is involved in the transport of synaptic vesicles [3]. Drebrin is a postsynaptic protein that has a role in the formation of cytoskeletal scaffolding in the dendrites and may be associated with synaptic changes in the context of long-term potentiation and Alzheimer's disease [11].

	Cytoskeleton	Cell cortex	Paraspeckles	Exon-exon junctions	Golgi apparatus	RNA binding	MAPKKKK activity	Neuronal process	SynaptomeDB
All	**31**	**10**	**3**	**3**	**18**	**21**	**3**	**13**	**17**
Cytoskeleton		9		1	4	2	1	8	9
Cell cortex					2			5	6
Paraspeckles						3		1	1
Exon-exon junctions						3			
Golgi apparatus						3	1	3	
RNA binding								1	3
MAP kinase kinase kinase kinase activity								2	2
Neuronal process									8

Figure 3.1: *SAH domain containing human functional groups identified with the consensus method*

3.2 Second thesis group: The development of the CoNSEnsX server

1. I reimplemented and added new features to the existing CoNSEnsX server.

For ease of development and sustainability, I re-implemented the existing CoNSEnsX server in Python and created the Docker environment for the server and the new database. If the task allowed, I used standard libraries (PDB, STAR-NMR format file management, graphing, principal component analysis) when implementing most functionalities. I implemented RDC (residual dipolar couplings) back-calculation to handle and display the data sets from one measurement together and the back-calcuation of side chain order parameters. I also provided a completely new graphical interface to the server.

2. I implented a subensemble selection algorithm in the CoNSEnsX+ server.

One of the main new functionalities of the CoNSEnsX+ server is selection algorithm, in which I maximize the correspondence of the user-specified parameter sets along the selected compliance measure (correlation, rmsd, q-value). The method can be used to produce sub-sensembles that have good/better correspondence to the selected experimental parameters, also usually much smaller in size than the initial ensemble. The smaller size of the sensebles also reduces the overfitting of the parameters and the computational requirements for their processing.

After the selection process, the metrics for the selected ensemble are displayed and the selected ensemble becomes available for download. In addition, the selected ensemble is compared to the initial ensemble by PCA analysis, the visualization of which becomes visible after selection.

3. I showed that the implemented selection method is capable of producing non-overfitted sub-ensembles reflecting biologically relevant motions.

The biological relevance of the selection algorithm implemented in the CoNSEnsX+ server was tested on several ensembles generated with molecular dynamics simulations, by comparing conformational spaces sampled from different ensembles. Based on the principal component analysis of the initial and selected ensembles, it can be said that the movements observed in the reference ensemble can be well represented by smaller ensembles, which also indicates over-fitting of the reference ensemble. It is possible to create a much smaller ensemble which represents well

the conformational space of the initial one. The algorithm was also tested with with intrinsically disordered protein (IDP) protein structures using the 9AAC α-synuclein ensemble of the pE-DB [13] database. The principal component analysis of the initial and selected ensembles proved here that the selection algorithm is able to produce a smaller sub-ensebmle, even in the case of disordered proteins, while preserving their relevant properties.

3.3 Third thesis group: Investigation of the dynamics of PDZ domains

1. I created dynamic structural ensembles of PDZ domains with molecular dynamics simulation.

During the analysis of the studied ensembles, there was a need for a larger ensemble that could be used as a reference, representing the PDZ domain in general. To compile this, I searched for additional ensembles for which experimentally determined chemical shifts are available (2KPL, 2LC6, 2LC7, 2LOB, 1GM1, 1OZI, 2KPK, 1VJ6, 1UM7) to our expand existing NOE-based PSD-95 PDZ1 and PDZ2 ensembles [7]. I then ran short (20 ns) molecular dynamics simulations without constraining parameters, and then used the selection function of the CoNSEnsX$^+$ server to maximize the compliance with the chemical shifts Hα and Cα.

PDB ID	Description	BMRB ID	Subens. size
Short simulations			
2KPK	MAGI-1 PDZ1	16558	11
2KPL	MAGI-1 PDZ1 / E6CT	16559	7
1GM1	Second PDZ Domain (PDZ2) of PTP-BL	5131	9
1VJ6	PDZ2 from PTP-BL in complex with the C-terminal ligand from the APC protein	5131	12
2LC6	Solution structure of Par-6 Q144C/L164C	17599	13
2LC7	Solution structure of the isolated Par-6 PDZ domain	17600	17
1OZI	The alternatively spliced PDZ2 domain of PTP-BL	5762	11
1UM7	Solution structure of the third PDZ domain of synapse-associated protein 102	11198	21
Long simulations			
2KPK	MAGI-1 PDZ1	16558	11
2KPL	MAGI-1 PDZ1 / E6CT	16559	14
1GM1	Second PDZ Domain (PDZ2) of PTP-BL	5131	25
1VJ6	PDZ2 from PTP-BL in complex with the C-terminal ligand from the APC protein	5131	29

Table 3.2: *Ensembles made with molecular dynamics simluation and the size of thei subensembles after the selection process*

2. For a comparative analysis of the motions of the PDZ domains, I compiled a reference ensemble called the "PDZ core".

In order to concatenate the produced ensembles into a common ensemble, their structures had to be fitted first. The fit applied to several structures was made with MAMMOTH-Mult [8]. For this, we used the first models of each set and the first chains of multi-chain structures. Structures containing tandem PDZ domains have been splitted up so that the PDZ domains can be used separately. Since it was not possible to fit all 152 structures at the same time with the MAMMOTH-Mult program, the alignment was done in several rounds, and the results of each round were concatenated, retaining the parts that can be found in all the used structures. Using the resulting alignment, I reduced all PDB models and ensembles to include only the common amino acids obtained during the alignment, which will be one of the bases for further comparisons, which we named the "PDZ core".

3. By comparative analysis of the examined PDZ domains, I showed that the dynamic changes due to ligand binding are domain-dependent.

Contrary to our expectations, the distribution of free and bound forms along the open-close motion is not uniform for each PDZ domain. In the case of PSD-95 PDZ1 and PDZ2, the free form samples the entire space along this movement, the bound form moving in the part corresponding to the closed conformation of this space. However, the bound form of the PDZ3 domain of PSD95, in contrast, stabilizes in the open form in the ligand-bound state. This observation remains true even when examining the structures available in PDB for these domains, although in this case the differences are less spectacular because they represent a substantially smaller conformational space.

This observation is also true for two other populations selected on the basis of chemical shifts, as well as for the initial PDB models used for them. The only exception is the ensemble pair with PDB IDs 1GM1-1VJ6 (the second PDZ domain of the PTP-BL protein, in the free and bound state), where the elements of the original populations are separated only along the third principal component.

Chapter 4

Publications

4.1 Publications serving as a basis for the present work

I D. Dudola, B. Kovács, and Z. Gáspári. Evaluation and Selection of Dynamic Protein Structural Ensembles with CoNSEnsX. *Methods Mol Biol*, 2112:241–254, 2020.

II Á. Kovács, D. Dudola, L. Nyitray, G. Tóth, Z. Nagy, and Z. Gáspári. Detection of single alpha-helices in large protein sequence sets using hardware acceleration. *J. Struct. Biol.*, 204(1):109–116, 10 2018.

III D. Dudola, G. Tóth, L. Nyitray, and Z. Gáspári. Consensus Prediction of Charged Single Alpha-Helices with CSAHserver. *Methods Mol. Biol.*, 1484:25–34, 2017.

IV D. Dudola, B. Kovács, and Z. Gáspári. CoNSEnsX+ Webserver for the Analysis of Protein Structural Ensembles Reflecting Experimentally Determined Internal Dynamics. *J Chem Inf Model*, 57(8):1728–1734, 08 2017.

V D. Dudola, A. Hinsenkamp, and Z. Gáspári. Ensemble-based analysis of the dynamic allostery in the PSD-95 PDZ3 domain in relation to the general variability of PDZ structures. *Int J Mol Sci*, 21(21), 2020

Bibliography

[1] M. Batchelor, M. Wolny, L. Dougan, E. Paci, P. J. Knight, and M. Peckham. Myosin tails and single α–helical domains. *Biochem. Soc. Trans.*, 43(1):58–63, Feb 2015.

[2] P. Carter, C. A. Andersen, and B. Rost. DSSPcont: Continuous Secondary Structure Assignments for Proteins. *Nucleic Acids Res.*, 31(13):3293–3295, Jul 2003.

[3] S. D. Fenster, W. J. Chung, R. Zhai, C. Cases-Langhoff, B. Voss, A. M. Garner, U. Kaempf, S. Kindler, E. D. Gundelfinger, and C. C. Garner. Piccolo, a presynaptic zinc finger protein structurally related to bassoon. *Neuron*, 25(1):203–214, Jan 2000.

[4] S. Griep and U. Hobohm. PDBselect 1992-2009 and PDBfilter-select. *Nucleic Acids Res*, 38(Database issue):D318–319, Jan 2010.

[5] Baruch Z. Harris and Wendell A. Lim. Mechanism and role of pdz domains in signaling complex assembly. *Journal of Cell Science*, 114(18):3219–3231, 2001.

[6] Albert Hung and Morgan Sheng. Pdz domains: Structural modules for protein complex assembly. *The Journal of biological chemistry*, 277:5699–702, 03 2002.

[7] B. Kovács, N. Zajácz-Epresi, and Z. Gáspári. Ligand-dependent intra- and interdomain motions in the PDZ12 tandem regulate binding interfaces in postsynaptic density protein-95. *FEBS Lett*, 594(5):887–902, 03 2020.

[8] Dmitry Lupyan, Alejandra Leo-Macias, and Angel R. Ortiz. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics*, 21(15):3255–3263, 06 2005.

[9] Michelle Peckham and Peter J. Knight. When a predicted coiled coil is really a single α-helix, in myosins and other proteins. *Soft Matter*, 5:2493–2503, 2009.

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[11] T. Shirao, K. Hanamura, N. Koganezawa, Y. Ishizuka, H. Yamazaki, and Y. Sekino. The role of drebrin in neurons. *J. Neurochem.*, 141(6):819–834, 06 2017.

[12] S. Sivaramakrishnan, B. J. Spink, A. Y. Sim, S. Doniach, and J. A. Spudich. Dynamic charge interactions create surprising rigidity in the ER/K alpha-helical protein motif. *Proc. Natl. Acad. Sci. U.S.A.*, 105(36):13356–13361, Sep 2008.

[13] M. Varadi, S. Kosol, P. Lebrun, E. Valentini, M. Blackledge, A. K. Dunker, I. C. Felli, J. D. Forman-Kay, R. W. Kriwacki, R. Pierattelli, J. Sussman, D. I. Svergun, V. N. Uversky, M. Vendruscolo, D. Wishart, P. E. Wright, and P. Tompa. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.*, 42(Database issue):D326–335, Jan 2014.