Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

Detection of single alpha-helices in large protein sequence sets using hardware acceleration

Ákos Kovács[1,*], Dániel Dudola[1,*], László Nyitray[2], Gábor Tóth[3], Zoltán Nagy[1], Zoltán Gáspári[1]

[1]Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary
[2]Department of Biochemistry, Eötvös Loránd University, Budapest, Hungary
[3]Department for Research and Development, National Research, Development and Innovation Office, Budapest, Hungary

[*]These authors contributed equally to this work.

Correspondence to: nagy.zoltan@itk.ppke.hu, gaspari.zoltan@itk.ppke.hu

Abstract
Single alpha-helices (SAHs) are increasingly recognized as important structural and functional elements of proteins. Comprehensive identification of SAH segments in large protein datasets was largely hindered by the slow speed of the most restrictive prediction tool for their identification, FT_CHARGE on common hardware. We have previously implemented an FPGA-based version of this tool allowing fast analysis of a large number of sequences. Using this implementation, we have set up of a semi-automated pipeline capable of analyzing full UniProt releases in reasonable time and compiling monthly updates of a comprehensive database of SAH segments. Releases of this database, denoted CSAHDB, is available on the CSAHserver 2 website at *csahserver.itk.ppke.hu*. An overview of human SAH-containing sequences combined with a literature survey suggests specific roles of SAH segments in proteins involved in RNA-based regulation processes as well as cytoskeletal proteins, a number of which is also linked to the development and function of synapses.

Keywords: Single alpha-helix; FPGA; protein structure prediction; cytoskeleton; RNA processing

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

Introduction

The Single Alpha-Helix (SAH), previously also called Charged Single Alpha-Helix (CSAH) has been identified as a structural motif more than ten years ago (Knight et al. 2005, Sivaramakrishnan et al. 2008, Baboolai et al. 2009, Süveges et al. 2009, Peckham and Knight 2009). Its characteristic structural feature is the presence of alternating blocks of positively and negatively charged residues, most often Arg, Lys and Glu. The name CSAH was coined to reflect this, although a considerable net charge of these segments is not a distinctive feature, thus, the name SAH is today more widely used in the literature. SAH segments display a considerable helical propensity in aqueous solution in their monomeric form. Their stability has been investigated using a number of designed constructs, leading to the conclusion that $K_i \rightarrow E_{i+4}$ ion pairs are more stabilizing than $E_i \rightarrow K_{i+4}$ pairs (Baker et al. 2015) and that E-R pairs are favored over E-K ion pars (Wolny et al. 2017). SAH segments are relatively rare in proteomes (Gáspári et al. 2012) and their abundance and exact physiological role has been most thoroughly investigated in motor proteins (Simm et al. 2017), where they can act as an extension of the lever arm (Spink et al. 2008)

To our knowledge, there are currently three approaches to predict the presence of SAH segments in protein sequences: SCAN4CSAH (Süveges et al. 2009) and Waggawagga (Simm et al. 2015, Simm and Kollmar, 2018) use scoring schemes to discriminate between favored and unfavored interactions, whereas FT_CHARGE (Süveges et al. 2009) detects oppositely charged residues alternating with specified frequencies using Fourier transformation. SCAN4CSAH and FT_CHARGE were developed in parallel and are integrated into a single web service to offer consensus prediction (Gáspári et al. 2012). The idea behind this was to reduce the false positive rate of SAH prediction, as FT_CHARGE detects much less segments than SCAN4CSAH. We regard the use of a consensus as an asset, but it is clear that the cost of this, especially the exclusion of some of the results of the more stringent method is a likely elevated false negative rate.

SCAN4CSAH is not only more permissive than FT_CHARGE but is also much faster. In fact, the use of FT_CHARGE on large protein sets alone was, until very recently, highly limited by its low speed. Therefore, the suggested offline detection pipeline included a SCAN4CSAH run first and then invoking FT_CHARGE on those sequences only in which a SCAN4CSAH hit was found. To overcome this limitation, we have recently implemented the FT_CHARGE algorithm on FPGA, achieving a speedup of three orders of magnitude (Nagy et al. 2016). This allowed us to set up a semi-automated pipeline capable of processing monthly releases of the UniProt database without compromising the performance of the individual detection methods. This also allows us

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

to provide a detailed analysis and comparison of the results obtained with the two methods.
We have also performed an evaluation of the predictive power of our method and have adjusted the default parameters to yield better predictions. It should be noted that the number of reliably identified and experimentally characterized SAH segments is still low to make a comprehensive sensitivity-specificity analysis.

Materials and Methods

The performance of the FT_CHARGE method was tested on a small dataset of 9 SAH segments that could be unambiguously identified from the literature. Because the boundaries of the segments tested experimentally could not be, in general, regarded as the longest SAH segments or the segments with highest helicity, we restricted our analysis on the true positive hits in terms of residues. We have run FT_CHARGE with different combinations of parameters, varying the minimum amplitude and the maximal P-value for segment identification. These parameters are described in detail in our previous publications (Süveges et al. 2009, Gáspári et al. 2012). Briefly, the amplitude is the height of the largest peak as identified with the Fourier transform, and the P-value measures the significance of the peak height in a segment containing given fractions of positively and negatively charged amino acid residues. P-values are calculated based on an EVD (extreme value distribution) fit to the amplitudes of segments built up from A, R and E residues with different R and E fractions, 5000 for each 10%-wide range.

All analysis was performed on the full UniProt database, release 2018_03 (The UniProt Consortium 2018). Scripts were run separately on the SwissProt and TrEMBL parts, then the relevant result were merged.
Technical details of the FPGA (field-programmable gate array) implementation of FT_CHARGE has been described (Nagy et al. 2016). It provides all the functionality of the original Perl version of the algorithm and produces the same detailed output. The acceleration achieved on FPGA enabled us to analyze the full UniProt database with this method, for segment lengths 32 and 64. SCAN4CSAH was run separately with default parameters on all sequences. The positions of SAH segments were determined as described, merging the results of FT_CHARGE obtained with different window sizes and applying the helicity filter based on Chou-Fasman scores (described in detail in Dudola et al. 2017). In brief, the filter is based on an EVD fit to the average Chou-Fasman helical propensities (Chou & Fasman 1977) of helices at least 15-residue long

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

as obtained from the DSSP assignment of PDB Select chains (Griep & Hobohm 2010). By default, SAH segments with a permissive threshold of P<=0.5 are retained in order to discard those that are clearly expected to be non-helical. In this work this step was also performed on the segments identified by FT-CHARGE before their merging. The default is to use a P-value that is half of the P-value used for filtering the full assembled SAH segments. This additional step was necessary to ensure that no long SAHs are rejected because of the inclusion of long FT-CHARGE segments with helix-breaker residues at the final check. For the consensus results, FT_CHARGE and SCAN4CSAH segments were combined using the intersection of the predicted segments with the appropriate minimum length, determined as the shortest segment/window size set for any of the two algorithms. The minimum length of SAH segments reported in the database and analyzed in this study is 30 residues.

The semi-automated web service was set up using standard Unix tools, in-house scripts and example code provided by the UniProt consortium.

Analysis of basic properties of SAH segments was performed on two nonredundant datasets, one derived from the consensus and the other from the FT_CHARGE-only results. For the consensus dataset, redundant sequences were removed by running CD-hit on all FASTA files containing the masked sequences of SAH-containing proteins with a similarity cutoff of 70%. Using the masked sequences ensured that the SAH segments do not influence redundancy filtering. For the FT_CHARGE-only dataset, all proteins in which no SAH segments were detected by the consensus method were removed from the sequences first and redundancy filtering with CD-hit was performed on this set as described above. Analysis of residue and ion pair abundance was performed with in-house Perl scripts.

Statistical overrepresentation analysis of Gene Ontology (Ashburner et al. 2000, The Gene Ontology Consortium 2017) terms in the protein sets was performed using the PANTHER web service (Mi et al. 2017) using the AmiGO site (Carbon et al. 2009) using the Fisher exact test with FDR correction (available as of December 5, 2017),

Results and Discussion

Evaluation of the predictions by FT_CHARGE

At the time of its first implementation, FT_CHARGE was specifically designed to restrict the number of predicted SAH segments by applying strict parameters. This consideration was applied in order to minimize the number of false positives in the case of a newly described structural motif and avoid the overestimation of its significance.

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

However, the fact that both SCAN4CSAH and FT_CHARGE focused only on the charged residues, several potential false positive segments were still identified with a regular repetitive charge pattern but containing a high number of proline residues. An example of such a segment is a 333-residue long region of RNA-binding protein 12B (RB12B_HUMAN), described in one of our previous surveys (Gáspári et al. 2012). In order to increase the stringency of FT_CHARGE, a helicity filter was introduced that is capable of discarding segments with low overall helical propensity (Dudola et al. 2017). On the other hand, the recent identification of a number of SAHs in various proteins highlighted that the method with the original parameterization is unable to identify existing SAHs, thus, a re-evaluation of the parameters is necessary. Unfortunately, the number of natural SAH segments characterized experimentally is still low and the boundaries of these are not explored systematically because of the disproportionally large experimental resources needed to perform residue-by-residue scans. In addition, defining negative controls is both trivial - any non-helical segment can be selected - and highly sophisticated - best negative controls are segments that closely match the properties of SAHs yet do not form stable single helices. Therefore, here we focused on the identification of residues in SAH segments in 9 proteins were such segments have been characterized experimentally. Even one of these segments, described in the spliceosomal protein Snu23 (Ulrich et al. 2016) displays a very low helical propensity as measured by CD spectroscopy (~20% helix at 6 °C), hardly corresponds to the properties expected from a SAH, although we are not aware of criterions formulated on a similar basis. We have selected a set of 10 negative controls also, including globular all-alpha, coiled coil and disordered proteins as well as RNA-binding protein 12B, a proline-rich segment that was predicted to contain a long SAH before the introduction of the helicity filter (Table S1).

Table 1. List of experimentally characterized SAH segments used for performance optimization

Name	**UniProt ID**	**Uniprot AC**	**SAH start**	**SAH end**	**reference**
Caldesmon	A0A1L1RXH5	A0A1L1RXH5_CHICK	196	252	Wang & Wang, 1996
GCP60	Q9H3P7	GCP60_HUMAN	183	238	Süveges et al. 2009.
INCENP	P53352	INCE_CHICK	503	715	Samejima et al. 2015.
MAP4K4	O95819	M4K4_HUMAN	417	480	Süveges et al. 2009.
MFAP1	P55081	MFAP1_HUMAN	267	344	Ulrich et al. 2016
Myosin 6	Q9UM54	MYO6_HUMAN	915	980	Spink et al. 2008.
Myosin 7	P97479	MYO7A_MOUSE	866	935	Li et al. 2017.
Myosin 10	Q9HD67	MYO10_HUMAN	813	909	Wolny et al. 2014.

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

Snu23	G0S6R0	G0S6R0_CHATD	131	164	Ulrich et al. 2016

After invoking FT_CHARGE on these nine protein sequences with varying parameters A (amplitude) from 0 to 20 in steps of 0.5 as well as with P-values 0.001, 0.005, 0.01, 0.05, 0.1, and 0.5 we have chosen A=7 and P=0.05 as new default parameters (Table S2). We note that no SAH segment in Snu23 is identified with these parameters and in MFAP1 a segment before and not overlapping with the experimentally characterized SAH region is predicted (Table S3). While the former was expected, the latter observation is somewhat surprising although it should be noted that the presence of another/longer SAH segments is compatible with the experimental results described in Ulrich et al. 2016.

The set of our negative control sequences produced no predicted SAH segments even with the most permissive parameters. We note, however, that at present our method is not able to efficiently discriminate between the designed (*'de novo'*) SAH and SAH-like sequences analyzed recently where issues like low solubility and aggregation strongly influence whether these truly behave as SAHs or not. Addressing such issues would require the introduction of other filters that take into account other features besides charge distribution, when more such experiments become available.

Performance of the FPGA implementation of FT_CHARGE on UniProt

The FPGA implementation was run on a Digilent ZedBoard (http://zedboard.org/product/zedboard) equipped with a Xilinx Zynq XC7Z020 FPSoC (Field Programmable System-on-Chip) device, 512MB on-board memory and Gigabit Ethernet interface.

The speed of the FPGA-based FT_CHARGE algorithm allows monthly reanalysis of the full UniProt database (Table S4). In order to account for all possible changes between releases a full analysis is preferred.

Setup of the semi-automated pipeline to generate CSAHdb

We have set up an analysis pipeline capable of performing comprehensive analysis of the full UniProt database on a monthly basis. Upon the availability of a new UniProt release, a notification is sent to the operators to activate the FPGA platform and initiate the processing pipeline. The pipeline downloads the data, sends it to the FPGA and receives the processed output of the FT_CHARGE algorithm. After a manual check of the success of the first part, a script is started that is responsible for invoking SCAN4CSAH and SAH segment determination. This includes running the script

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

csahdetect.pl and also obtaining fully annotated UniProt txt files for proteins with putative SAH segments.

We note that our decision, i.e. re-running the full analysis on each UniProt release instead of just on the newly added/updated sequences is based on the consideration that reliable and comprehensive identification of all UniProt changes would require an additional step with the possible introduction of novel types of errors. The speed of the FPGA-based analysis is so high that currently there would be no benefits from introducing such an extra step even if it would greatly reduce the number of sequences to be analyzed.

As a result, four kinds of files are generated for consensus SAHs and the proteins in which they are located:

- A list of SAHs identified. Each SAH is in a separate line containing the UniProt ID of the protein, the consensus location and the corresponding segments identified by SCAN4CSAH and FT_CHARGE
- A FASTA format file with the positions of the SAH segments masked with ‘x’ characters
- A FASTA format file containing the sequence of each SAH identified
- A fully annotated UniProt text file with the location of the SAH segments added to the FT section.

These types of files are generated separately for SwissProt and TrEMBL entries. For SAHs identified using FT_CHARGE only, the first three types of files are also provided in a separate directory. The files are available for download in a single zip archive from the site *csahserver.itk.ppke.hu*.

The net execution time of the full pipeline is within 48 hours even with using the original Perl implementation of all steps except FT_CHARGE. This means that the semi-automated pipeline, including the delays while awaiting manual intervention, can be completed well within a week, allowing monthly releases of the SAH database. We plan to make the pipeline fully automated in the near future.

SAH segments detected in the full UniProt by the different algorithms

We have analyzed all SAH segments in the full UniProt database that were not included in the consensus but were predicted by FT_CHARGE (Table 2). It should be noted that in a number of cases the SAH segment predicted by FT_CHARGE covers multiple segments detected by the consensus method. In such cases, we counted the SAH as ‘FT_CHARGE only’ if less than 50% of the FT_CHARGE detected residues were included in consensus SAH segments.

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

Table 2: SAH segments identified with the consensus method and FT_CHARGE only. For comparison, the results with both the adjusted and previously used parameters are given. Data refer to UniProt release 2018_03.

Dataset	Consensus SAHs		SAHs detected with FT_CHARGE only	
	SAH segments	Proteins with SAHs	SAH segments	Proteins with SAHs
Adjusted default parameters (A>=7, P<=0.05)				
SwissProt	1137	985	242	237
TrEMBL	121171	108455	30367	30044
Previously used parameters (A>=10, P<=0.01)				
SwissProt	629	555	78	77
TrEMBL	64327	57631	7753	7688

SAHs detected only by FT_CHARGE were identified as having less than 50% of their residues listed in the consensus identification. It should be noted that in a number of cases the FT_CHARGE-only SAH segments overlap multiple consensus-based SAHs. Nevertheless, it is safe to say that in this way we were able to predict the presence of about 10% more SAH segments than the consensus method.
Adjustment of the helicity filtering as described in the methods section resulted in fully consistent results meaning that there were no segments identified by the consensus method that were not overlapping with SAHs detected using FT_CHARGE only. Without the adjustment, nonhelical segments (i.e. those containing prolines) identified by FT_CHARGE could get incorporated into the full SAH regions leading to their rejection at the final helicity check step.
Naturally, at the present stage it is hard to provide an estimate of the error rate of our detection pipeline as very few SAHs are characterized experimentally, although the number of these is continuously growing.

We have analyzed some basic features of the identified SAH segments using nonredundant sets of consensus- and FT_CHARGE-only-detected SAH-containing proteins (Figure 1). Interestingly, the percentage of arginine residues within positively charged ones defined as $100*N_R(N_R+N_K)$ (where N_R represents the number of arginine and N_K the number of lysine residues in the SAH segment) ranges from 0 to 100 with a median around 50, independent of the length of the SAHs.). Intriguingly, there is no

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

apparent difference in the normalized abundance of different $i \rightarrow i+4$ ion pairs in the SAH segments. The lower abundance of ion pairs in FT_CHARGE-only detected segments can be explained on the basis that the consensus results reflect the scoring scheme of SCAN4CSAH favoring such interactions, whereas FT_CHARGE focuses on the regularity on the repeating charge pattern and detects SAHs with somewhat lower charge density. However, the count of FT_CHARGE-only detected SAHs used in this analysis is an order of magnitude lower than those in the consensus-based set.

Our observations showing uniform distribution of Arg and Lys residues as well as different ion pairs are apparently at odds with those obtained on experimental analysis of charged helical segments. One interpretation might be that our predictions have a larger false positive rate than desired as we identify segments with interactions not favored in single alpha-helical segments. Another possibility can be that the majority of the SAHs predicted here form - supposedly differently, but in principle largely - stable single helices when they are in the context of the full proteins under cellular conditions. This issue can only be resolved by detailed experimental investigation of potential SAH-containing proteins. In turn, our prediction methodology might need further adjustments as experimental data accumulate.

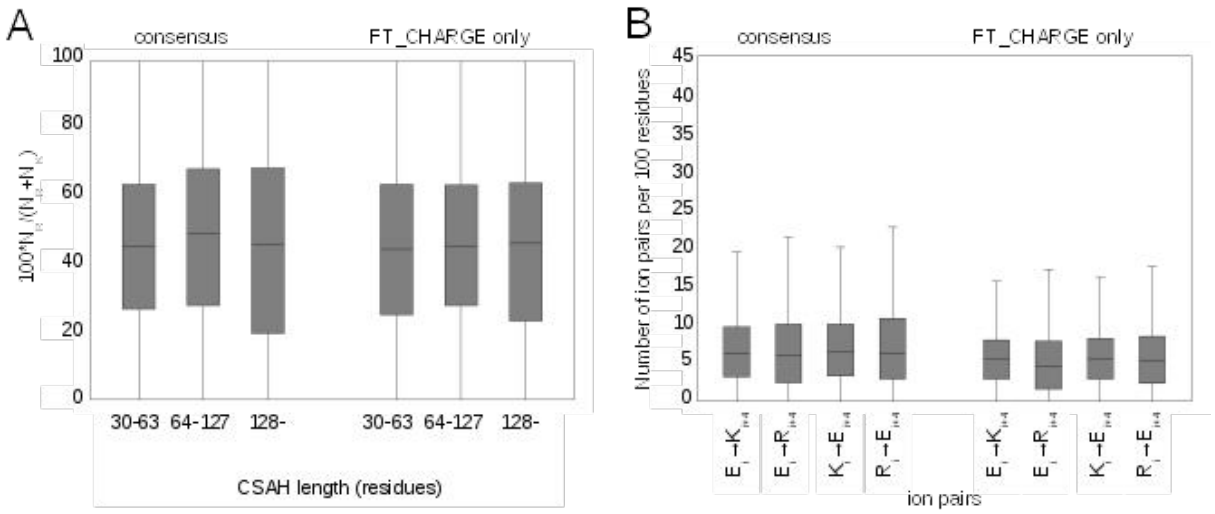

Figure 1. Charged residues and ion pairs in consensus and FT_CHARGE-only detected SAH segments (using UniProt release 2018_03), A) Distribution of $N_R/(N_R+N_K)$ percentages in SAH segments of different lengths. B) Normalized abundance of selected ion pairs (outliers not shown in the box plots).

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

Analysis of human SAH-containing proteins

Our analysis identified almost 400 human proteins with SAH segments with the consensus method and 500 using FT_CHARGE only, 111 of which were represented in AmiGO. FT_CHARGE predicts over 100 SAHs not recognized using the consensus (Table 3). Note that FT_CHARGE-only SAHs are defined as SAHs with less than 50% of their residues identified by the consensus.

Table 3. Number of SAH segments and SAH-containing proteins in human UniProt sequences (release 2018_03)

	SAH segments	*Proteins*
Consensus	*483*	*396*
FT_CHARGE	*527*	*473*
FT_CHARGE only	*138*	*128*

GO-term enrichment analysis of human proteins (background: all human proteins) yielded results consistent with earlier findings about the prevalence of cytoskeletal proteins as well as proteins involved in RNA processing (Tables S5-S6). To be explored in more detail we have selected four of the cellular components enriched in SAH-containing proteins: the paraspeckle, exon-exon junctions, the cytoskeleton and the cell cortex, as well as MAP kinase kinase kinase kinase activity. (Table 4). All these are obtained when analyzing consensus and FT-CHARGE-only results in PANTHER.

Table 4. Overview of selected functional groups of human SAH-containing proteins obtained with the consensus method. Besides GO terms for cellular components and molecular functions, proteins with annotated neuronal functions and/or listed in SynaptomeDB are also shown.

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

	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

Table 5. Position and sequence of SAH segments identified in the proteins discussed in detail.

Protein	region	sequence
Q9HAU5 (UPF2)	48-115	EVSKAPEDKKKRLEDDKRKKEDKERKKKDEEKVKAEEESKKKEEEEKKKHQEEER KKQEEQAKRQQEE
Q9H1J1 (UPF3A)	237-303	EERRRRELEKKRLREEEKRRRREEERCKKKETDKQKKIAEKEVRIKLLKKPEKGE EPTTEKPKERGE
Q9BZI7 (UPF3B)	209-269	RMREEKREERRRRREIERKRQREEERRKWKEEEKRKRKDIEKLKKIDRIP
P45379 (TNNT2)	142-185	AERAEQQRIRNEREKERQNRLAEERARREEEENRRKAEDEARKK
P13805 (TNNT1)	112-151	EQQRFRTEKERERQAKLAEEKMRKEEEEAKKRAEDDAKKK
P45378 (TNNT3)	116-155	EQQRIRAEKERERQNRLAEEKARREEEDAKRRAEDDLKKK
Q16181 (SEPT7)	366-419	KEKVQKLKDSEAELQRRHEQMKKNLEAQHKELEEKRRQFEDEKANWEAQQRILE

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

Q15811 (ITSN1)	630-662	LKQKEQERKIIELEKQKEEAQRRAQERDKQWLE
	668-706	DEHQRPRKLHEEEKLKREESVKKKDGEEKGKQEAQDKLG
Q9NZM3 (ITSN2)	671-737	KLKEIERKRLELMQKKKLEDEAARKAKQGKENLWKENLRKEEEEKQKRLQEEKTQ EKIQEEERKAEE
O95819 (MAP4K4)	378-480	EQQLREQEEYKRQLLAERQKRIEQQKEQRRRLEEQQRREREARRQQEREQRRREQ EEKRRLEELERRRKEEEERRRAEEEKRRVEREQEYIRRQLEEEQRHLE
Q9UKE5 (TNIK)	364-459	RSEALRRQQLEQQQRENEEHKRQLLAERQKRIEEQKEQRRRLEEQQRREKELRKQ QEREQRRHYEEQMRREEERRRAEHEQEYIRRQLEEEQRQLE
Q8N4C8 (MINK1)	396-467	RRIEEQKEERRRVEEQQRREREQRKLQEKEQQRRLEDMQALRREEERRQAEREQE YKRKQLEEQRQSERLQR
Q86SQ0 (PHLDB2)	1034-1097	RIEEMERLLKQAHAEKTRLLESRERFMEAKKRALEEEKRRREILEKRLQEETSQR QKLIEKEVK
Q08AD1 (CAMSAP2)	1197-1237	KQQLEAEMEHKKEETRRKTEEERQKKEDERARREFIRQEYM
Q9Y6V0 (PCLO)	1212-1251	KPLPEEKKLIPEEEKIRSEEKKPLLEEKKPTPEDKKLLPE
Q16643 (DBN1)	173.238	KRINREQFWEQAKKEEELRKEEERKKALDERLRFEQERMEQERQEQEERERRYRE REQQIEEHRRK
Q9UDT6 (CLIP2)	413-485	KLQRARLLVESVRKEKVDLSNQLEEERRKVEDLQFRVEEESITKGDLETQTQLEH ARIGELEQSLLLEKAQAE

Proteins involved in RNA-related processes: paraspeckle and EJC components

The occurrence and role of SAH segments in paraspeckle-containing proteins has been previously investigated in detail by sequence analysis and molecular modeling. This analysis suggested that SAH segments might play a role in the exact positioning of core paraspeckle protein dimers along the NEAT1 long noncoding RNA (Dobson et al. 2015). PANTHER analysis also identified three proteins in exon-exon junctions (Figure 2). All three are involved in nonsense-mediated mRNA decay. In a detailed study on the role of intrinsic disorder in NMD, two of these have been previously reported to contain a SAH segment (Kalmar et al. 2012), whereas UPF3A, a close homolog of UPF3B, has been added to the list in our analysis. In UPF2, the SAH region is at the N-terminus (residues 48-115) and overlaps a coiled coil region (54-134) annotated in UniProt. This scenario is reminiscent to that observed in paraspeckle proteins where the SAH was also identified as a continuation of a long coiled coil and a similar role of the SAH segment can be proposed. The position of the N-terminal helix in the X-ray structure of the first MIF4G domain (PDB id 4CEM, residues 122-149, Clerici et al. 2014.) of this protein seems to be compatible with the presence of a rod-like structure pointing away from the domain.

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

In both UPF3A and UPF3B the SAH is in the central region of the proteins preceded by a structured and followed by a potential intrinsically disordered part. Available structural data on UPF2/3 and their complexes do not provide hints on the position or role of the SAH-containing part of these proteins (Kadlec et al. 2004, Buchwald et al. 2010, Melero et al. 2012, Melero et al. 2014). However, it is clear that there are multiple modes of interactions possible between the components of the NMD apparatus (Melero et al. 2014). As the UPF2-UPF3 complex links the exon-junction complex to SMG1, we propose that the SAH segments in UPF2/3 proteins play a role in defining the proper distance within the NMD machinery between the EJC and effector components.

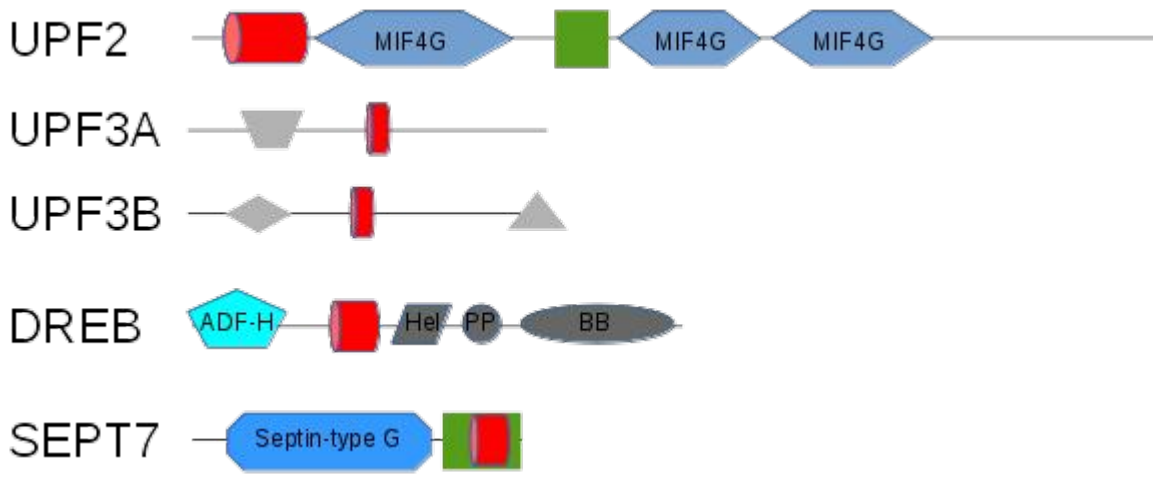

Figure 2. Schematic composition of selected SAH-containing proteins Red cylinders represent SAH segments, other UniProt-annotated domains are depicted with blue shapes, other regions with grey shapes. For the NMD-involved proteins UPF3A and UPF3B these correspond to regions with solved structures in the PDB, for DREB (Drebrin) these are depicted based on the annotation in Worth et al. 2013. Green boxes represent annotated coiled coil segments not or only partially overlapping with the predicted SAHs.

Cytoskeletal proteins

The consensus method identified 17 proteins containing SAHs, joined by one protein in which only FT_CHARGE predicts the presence of a single helix. It is of note that all three troponin T proteins (TnC, TnI and TnT) are among the hits.

Perhaps the best studied of these is cardiac troponin T. Still, a detailed structural characterization of the region identified here as SAH is still lacking. In cTnT, the predicted SAH region is located at the N-terminal region of the central domain that is involved in tropomyosin binding and supposed to be mainly alpha-helical (Katrukha

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

2013). Importantly, a deletion mutant containing only 3 of 4 consecutive Glu residues causes serious structural disruptions in cardiac muscle (Moore et al. 2013).

Septin 7 is a member of the septin family of cytoskeletal proteins capable of forming filaments and rings that can act as scaffolds and diffusion barriers and are involved in many processes including cell division and the development of dendritic spines (Mostowy & Cossart 2012). Septins can assemble to hexamers and octamers with no polarity and these are believed to associate via their coiled coil domains (Weirich et al. 2008). The SAH segment in septin 7 is located near its C-terminus, within its coiled coil region. It is interesting to note that Septin 7 is the terminal protein in the hexamers and thus the SAH might be important in directing the higher-order interaction of hexamers with each other. Interestingly, no SAHs were detected in other septins even in other organisms in SwissProt.

Intersectins are cytoskeletal scaffold proteins involved in endocytosis and with suggested roles in autism and Down syndrome (Hunter et al. 2013). Their predicted SAH domain is located in the N-terminal half of the central coiled coil region between blocks of 2 EH and 5 SH3 domains, characteristic of the protein family. The coiled coil region has been shown to interact with a number of proteins and intersectins are also capable of homodimerization (Wong et al. 2012). The exact role of the putative SAH region can thus be manifold.

Using FT_CHARGE only, CAP-Gly domain-containing linker protein 2 (CLIP2) is also identified as harboring a SAH segment and our PANTHER analysis on the enlarged protein set (including SAHs identified only by FT_CHARGE) lists it at a cytoskeletal component. This protein seems to be involved in brain-specific organelle translocation (De Zeeuw et al. 1997). Its predicted SAH segment (residues 417-480 according to the consensus prediction) is part of the first of its three UniProt-annotated coiled coil regions after the two CAP-Gly domains. To our knowledge, no detailed experimental structural information is available on this protein apart from the structure of its CAP-Gly domains (PDB entries 2CP2 and 2CP3). It is interesting to note that CLIP2 proteins from a number of different species are found by FT_CHARGE but not using our consensus-based approach.

Cell cortex is defined in as the region immediately below the plasma membrane. It usually contains actin filaments and other associated proteins (Biro et al. 2013). Our PANTHER analysis identified 4 proteins with SAH segments that are localized here, all of them also components of the cytoskeleton.

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

MAP kinase kinase kinase kinases

Using the optimized parametrization, three proteins with MAP kinase kinase kinase activity are predicted to bear SAHs: Mitogen activated kinase kinase kinase 4 (MAP4K4) with an experimentally characterized SAH segment (Süveges et al. 2009), the Traf2- and Nck-interacting kinase (TNIK) and misshapen-like kinase 1 (MINK1). In all three proteins, the SAH is located between the N-terminal kinase and the C-terminal CNH domains. TNIK is involved in the organization of the cytoskeleton (Taira et al. 2004) and is abundant in dendritic spines (Burette et al. 2015), whereas MINK1 is essential for cytokinesis (Hyodo et al. 2012). The role of the SAH domain in these proteins remains at present elusive.

Proteins involved in neuronal processes

Although SAH-containing proteins do not seem to be significantly enriched in proteins involved in neuronal processes, it is of note that a number of our hits are associated with such. This is because of the involvement of specific cytoskeletal elements (like CLIP2, Myosin VI and caldesmon) and cell cortex-located proteins (e.g. protein piccolo and drebrin) in development of cell appendages like dendrites and involvement in synaptic vesicle release. 17 human proteins identified in our SAH analysis are listed in SynaptomeDB and for 13 we found implications in their UniProt annotation to be linked to neural development (Table 4), e.g. for Pleckstrin homology-like domain family B member 2 (PHLDB2) is involved in the assembly of the postsynaptic complex in neuromuscular junctions (Madhavan and Peng, 2005), whereas Calmodulin-regulated spectrin-associated protein 2 (CAMSAP2) binds to the minus end of microtubules and plays a part in the regulation of neuronal polarity (Yau et al. 2014). We are, however, not aware of structural/functional studies that could account for the exact role of the SAH-containing segment in these proteins.

Protein piccolo (PCLO) is located in presynaptic active zones and is required for synaptic vesicle trafficking (Fenster et al. 2000). Its region containing the predicted SAH segment (residues 1212-1251 in the consensus prediction) has been identified as a coiled coil and treated as such in an attempt to build a full structural model of the protein, contributing to its estimated 80 nm-long elongated shape (Gundelfinger et al. 2014). We note that despite passing our helicity filter, this region contains several proline residues and thus it is questionable whether it is able to adopt a straight helical structure. Nevertheless, it is located between two segments with strong coiled coil propensity according to COILS (Lupas 1991) and might act as a hinge region with still considerable helical propensity. Apart from this speculation, the exact function of the

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

putative SAH segment in this protein apart from its generally assumed spacer role remains to be established.

Drebrin is a postsynaptic protein playing a part in shaping the cytoskeletal scaffold in dendrites and is associated with synaptic changes in long-term potentiation and Alzheimer's disease (Shiaro et al. 2017). Its SAH segment is located in of a coiled coil (CC) region annotated in Pfam. This region, together with an adjacent helical segment (Hel) binds actin and thus one drebrin molecule can be able to link two actin filaments together. In this context, the role of the SAH segment might be interpreted as necessary to maintain the distance between the two filaments straddled by drebrin and/or precluding multiple interactions with the same filament. The interaction site on the CC domain can be blocked by intramolecular binding of the C-terminal 'BB" domain to the ADF-CC domain boundary, leaving only one site accessible to actin. Phosphorylation of Ser142 located before predicted SAH (residues 179-216 in the consensus prediction) relieves the other actin binding site (Worth et al. 2013).

Conclusions

Using an FPGA implementation of the inherently slow and stringent method FT_CHARGE, we have established a semi-automated pipeline that allows identification of charged single alpha-helical segments in full UniProt monthly releases without any compromises. Thus, a comprehensive set of SAH-containing proteins can be made available for further analysis. We note that based on similar considerations, a command-line version of Waggawagga, suitable for the analysis of large datasets, has been published recently (Simm and Kollmar, 2018). Our present overview of human SAH-bearing proteins reveals that this comprehensive analysis together with our growing knowledge of individual proteins can lead to novel insights of the distribution and biological role of SAH segments.

Acknowledgements

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

Funded by the Hungarian Human Resources Development Operational Programme (EFOP-3.6.2-16-2017-00013). Z.G is a recipient of a János Bolyai Research Fellowship from the Hungarian Academy of Sciences.

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25-29.

Baboolai, T.G., Sakamoto, T., Forgacs, E., White, H.D., Jackson, S.M., Takaki, Y., Farrow, R.E., Molloy, J.E., Knight, P.J., Sellers, J.S., Peckham, M. 2009. The SAH domain extends the functional length of the moysin lever. Proc Natl. Acad. Sci. USA 106: 22193-22198.

Baker, E.G., Bartlett, G.J., Crump, M.P., Sessions, R.B., Linden, N., Faul, C.F.J., Woolfson, D.N. 2017. Local and macroscopic electrostatic interactions in single α-helices. Nat. Chem. Biol. 11, 221-228.

Biro M., Romeo, Y., Kroschwald, S., Bovellan, M., Boden, A., Tcherkezian, J., Roux, P.P., Charras, G., Paluch, E.K. 2013. Cell cortex composition and homeostasis resolved by integrating proteomics and quantitative imaging. Cytoskeleton 70, 741-754.

Buchwald, G., Ebert, J., Basquin, C., Sauliere, C., Jayachandran, U., Bono, F., Le Hir, H., Conti, E. 2010. Insights into the recruitment of the NMD machinery from the crystal structure of a core EJC-UPF3b complex. Proc. Natl. Acad. Sci. USA 107, 10050-10055.

Burette, A.C., Phend, K.D., Burette, S., Lin, Q., Liang, M., Foltz, G., Taylor, N., Wang, Q., Brandon, N.J,. Bates, B., Ehlers, M.D., Weinberg, R.J. 2015. Organization of TNIK in dendritic spines. J. Comp. Neurol. 523, 1913-1927.

Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., Lewis, S., AmiGO Hub, Web Presence Working Group. 2009. AmiGO: online access to ontology and annotation data. Bioinformatics.25, 288-289.

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

Clerici, M., Deniaud, A., Boehm, V., Gehring. N.H., Schaffitzel, C., Cusack, S. 2014. Structural and functional analysis of the three MIF4G domains of nonsense-mediated decay factor UPF2. Nucleic Acids Res. 42, 2673-2686.

Chou, P.J., Fasman, G.D. 1977. Secondary structural prediction of proteins from their amino acid sequence. Trends Biochem. Sci. 2, 128-131.

De Zeeuw, C.I., Hoogenraad, C.C., Goedknekt, E., Hertzberg, E., Neubauer, A., Grosveld, F., Galjart, N. 1997. CLIP-115, a novel brain-specific cytoplasmic linker protein, mediates the localization of dendritic lamellar bodies. Neuron 19, 1187–1199.

Dobson, L., Nyitray, L., Gáspári, Z. 2015. A conserved charged single alpha-helix with a putative steric role in paraspeckle formation. RNA 21, 2023-2029.

Dudola, D., Tóth, G., Nyitray, L., Gáspári, Z. 2017. Consensus prediction of charged single alpha-helices with CSAHserver. In: Zhou, Y., Kloczkowski, A., Faraggi, E., Yang, Y. (Eds): Prediction of Protein Secondary Structure (Methods Mol. Biol. Vol. 1847), Humana Press, New York, pp. 25-34.

Fenster, S.D., Chung, W.J., Zhai, R., Cases-Langhoff, C., Voos, B. Garner, A.B., Kaempf, U., Kindler, S. Gundelfinger, E.D., Garner, C.C. 2000. Piccolo, a presynaptic zinc finger protein structurally related to Bassoon. Neuron, 25, 203-214.

Gáspári, Z., Süveges, D., Perczel, A., Nyitray, L., Tóth, G. 2012. Charged single alpha-helices in proteomes revealed by a consensus prediction approach. Biochem. Biophys. Acta - Proteins and Proteomics 1824, 637-646.

Griep, S., Hobohm, U. 2010. PDBSelect 1992-2009 and PDBfilter-select. Nucleic Acids Res. 38, D318-D319.

Gundelfinger, E.D., Reissner, C. Garger, C.C. 2016. Role of Bassoon and Piccolo in assembly and molecular organization of the active zone. Front. in Synaptic Neurosci. 7, 19.

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

Hunter, M.P., Russo, A., O'Bryan, J.P. 2013. Emerging role of Intersectin (ITSN) in regulating signaling and disease pathways. Int. J. Mol. Sci. 14, 7829-7852.

Hyodo, T., Ito, S., Hasegawa, H., Asano, E., Maeda, M., Urano, T., Takahashi, M., Hamaguchi, M., Senga, T. 2012. Misshapen-like kinase 1 (MINK1) is a novel component of striatin-interacting phosphatase and kinase (STRIPAK) and is required for completion of cytokinesis. J. Biol. Chem. 287, 25019-25029.

Kadlec, J., Izaurralde, E., Cusack, S. 2004. The structural basis for the interaction between nonsense-mediated mRNA decay factors UPF2 and UPF3. Nat. Struct. Biol. 11, 330-337.

Kalmar, L., Acs, V., Silhavy, D., Tompa. P. 2012. Long-range interactions in nonsense-mediated mRNA decay are mediated by intrinsically disordered protein regions. J. Mol. Biol. 424, 125-131.

Katrukha, I.A. 2013. Human Cardiac troponin complex. Structure and functions. Biochemistry (Moscow) 78, 1447-1465.

Knight, P.J. , Thirumurugan, K., Xu, Y., Wang, F., Kalverda, A.P., Stafford, W.F. III, Sellers, J.R., Peckham, M. 2005. The predicted coiled coil domain of moysin 10 forms a novel elongated domain that lengthens the head. J. Biol. Chem. 280, 34702-34708.

Li, J., Chen, Y., Deng, Y., Unarta, I.C., Lu, Q., Huang, X., Zhang, M. 2017. Ca^{2+}-induced rigidity change of the myosin VIIa IQ motif-single α helix lever arm extension. Structure 25, 579-591.

Lupas, A., Van Dyke, M., Stock, J. 1991. Predicting coiled-coils from protein sequences. Science 252, 1162-1164.

Madhavan, R., Peng, H.B. 2005. Molecular regulation of postsynaptic differentiation at the neuromuscular junction. IUBMB Life 57, 719-730.

Melero, R., Buchwald, G., Castano, R., Raabe, M., Gil, D., Lazaro, M., Urlaub, H., Conti, E., Llorca, O. 2012. The cryo-EM structure of the UPF–EJC complex shows UPF1 poised toward the RNA 3′ end. Nat. Struct. Mol. Biol. 19, 498-505.

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

Melero, R., Uchiyama, A., Castano, R., Kataoka, Kurosawa, H., Ohno, S., Yamashita, A., Llorca, O. 2014. Structures of SMG1-UPFs complexes: SMG1 contributes to regulate UPF2-dependent activation of UPF1 in NMD. Structure 22, 1105-1119.

Moore, R.K., Grinspan, L.T., Jimenez, J., Guinto, P.J., Ertz-Berger, B., Tardiff, J.C. 2013. HCM-linked Δ160E cardiac troponin T mutation causes unique progressive structural and molecular ventricular remodeling in transgenic mice. J. Mol. Cell. Cardiol. 58, 188-198.

Mostowy, S., Cossart, P. 2012. Septins: the fourth component of the cytoskeleton. Nat, Rev. Mol. Cell. Biol. 13, 183-194.

Mi, H., Huang. X., Muruganujan. A., Tang, H., Mills, C., Kang, D., Thomas, P.D. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 45, D183-D189.

Nagy, Z., Gáspári, Z., Kovács, Á. 2016. Accelerating a charged single alpha-helix search algorithm in protein sequences using FPGA. In: Tetzlaff, R. (Ed.) CNNA 2016. VDE Verlag GmbH, Berlin, pp. 117-118.

Peckham, M., Knight, P.J. 2009. When a coiled coil is really a single alpha helix, in myosins and other proteins. Soft Matter 5, 2493-2503.

Samejima, K., Platani, M., Wolny, M., Ogawa, H., Vargiu, G., Knight, P.J., Peckham, M., Earnshaw, W.C. 2015. The inner centromere protein (INCENP) coil Is a single α-helix (SAH) domain that binds directly to microtubules and is important for chromosome passenger complex (CPC) localization and function in mitosis. J. Biol. Chem. 290, 21460-21472.

Shiaro, T., Hanamura, K., Koganezawa, Ishizuka, Y., Yamazaki, H., Sekino, Y. 2017. The role of drebrin in neurons. 141, 819-834.

Simm, D., Hatje, K., Kollmar, M. 2015. Waggawagga: comparative visualization of coiled-coil predictions and detection of stable single α-helices (SAH domains). Bioinformatics 31, 767-769.

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

Simm, D., Hatje, K., Kollmar, M. 2017. Distribution and evolution of stable single α-helices (SAH domains) in myosin motor proteins. PLoS ONE 12, e0174639.

Simm, D., Kollmar, M. 2018. Waggawagga-CLI: A command-line tool for predicting stable single α-helices (SAH-domains), and the SAH-domain distribution across eukaryotes. PLoS ONE 13, e0191924

Sivaramakrishnan S., Spink B.J., Sim, A.Y., Doniach, S., Spudich, J.A. 2008. Dynamic charge interactions create surprising rigidity in the ER/K alpha-helical protein motif. Proc. Natl. Acad. Sci. USA 105, 13356-13361.

Spink, B.J., Sivaramakrishnan S., Lipfert, J., Doniach, S., Spudich, J.A. 2008. Long single alpha-helical tail domains bridge the gap between structure and function of myosin VI. Nat. Struct. Mol. Biol. 15, 591-597.

Süveges, D., Gáspári, Z., Tóth, G., Nyitray, L., 2009. Charged single α-helix: a versatile protein structural motif. Proteins 74, 905-916.

Taira, K., Umikawa, M., Takei, K., Myagmar, B-E., Shinzato, M., Machida, M., Uezato, H., Nonaka, S., Kariya, K. 2004. The Traf2- and Nck-interacting kinase as a putative effector of Rap2 to regulate actin cytoskeleton. J. Biol. Chem. 279, 49488-49496.

The Gene Ontology Consortium, 2017. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 45, D331-D338.

The UniProt Consortium, 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45, D158–D169.

Ulrich, A.K.C., Seeger, M., Schütze, T., Bartlick, N., Wahl, M.C. 2016. Scaffolding the spliceosome with single α-helices. Structure 24, 1972-1983.

Wang, E., Wang, A.C-L. 1996. (i, i+4) ion pairs stabilize helical peptides derived from smooth muscle caldesmon. Arch. Biochem. Biophys. 329, 156-162.

Published in final edited form as: *J Struct Biol.* 204 (2018) 109-116, doi: 10.1016/j.jsb.2018.06.005

Weirich, C.S., Erzberger, J.P., Barral, Y. 2008. The septin family of GTPases: architecture and dynamics. Nat. Rev. Mol. Cell. Biol. 9, 478-489.

Wolny, M., Batchelor, M., Knight, P.J., Paci, E., Dougan, L., Peckham, M. 2014. Stable single α-helices are constant force springs in proteins. J. Biol. Chem. 289, 27825-27835.

Wolny, M., Batchelor, M., Bartlett, G.J., Baker, E.G., Kurzawa, M., Knight, P.J., Dougan, L., Woolfson, D.N., Paci, E., Peckham, M. 2017. Characterization of long and stable de novo single alpha-helix domains provides novel insight into their stability. Sci. Rep. 7, 44341.

Wong, K.A., Wilson, J., Russo, A., Wang, L., Okur, M.N., Wang, X., Martin, N.P., Scappini, E., Carnegie, G.K., O'Bryan J.P. 2012. Intersectin (ITSN) family of scaffolds function as molecular hubs in protein interaction networks. PLoS ONE 7, e36023

Worth, D.C., Daly, C.N., Geraldo, S., Oozer, F. Gordon-Weeks, P.R. 2013. Drebrin contains a cryptic F-actin–bundling activity regulated by Cdk5 phosphorylation. J. Cell Biol. 202, 793-806.

Yau, K.W., van Beuningen, S.F.B., Cunha-Ferreira, I. Cloin, B.M.C., van Battum, E.Y., Will, L., Schatzle, P., Tas, R.P., van Krugter, Katrukha, E.A., Jiang, K., Wulf, P.S., Mikhalyova, M., Harterink, M., Pasterkamp, R.J., Makhmanova, A., Kapitein, L.C., Hoogenraad, C.C. 2014. Microtubule minus-end binding protein CAMSAP2 controls axon specification and dendrite development. Neuron 82, 1058-1073.