Mapping the amino acid composition of the peripheral membrane binding interface of PH domains

Kamilla Ormevik Jansen

Thesis is submitted in partial fulfilment of the requirements for the degree of Master of Science

Department of Biological Sciences
Faculty of Mathematics and Natural Sciences
University of Bergen

November 2021

Acknowledgements

I would like to thank my main supervisor Nathalie Reuter for the opportunity to do a master project in bioinformatics. Thank you for letting me be a part of the Reuter group at CBU and to be able to learn from you. Your knowledge and skill is incredibly inspiring. Thank you for your feedback and guidance on writing my thesis. I would also like to thank you for introducing me to Thibault Tubiana, and choosing him as my co supervisor.

I would like to give my biggest gratitude to my co supervisor Thibault Tubiana. I am forever grateful for all your help and support throughout my project. You have always been there for me when I have needed it. Thank you for teaching me how to program, thank you for your help in the writing process, and thank you for everything in between. Thank you for calming my nerves whenever they were running wild. You always say the right thing. Thank you for sharing and discussing with me and for being so patient when I had troubles understanding things.

I also want to thank everyone in the Reuter group for letting me feel welcomed and included, even in the lockdown period. Thank you for creating a warm and friendly work environment. A special thanks to Dandan Xue for sharing her office with me and helping me in the process of learning to program. Thank you for being a good friend and colleague.

I want to thank Lill Kristin Knudsen, not only for her incredible help in the end of my master project, but for everything she has helped me with throughout my years of studying. You have been there for me since the beginning of my bachelor studies and followed me to the end of my master. It has been a roller-coaster, and I am so grateful for your support and motivation.

At last, I would like to thank my parents, Evy Ormevik and Trond Jansen, and my partner Kim Villanger for their support throughout this journey. Thank you for always believing in me.

Bergen, November 2021
Kamilla Ormevik Jansen

Table of Content

Abbreviations

General abbreviations

AA – Amino acid

β1, β2, β3 – First, second and third beta strand and so on..

IBS – Interfacial binding site

MSA – Multiple sequence alignment

PDB – Protein Data Bank

PH domain – Pleckstrin Homology domain

PMP – Peripheral membrane protein

PIPs – Phosphatidylinositols

PIP2 – Phosphatydylinositol 4,5-biphosphate, Phosphatidylinositol 3,4-biphosphate, Phosphatidylinositol 3,5-biphosphate

PIP3 - Phosphatidylinositol 3,4,5-trisphosphate

PMP – Peripheral membrane protein

X – Any amino acid

Abbreviations made for this project

B1 – Canonical membrane binding

B2 – Non-canonical membrane binding

L1 and Loop 1 - Loop linking β1/β2

L2 and Loop 2– Loop linking β3/β4

L3 and Loop 3 – Loop linking β5/β6

M1 – Motif for canonical membrane binding KX(K/R)XR

M2 – Motif for non-canonical membrane binding (K/R)X(W/Y/F)

X(n) – Amino acids between K and (K/R) in the beginning of M1, where X is any amino acid and n is a number

X(-2) – Amino acid between (K/R) and R in the end of M1, it is called X(-2) because it is the second AA counting backwards

Summary

Biological membranes are important in the organization of cells, and the main components of membranes are different kinds of lipids. Peripheral proteins are soluble and interact transiently and bind reversibly to lipids in membranes. Pleckstrin Homology (PH) domains are protein domains found in peripheral membrane proteins, which bind to lipids called phosphatidylinositides (PIPs). Different PH domains have a well conserved structure but low sequence identity. About 15% of PH domains bind to PIPs in membranes with high affinity, but varying specificity. Characterization of the membrane binding sites in terms of amino acid content and structure is still incomplete. Thus, the main goal of this project is to fill a part of this knowledge gap by mapping the amino acid composition of the membrane binding interface (IBS) of PH domains. This was done using a bioinformatical approach as it allows to gather and analyze large datasets of protein structures and sequences.

To address the aim of the project the work was divided into four subparts. I first collected datasets of PH domain sequences and structures, and aligned them. Secondly, datasets and multiple sequence alignments were analyzed to find PH domains with sequence patterns described in the literature to be important for membrane binding. The length of the IBS loops was calculated and we made an inventory of the PH domain secondary structure elements. Thirdly, the amino acid composition of the membrane binding interface was mapped. Lastly, the level of conservation of amino acids in the peripheral membrane binding interface region was calculated.

The main goal of mapping the amino acid composition of the membrane binding interface of PH domains was successfully accomplished. The most important results and conclusions from the project are: 1) The level of basic amino acids follows the PIP binding pattern for canonical and non-canonical PIP binding. 2) Lysines are more common than arginines in the IBS, with only one structurally positioned exception. 3) Glycine is common in the IBS used for both types of specific membrane binding. 4) The amino acids important for specific membrane binding are also highly conserved in the PH domains which do not have the known amino acid patterns for specific binding to PIPs in membranes.

1. Introduction

1.1 Introduction to membranes

Biological membranes are important in the organization of cells. They provide a barrier between what is inside and what is outside of a given cell or organelle. All cells are surrounded by a cell membrane, and multiple organelles such as mitochondria, Golgi apparatus, endoplasmatic reticulum (ER), and vesicles are surrounded by membranes. The membranes are diverse, and they participate in many different tasks. They can take part in various tasks involving examples from energy providing, signalling, trafficking, to cell maintenance (Lemmon, 2007; Luckey, 2014). Biological membranes are made of different types of lipids, proteins and carbohydrates. The main component is usually different kinds of lipids like phospholipids, sphingolipids and sterols (Luckey, 2014). The different membranes can be distinguished from each other by the presence or absence of specific lipids (Lemmon, 2008; Falkenburger *et al.*, 2010). The lipids are organised in a bilayer consisting of two leaflets, where the lipids are amphipathic, meaning that they have a hydrophilic head-group that faces towards the water, and a hydrophobic tail-group, which is positioned tail to tail making the bilayer (Luckey, 2014). The hydrophobic effect explains the spontaneous formation of the membranes as the lipids aggregate to face their hydrophobic tails away from the water (Luckey, 2014). The fluid mosaic model explains that the lipids provide a sea where proteins are scattered like a mosaic pattern, and lateral movement of membrane components can happen because of the fluidity (Singer and Nicolson, 1972). Lipids can also flip-flop between the two membrane leaflets (Gurtovenko and Vattulainen, 2007).

1.1.1 Membrane proteins

The proteins scattered in the membrane can be either integral proteins or peripheral proteins. The integral proteins are embedded into the membrane and cannot be easily removed. These proteins can either be transmembrane (bitopic or polytopic) or set in one of the sides of the membrane (monotopic). The bi- and polytopic proteins usually function as channels or receptors, and the monotropic can have enzymatic functions (Blobel, 1980; Allen *et al.*, 2019).

The peripheral proteins on the other hand are soluble and they interact transiently and bind reversibly to the membranes (Fuglebakk and Reuter, 2018). These peripheral proteins are often involved in signalling and trafficking (Cho and Stahelin, 2005; Lemmon, 2008). The peripheral proteins can interact with membranes using different strategies. They can contain different membrane-targeting domains that can bind to the specific lipids in the membranes, as mentioned above, or use covalently attached lipid anchors. Some PMPs can use a part of the protein surface, like loops or amphipathic α-helices to bind reversibly to the membrane (Cho and Stahelin, 2005; Lomize *et al.*, 2007). The most common lipid binding targets for the membrane-targeting domains are acidic phospholipids. The phospholipid-protein domain interactions can be divided into two types. The first type of binding is very specific and involve specific recognition of a stereoisomer in a binding reaction. Stereoisomers are isomers who have the same constituents, but have them arranged in different orientations. In the second, the interaction is non-specific and involve attraction to the physical properties of the membrane which includes charge, amphiphilicity and curvature. Specificity can also be added by the requirement of certain second messengers for membrane binding. The binding-mechanism for the different membrane-targeting domains lay somewhere in between these extremes (Lemmon, 2008).

1.1.2 Phospholipids

Phospholipids have a glycerol backbone with two fatty acid chains connected by ester linkages, and a polar phosphate headgroup. A type of phospholipid is the phosphatidylinositol (PI), which has an inositol ring connected to the phosphate (Falkenburger *et al.*, 2010). Some of the phosphatidylinositides are phosphorylated at the 3, 4 and/or 5 position on the inositol ring, making different phosphoinositides (PIPs). Figure 1 shows the structural formula of three PIPs: phosphatidylinositol 4,5-biphosphate (PIP(4,5)2), phosphatidylinositol 3,5-biphosphate (PIP(3,5)2) and phosphatidylinositol 3,4,5-triphosphate (PIP(3,4,5)3). The phosphoinositides are called PIP1, PIP2 or PIP3 in respect to the number of phosphate groups on the inositol ring.

These phosphoinositides combined only account for less than one percent out of all phospholipids in mammalian cells (Lemmon, 2008). Because these phosphoinositides are so rare, they are the perfect way to make diverse and specialized membranes where PMPs can distinguish them from others and bind to them with high specificity. Another feature of the

specific phosphoinositide-protein binding is that it keeps the PMPs from becoming active until it reaches the correct membrane, while passing through the different membranes in the path of being synthesised and processed (Falkenburger *et al.*, 2010).

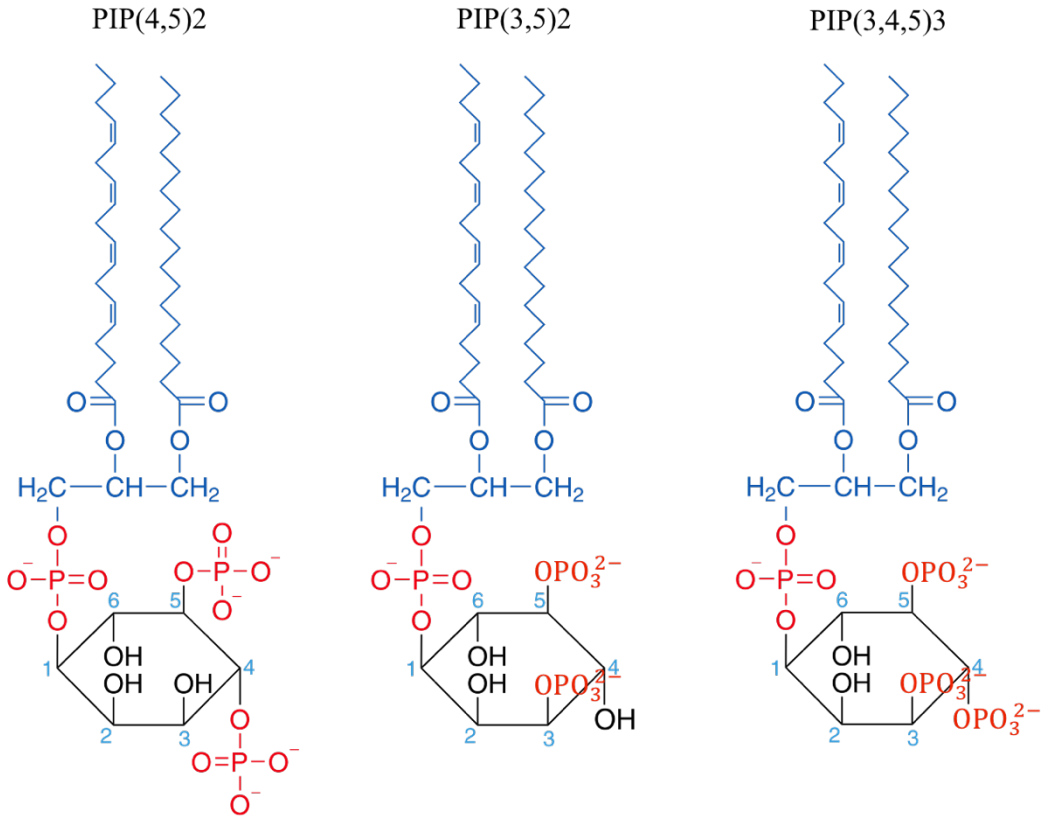

Figure 1 Phosphatidylinositol 4,5-biphosphate (PIP(4,5)2), Phosphatidylinositol 3,5-biphosphate (PIP(3,5)2) and Phosphatidylinositol 3,4,5-triphosphate (PIP(3,4,5)3). *The glycerol backbone and the fatty acid chains are blue, inositol ring is black, and the phosphate groups are red. The phosphate groups are in expanded form in PIP(4,5)2 and in condensed from in PIP(3,5)2 and PIP(3,4,5)3 for better visualisation. Altered from Wikipedia Commons (2021).*

1.2 Introduction to Pleckstrin Homology domains

The Pleckstrin Homology (PH) domain was first identified in 1993 by Haslam, Koide & Hemmings. The domain was identified twice in pleckstrin, which is the main component in blood platelets. Pleckstrin accounts for about one percent of the total protein in the platelet cells, and it is the main substrate for protein kinase C (PKC). PKCs are enzymes involved in protein regulation, by phosphorylating serine and threonine residues in target proteins. Pleckstrin has two homologous PH domains, one at each terminal, which is how the PH domain

got its name (Haslam, Koide and Hemmings, 1993; Webb, Hirst and Giembycz, 2000; Lian *et al.*, 2009)

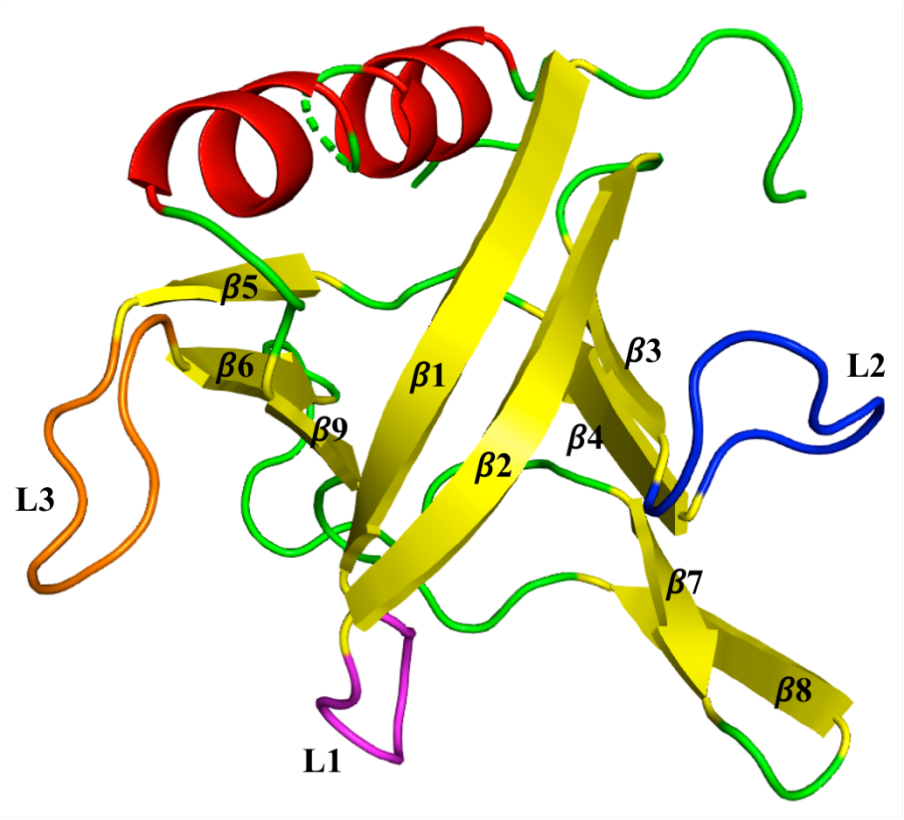

Figure 2 GRP1 PH domain*. β-strands are yellow, Loop 1 linking β1/β2 is magenta, Loop 2 linking β3/β4 is blue, Loop 3 linking β5/β6 is orange, all other loops are green and the C-terminal α-helix is red. CATH ID: 1fgyA00. The figure is made using PyMOL (Schrödinger, 2015)*

The PH domain is approximately 100-120 amino acids (AAs) long. The sequence identity between PH domains is very low, usually less than 30% (Cho and Stahelin, 2005). Despite this low sequence identity they have a well conserved structure. The number of β-strands and α-helices vary, but the domains usually contain a seven-stranded β-sandwich and a C-terminal α-helix (Timm *et al.*, 1994; Cho and Stahelin, 2005; Lemmon, 2008). The domain usually has two membrane-binding loops, but they can have more. The first one is located between the first two β-strands, the second one is located on the opposite side of the structure, often between β-strands three and four (Lemmon, 2008). Figure 2 shows a PH domain with three binding loops coloured.

PH domains can bind to the seven different phosphoinositides in cellular membranes, but some have a clear preference for PIP2 and PIP3 (Cho and Stahelin, 2005; Corey, Stansfeld and Sansom, 2019). Only about 15% of PH domains bind to PIPs with high affinity, and with varying specificity. In some cases binding of PH domains initiate clustering of PIPs in the membranes to increase the affinity, and many PH domains bind to three-five PIPs simultaneously (Cho and Stahelin, 2005; Yamamoto *et al.*, 2020). A lot is known about how and why many of the different PH domains bind to membranes, but the function of some of them are still unknown (Cho and Stahelin, 2005).

1.2.1 Canonical membrane binding, B1

About ten percent of PH domains which bind with high affinity and specificity usually does this in a common/canonical way, and they bind to the three PIPs with vicinal inositol-ring phosphate groups (PI(3,4)P2, PI(4,5)P2 and PI(3,4,5)P3) (Lemmon, 2007). Vicinal means that two functional groups are connected to two adjacent carbon atoms (Clayden, Greeves and Warren, 2012). Side chains from two conserved basic residues, one positioned in the end of the first β-strand (β1) and the other in the middle of the second β-strand (β2), make hydrogen bonds with the two adjacent lipid phosphate groups (Lemmon, 2007). This type of binding has a sequence fitting the motif of KX(n)(K/R)XR, (M1), where X is any amino acid, and K and R are lysine and arginine, respectively. The position of this binding motif in the structure can be seen in Figure 3.

The starting lysine positioned at the end of β1 binds to both vicinal inositol ring phosphates in PI(3,4)P2, PI(4,5)P2 or PI(3,4,5)P3, and the ending arginine in the middle of β2 binds to phosphate 3 or 5 depending on the PIP (Ferguson *et al.*, 2000; Lemmon, 2008). The (K/R) in the motif and other basic residues present in this region and elsewhere in the domain can make hydrogen bonds with phosphates outside of the vicinal phosphate pair in, for example with the 1-phosphate connecting the fatty acid chains and the inositol ring of the PIP (Figure 1) to increase the binding energy and/or specificity (Lemmon, 2007, 2008). Loop 2 linking β3/β4 also binds together with Loop 1 in this canonical binding way, and it usually contains many positive supporting amino acids (Naughton, Kalli and Sansom, 2018).

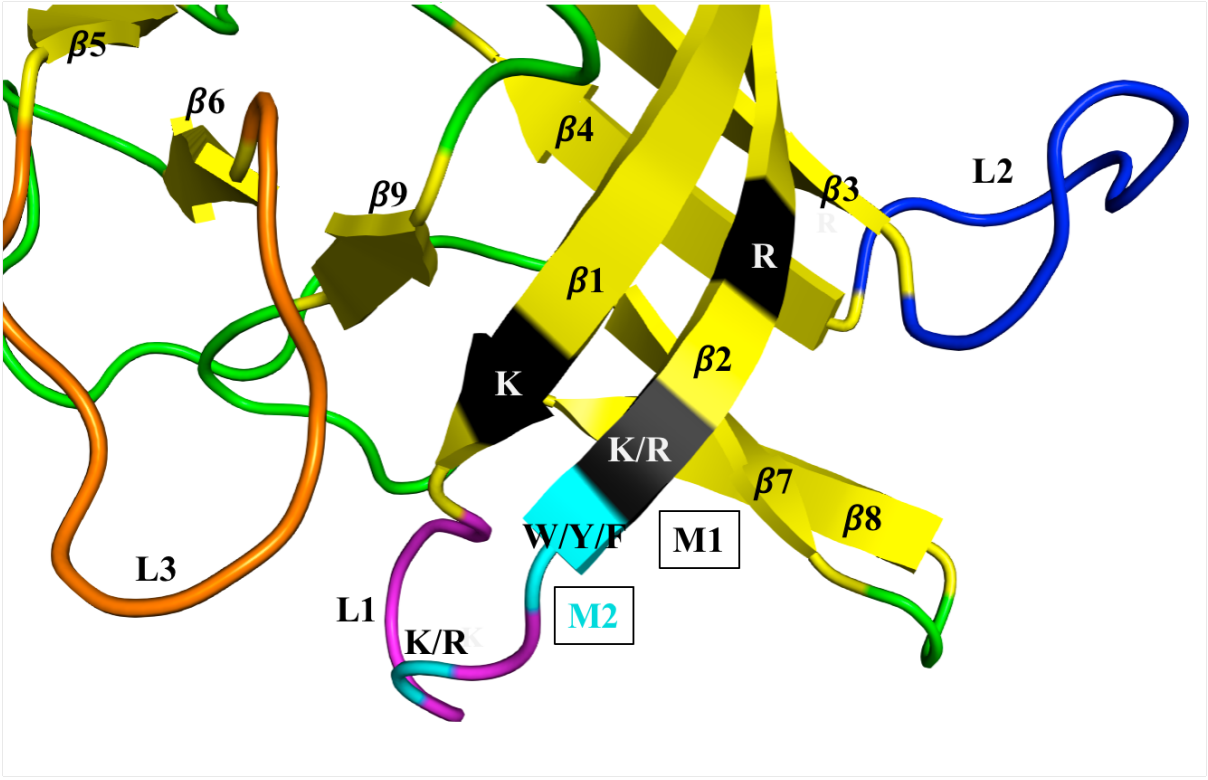

Figure 3 GRP1 PH domain with both canonical and non-canonical motif. *The position of amino acids in binding motifs in the structure is shown. Motif for canonical binding motif (M1) is KX(n)(K/R)XR and it is black. Starting K in the end of β1, (K/R) in the beginning/middle of β2 and the ending R in the middle of β2. The motif for non-canonical binding (M1) is (K/R)X(W/Y/F), The (K/R) is in the middle of L1 and the (W/Y/F) is positioned at the end of L1 or beginning of B2 (turquoise). β-strands are yellow, Loop 1 linking β1/β2 is magenta, Loop 2 linking β3/β4 is blue, Loop 3 linking β5/β6 is orange, all other loops are green. CATH ID: 1fgyA00. The figure is made using PyMOL (Schrödinger, 2015).*

1.2.2 Non-canonical membrane binding, B2

Some PH domains bind to PIPs in another way, a non-canonical binding (Naughton, Kalli and Sansom, 2016). This type of binding uses Loop 1 linking β1/β2, and Loop 3 linking β5/β6. This non-canonical binding requires a short binding motif (M2) located at the end of Loop 1 and beginning of β2, (K/R)X(W/Y/F), which can be seen in Figure 3. When a PH domain binds in this fashion there is usually a long Loop 3 containing many positive amino acids. Supporting amino acids elsewhere in the domain can also interact with the PIP. Interestingly, this binding site can bind phosphatidyl serine as well as PIPs (Naughton, Kalli and Sansom, 2018).

Some PH domains bind in either the canonical (B1) or the non-canonical way (B2), and some can bind in both ways. They can bind using one binding site at the time, or bind PIPs in both of them simultaneously. They can bind the same type of PIP or two different types in each of

the binding sites (Naughton, Kalli and Sansom, 2018; Yamamoto *et al.*, 2020). Generally the PH domain/lipid interaction is tighter when using the canonical binding, and the binding is weaker or more loose when binding in the non-canonical way. There is also normally a preference towards B1 binding when both the binding motifs are present, and the B2 site is more of a supporting site (Naughton, Kalli and Sansom, 2018). In the cases when the canonical binding (B1) is weak, the non-canonical binding (B2) can compensate to give an overall higher affinity. According to Naughton, Kalli and Sansom (2018), the PH domains that prefer B1 but can also bind with B2 have tyrosine or phenylalanine in their binding motif (M2) instead of the most common tryptophan.

1.2.3 Other types of PH domain bindings

Outside of the canonical binding for high specificity and high affinity PH domains and non-canonical binding, there are multiple other examples of ways for PH domains to bind to PIPs and other molecules. As mentioned earlier, most PH domains bind to all PIPs with low affinity and specificity, without a clear function (Lemmon, 2007). Interestingly, it has been reported that the canonical binding pocket is observed to be less defined in PH domains with low specificity like pleckstrin and spectrin (Hurley, 2006). Others have been shown to bind much more strongly too free Inositol-phosphate (Ins(1,4,5)P3, Ins(1,3,4,5)P4) than to their corresponding phosphoinositides (PI(4,5)P2, PI(3,4,5)P3) (Lemmon, 2007). A split PH domain is thought to bind other PIPs like PI(3)P1 (Lemmon, 2008). While PH domains usually do not have large hydrophobic protrusions (Hurley, 2006), it has been suggested that hydrophobic residues in the PH domain in some cases might penetrate the interfacial region to make hydrophobic interactions (Cho and Stahelin, 2005; Lemmon, 2007). Some PH domains bind both to PIPs with low specificity and to another protein (small g-protein) simultaneously, on two different surfaces on the PH domain. There are also some PH domains which do not bind PIPs or membranes at all, some cases with protein-protein binding, and some cases where the domain binds DNA (Hurley, 2006; Lemmon, 2007; Okuda *et al.*, 2017).

As recited above, the function and mechanism of the PH domain are quite diverse. It has been suggested that the PH domain is more of an example of conserved structural fold rather than functional conservation, where the β-sandwich structure is a stable scaffold where many different binding functions can be assigned (Lemmon, 2007).

1.3 Presentation of the master project

PH domains are very diverse. They have a low sequence identity, around 30% and bind to membranes, proteins or DNA, where they bind in various ways using different strategies. What they have in common is a well conserved structure and it is thought that around 15% bind to PIPs with high affinity, with varying specificity (Cho and Stahelin, 2005). Most studies done on PH domains involve only one or a few PH domains, and they are usually very close to each other in either function, sequence or evolution (Hyvönen *et al.*, 1995; Lemmon and Ferguson, 2000; Ceccarelli *et al.*, 2007; Jian *et al.*, 2015; Yamamoto *et al.*, 2016; Naughton, Kalli and Sansom, 2018). Other than the conserved PH domain structure there is a lack of general knowledge. There is especially a gap concerning the characterization of the membrane binding sites in terms of amino acid content and structure. This is a gap that needs to be filled, and a part of what this project will try to enlighten. To fill this gap it is important to look at as many PH domains as possible. To do this I will use a bioinformatical approach as it allows to gather and analyze large datasets of protein structures and sequences. More PH domains can be studied at the same time. Bioinformatics is defined as "conceptualizing biology in terms of molecules and applying informatics techniques to understand and organize the information associated with these molecules, on a large scale. In short, bioinformatics is a management information system for molecular biology and has many practical applications" (Luscombe, Greenbaum and Gerstein, 2001). Bioinformatic databases contain lots of biological information, and in this project databases concerning protein structure and sequence are the most useful to address the PH domain. Bioinformatic methods like multiple sequence alignment gives the possibility to compare this structural and sequential information and result in new knowledge about PH domains.

The PH domains are diverse and bind to different types of molecules. Thus, a master project about every aspect of PH domains would be too much to grasp over. I have therefore limited the project to look specifically on the peripheral membrane binding interface of the PH domains. The aim of the project is to map the amino acid composition of the membrane binding interface of PH domains. To address the aim the work is divided into four goal parts. I will first collect datasets of PH domain structures and sequences, and align them. Secondly, datasets and multiple sequence alignments will be analyzed to find PH domains with sequences matching motif for canonical and non-canonical membrane binding. I will also find positions and lengths of the binding loops of the interfacial binding site and quantify variations in secondary

structures of PH domains. Thirdly, the amino acid composition of the membrane binding interface will be mapped. The hypotheses for part two and three are that the membrane-binding loops are longer and contain more basic amino acids in the PH domains where they are used for peripheral membrane-binding. Lastly, the level of conservation of amino acids in the peripheral membrane binding interface region, used for both canonical and non-canonical membrane binding, will be found.

2. Materials and Methods

The materials and methods section is divided into two parts. In the first part I will give an introduction and explanation of the methods I have used for sequence alignment and the main database I have extracted data from. In the second part I will present the workflow and methodology for the master project.

2.1 Introduction to relevant tools and methods in bioinformatics

2.1.1 Sequence alignment without structural information

Sequence alignment is a method used to compare biological sequences to each other. Throughout evolution sequences that are related diverge from each other, but some regions may be conserved. These regions are the ones that are looked for by using sequence alignment. DNA, RNA or protein sequences can be compared to look for similarities in the pattern of the residues. Regions that are important for function or structure are usually conserved, while other regions tend to mutate more often and become more different between the sequences (Xiong, 2006).

Sequence homology is way of telling if a pair of sequences are descendant from a common evolutionary origin. If two protein sequences have more than 30% sequence identity, they are said to be homologous sequences. In proteins, sequence identity is the percentage of matches of the exact same amino acid aligned in the alignment. Sequence similarity on the other hand is the percentage of aligned amino acids with the same physiochemical properties (Xiong, 2006).

2.1.1.1 Dynamic programming

In 1970 Needleman & Wunsch presented a dynamic programming algorithm to globally align sequences. This type of alignment is done from the beginning to the end of the sequences, and the goal is to find the alignment that fits best across the entire sequence length. This alignment type works best for sequences of roughly the same length who are closely related (Needleman and Wunsch, 1970; Xiong, 2006). In 1981 Smith & Waterman presented another algorithm for alignment, which aimed to find the best local alignments between sequences. Local alignment finds local parts of the sequences with the highest similarity, and aligns these parts to each other, without considering the rest of the sequence length. This alignment type is most appropriate for aligning more divergent sequences to find conserved patches like domains or motifs (Smith and Waterman, 1981; Xiong, 2006).

For sequence alignment the optimal alignment between two sequences is found by using a scoring matrix, such as BLOSUM (Dayhoff, Schwartz and Orcutt, 1978) or PAM (Henikoff and Henikoff, 1992), where a score is assigned to each match and mismatch. The highest score will then give the best alignment. The way this is done is by making a matrix with two dimensions where the two sequences are the two axes. The residue matching is done row by row by using a particular scoring matrix. The scoring matrix, also called substitution matrix, is a set of values indicating the likelihood of a residue being substituted to another one. The score from the previous row is taken into account in the next row, and this continues until the end of the sequences. The best alignment has the highest score. In evolution, some mutations happen more frequently than others. Substitutions, where one residue mutates to another, happens more often than deletions or insertions. Deletions are when amino acids disappear, and insertions are when amino acids are added to the sequence. Because of this, it is important that the introduction of gaps in the alignment does not happen too often. To avoid this, there is a penalty for adding gaps in the alignment. The gap penalty is usually quite high for opening a gap, and lower for extending a gap, depending on the particular scoring matrix used (Xiong, 2006).

Pairwise sequence alignment is the comparison of two sequences, and multiple sequence alignment (MSA) is the alignment of three or more sequences. In theory the best MSA result would come from using an exhaustive dynamic programming where all possible alignments are investigated, as explained above, but this approach would not be efficient. This dynamic

programming approach works best for pairwise sequence alignment. Because of this, heuristic algorithms have been developed. They are fast and provide reliable alignments (Xiong, 2006).

2.1.1.2 Progressive alignment

The most used heuristic algorithm type is progressive alignment, which was first introduced in 1984 by Hogeweg and Hesper. In progressive alignment all sequences are pairwisely aligned to each other, to make a distance matrix based on either the similarity or identity score of the pairwise alignments. This distance matrix is in turn used to build a guide tree representing the relationship between the sequences. The most closely related sequences found in the tree is aligned using dynamic programming first. The next closest sequence or sequence pair is then aligned, depending on the branching in the guide tree. The first alignment is reduced to a consensus sequence and either aligned with the third closest sequence or later aligned to the consensus sequence of the closest related sequence pair. This continues from the closest to the most distantly related sequences following the guide tree, until all sequences are aligned, making the complete MSA (Hogeweg and Hesper, 1984; Xiong, 2006; Tran, Quoc-Nam, Wallinga, 2017). An overview of how the sequences are aligned can be seen in Figure 4.

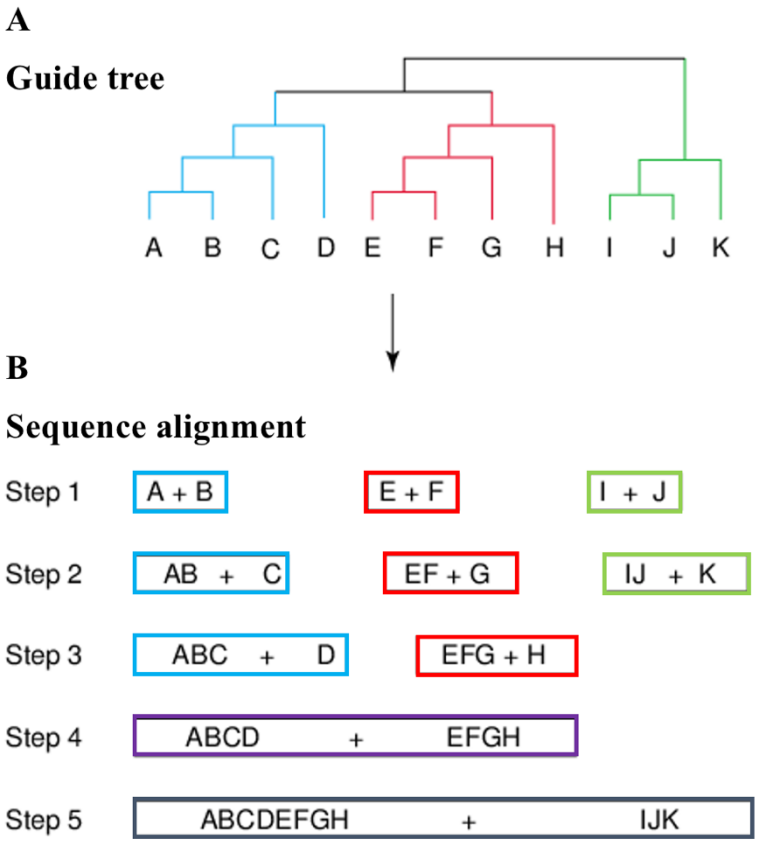

Figure 4 Progressive alignment. *A: all sequences have been pairwisely aligned to make a distance matrix which in turn is used to make the guide tree. The guide tree determines the order of how the sequences are aligned (B). Figure adapted from (Baldauf, 2003).*

In 1987 Feng & Doolittle introduced the rule “once a gap, always a gap”, which means that in the progressive alignment step, the gaps introduced in the closest related sequences are fixed and replaced by a neutral character (X) before moving on to the next alignment, to ensure that high-confidence gaps are not discarded to improve alignments of more distant sequences (Feng and Doolittle, 1987; Tran, Quoc-Nam, Wallinga, 2017).

2.1.1.3 T-Coffee

One type of progressive alignment method is T-Coffee (Notredame, Higgins and Heringa, 2000). The biggest difference between T-Coffee and other MSA methods is that rather than using a conventional scoring matrix to estimate scores when adding the sequences to the final MSA, the scores used are specific for every possible pair of residues in this combination of sequences and come from the extended library (Notredame, Higgins and Heringa, 2000).

In T-Coffee a library of pairwise alignments of all the query sequences is conducted. In the first version of T-Coffee the library computation was done by making two primary libraries of pairwise global and local alignments. The global library was made using Clustal W and the local library using Lalign (Needleman and Wunsch, 1970; Smith and Waterman, 1981; Huang and Miller, 1991; Thompson, Higgins and Gibson, 1994; Notredame, Higgins and Heringa, 2000).

The two pairwise primary libraries are then combined and followed by a heuristic process called library extension. Here a weight, also called consistency score, is assigned to each pair of residues, which reflects the degree to which those two residues align consistently with residues in the rest of the library. In the library extension, a triplet approach is used where each residue pair is checked against all the other sequences. The final weight for the residue pair will be the sum of all weights accumulated in examining the triplets, and it will reflect the information in the whole library, not just the information from that particular pairwise alignment (Notredame, Higgins and Heringa, 2000; Xiong, 2006).

In the current version of T-Coffee (v13.45.0.4846264) over 20 different third-party aligners are supported for making the library. The default aligner is called proba_pair, which is adapted from ProbCons (Do *et al.*, 2005; Di Tommaso *et al.*, 2011). Proba_pair uses a biphasic gap

penalty (Di Tommaso *et al.*, 2011). A biphasic gap penalty means that there are two sets of gap penalties, one for short gaps, and one for long gaps (Floden *et al.*, 2016).

After the library is done, the MSA is made using a standard progressive alignment algorithm. The query sequences are clustered to make the guide tree (Di Tommaso *et al.*, 2011). The sequences are aligned in the final MSA one by one by using the weights from the extended library. As the sequences are added to the MSA, they become fixed, and cannot be edited (Notredame, Higgins and Heringa, 2000; Taly *et al.*, 2011).

2.1.2 Sequence alignment with structural information

2.1.2.1 T-Coffee Expresso

T-Coffee Expresso is an alignment method that utilises both sequence and structure (Armougom *et al.*, 2006). Sequences are uploaded, and Expresso runs a Blast on the Protein Data Bank (PDB) to find corresponding structures to be used as templates. Structures with over 60% sequence identity to the entry sequences can be used. By default Expresso only considers X-ray structures, but it is possible to change the parameters (Taly *et al.*, 2011). When each entry sequence has been assigned a structure, Expresso starts a library computation. Pairwise structure-based alignments of the templates are first computed. The default structural aligner is SAP (Taylor, 2008), but other methods like TMalign and Mustang are also supported (Zhang and Skolnick, 2005; Konagurthu *et al.*, 2006). The two corresponding entry sequences are then aligned to their respective template structures. The induced pairwise alignment of the two entry sequences is then integrated in the library. The library will hold pairwise alignments of all the entry sequences guided by the structural information from the template structures. If no appropriate template structure is found, the default pairwise alignment method, proba_pair, is used instead of SAP for all alignments using this sequence, but its accuracy decreases significantly (Do *et al.*, 2005; Di Tommaso *et al.*, 2011; Taly *et al.*, 2011). When the library is finished, the MSA is assembled using the standard T-Coffee algorithm explained above (Armougom *et al.*, 2006).

2.1.2.2 PROMALS3D

Another alignment method which utilises both sequence and structure information is PROMALS3SD (Pei, Kim and Grishin, 2008). Sequences are uploaded, and PROMALS3D

aligns similar sequences, making groups of pre-aligned sequences, called clusters. One representative sequence is chosen from each cluster, and this sequence is now called a "target" sequence. The target sequences are then used in a PSI-BLAST search of the UniRef90 database for additional homologs (Suzek *et al.*, 2015). The UniRef100 database consists of all UniProtKnowledgebase records and some selected UniParc records, and the UniRef90 database is made from clustering UniRef100 sequences. Each cluster contains sequences with 90% sequence identity and UniRef90 contains one representative sequence from each of these clusters (Leinonen *et al.*, 2004; Suzek *et al.*, 2015; Bateman *et al.*, 2021).

Then two things happen simultaneously to find sequence and structure constraints which will be subsequently used to build the MSA. To find the sequence constraints the target sequences and the additional homologs are subjected to secondary structure prediction using PSIPRED (McGuffin, Bryson and Jones, 2000). The target sequences are used in a hidden Markov model (HMM) of profile-profile alignment with predicted secondary structures to find posterior probabilities of residue matches. The sequence constraints are these posterior probabilities.

To find the structural constraints, the results from the PSI-BLAST search against UniRef90 using the target sequences, are used in a new PSI-BLAST search against the SCOP40 database, to find protein domain sequences with known structures (Andreeva *et al.*, 2020). Redundancy is removed, and the homologs which are kept are called homolog3Ds. There can be more than one homolog3D for each target sequence if it contains several distinct domains with known structures. Alignments between two homolog3Ds and alignments between a target sequence and a homolog3D are then done to find structural pairwise residue match constraints. The structural aligners used are DaliLite, FAST and TMalign (Holm and Park, 2000; Zhang and Skolnick, 2005; Zhu and Weng, 2005). The sequence and structure constrains are combined to make a consistency-based multiple sequence alignment (Pei, Kim and Grishin, 2008).

2.1.3 CATH: Protein Structure Classification Database

CATH is a database of hierarchical classification of protein domain structures from the Protein Data Bank (PDB) (Orengo *et al.*, 1997). The classification is done by combination of automated and manual methods. In CATH there are four major levels: Class, Architecture, Topology and Homologous superfamily, from highest to lowest level. The Class level is the most general level, and the domains are classified on their α-helix and β-sheet content in the five classes:

mostly alpha, mostly beta, mixed alpha/beta, few secondary structures and special (Orengo *et al.*, 1997; Sillitoe *et al.*, 2021). In the next level, Architecture, the structures are grouped on their overall 3D shape of the secondary structures, but not how they are connected. At the T level, Topology, the structures are classified on fold. At this level the structures have both the same shape and connectivity of their secondary structures. At the last major level, the Homologous superfamily level, the structures that are grouped together here are thought to have evolved from a common ancestor, thus they are described as homologous (Orengo *et al.*, 1997; Sillitoe *et al.*, 2021).

There are five more levels of classification after the first four major levels, which are abbreviated to SOLID. Here the domains are classified on their level of sequence identity.
The S level is called Sequence family, the O level Orthologous Family, the L level "Like" domain and I level is called Identical domain. They have sequence identities with at least 35%, 60%, 95% and 100%, respectively. The last level is the D level, called unique domains or domain count, it is there to give each domain a unique code (Sillitoe *et al.*, 2021). A representative for each level is chosen, and this is usually either the first entry in the list which is also generally the best resolved crystal structure, or a commonly known example of the family. These representatives are chosen as a paradigm and will not be changed with new releases of CATH (Orengo *et al.*, 1997).

2.2 Methodology and workflow of the master project

The main goal of the project is to map the amino acid composition of the peripheral membrane binding interface of PH domains. This goal will be addressed using a bioinformatical approach, as stated in the introduction. The reason for this is to be able to study structure and sequence information of many PH domains extracted from public databases. To reach the goal the project work was divided into four parts.

In the first part relevant datasets of PH domain sequences and structures were collected and the sequences in the datasets were aligned using MSA. In the second part the datasets and MSAs were analyzed to collect information about motif 1 for canonical membrane binding, motif 2 for non-canonical membrane binding, the lengths of the three binding loops and the number of

secondary structures in the different PH domains. In the third part the amino acid composition was calculated for the different datasets. The amino acid composition was mapped for the whole domains, the binding loops and the sequences matching the two binding loops for the different datasets. The fourth part of the project addresses the level of conservation of the amino acids in Loop 1 which is used for both canonical and non-canonical binding to PIPs in the membrane. The overview of the workflow can be seen in figure 5.

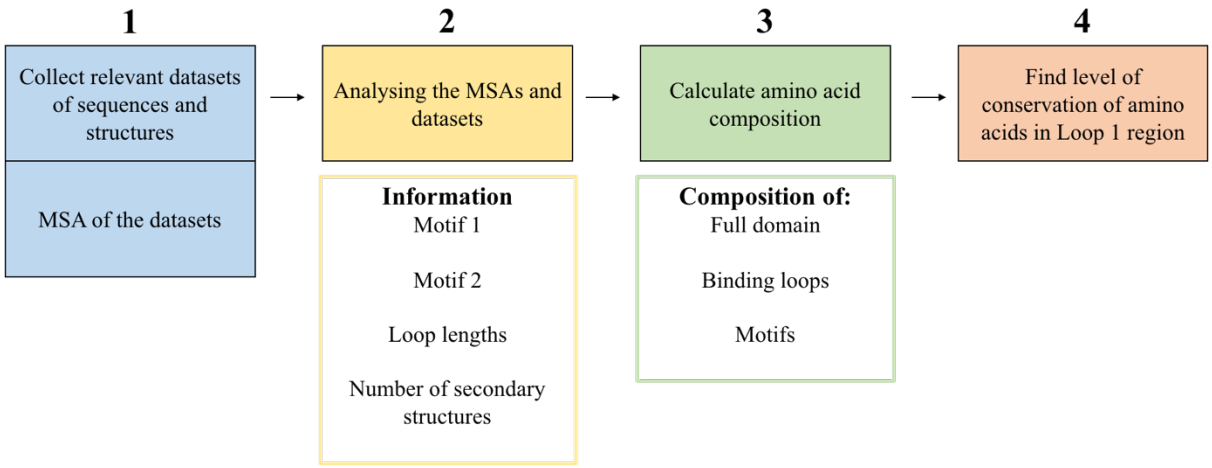

Figure 5 Workflow of the master project showing the four work parts .

2.2.1 Part 1 – Collecting relevant datasets and MSAs

2.2.1.1 Datasets derived from CATH sequence identity clusters

The main datasets used in this project were derived from the CATH superfamily 2.30.29.30 Pleckstrin homology domain (PH domain)/Phosphotyrosine binding domain (PTB domain) (Orengo et.al., 1997). The PDB files for this superfamily were downloaded from CATH-Plus Version 4.2, and there were 810 PDB files for the domains in this superfamily. The CATH cluster lists were downloaded from CATH-Plus version 4.3. The domains in CATH are clustered by sequence identities of 100, 95, 60 and 35 percent, and a representative from each cluster is chosen to form the S100, S95, S60 and S35 datasets, respectively. This means that the number of clusters is the same as the number of domain representative PDBs, and that the sequence identity between the domains in the S100, S95, S60 and S35 datasets is at most 100%, 95%, 60% or 35%, respectively. The sequences from each cluster were extracted using a python script that I wrote (Sequence_from_PDB_CATH.py, Jansen, 2021). Then the method of how

to sequence align the PH domains in the obtained datasets, when their sequence identity is so low, was addressed.

2.2.1.2 Alignment without structural information

To determine the best alignment method, eight PH domains, dataset I8, were randomly chosen for alignment from the CATH superfamily 2.30.29.20 (Orengo *et al.*, 1997). The I8 PDB files were opened in PyMOL (v2.4.2) and the sequences exported to FASTA, and were subsequently aligned using Clustal Omega (v1.2.4), T-Coffee (v11.00) and MUSCLE (v3.8.31) (Pearson and Lipman, 1988; Notredame, Higgins and Heringa, 2000; Edgar, 2004; Pettersen *et al.*, 2004; Waterhouse *et al.*, 2009; Sievers *et al.*, 2011; Schrödinger, 2015; Madeira *et al.*, 2019). Clustal Omega and MUSCLE alignments were made using the MSA tool on EMBL-EBI online (Edgar, 2004; Sievers *et al.*, 2011; Madeira *et al.*, 2019). T-Coffee alignment was made using simple MSA from the T-Coffee web tool (Notredame, Higgins and Heringa, 2000).

2.2.1.3 Alignment with structural information

Two structural alignment tools, were also used to determine the best suited method (Armougom *et al.*, 2006; Pei, Kim and Grishin, 2008). The alignments were made using T-Coffee Expresso (v11.00) MSA from T-Coffee web tool, and PROMALS3D from prodata.swmed web tool (Pei, Kim and Grishin, 2008). T-Coffee Expresso provides three different advanced options, SAP, TMalign and Mustang, for pairwise structural library computation, which were all tested (Taylor and Orengo, 1989; Zhang and Skolnick, 2005; Konagurthu *et al.*, 2006). Jalview (v2.11.1.3), Chimera (v1.15.0) and a custom visualisation script called colorMSAwithSSE (v1.0.0) were used for visualization and figure making for both alignments with and without structural information (Pettersen *et al.*, 2004; Waterhouse *et al.*, 2009; Tubiana, 2021). Both T-Coffee Expresso (v11.00 and v13.41.123.92238f3) with TMalign and PROMALS3D were used to align the datasets obtained from CATH.

2.2.2 Part 2 – Analyzing MSAs and datasets

The second part of the project revolved around analyzing multiple sequence alignments and the datasets obtained in part 1. Not all PH domains bind to membranes, and the aim of this project is to map the peripheral membrane binding interface of PH domains. Thus, the PH domains which bind to membranes have to be found. PH domains bind to PIPs in membranes using canonical or non-canonical membrane binding, and their sequences are characterized by

binding motifs present in binding loop 1 (Figure 3). PH domains from CATH cluster S95 were used for further analysis instead of S100 to avoid results from very similar PH domains to be weighed too much.

2.2.2.1 Motif for canonical membrane binding (M1), KX(K/R)XR

The S95 sequences from CATH obtained in part 1 were subjected to MEME Suites FIMO (v5.3.3) to find PH domains with sequences matching the canonical binding motif, KX(K/R)XR (Lemmon and Ferguson, 2000; Grant, Bailey and Noble, 2011). FIMO is an online tool to scan provided sequences for matches to a provided motif, and it was used with default parameters where the *P* value was less than 1e-4. The domains with the detected motifs were aligned using PROMALS3D (Pei, Kim and Grishin, 2008). The alignment showed that many of the sequences matching the motifs were not structurally positioned around a loop, where this binding motif should be positioned (Lemmon and Ferguson, 2000; Lemmon, 2008). The alignment was therefore cut at the beginning of β1 and at the end of β2, and the excerpt was ungapped using a python script I wrote (Sequence_from_MSA.py, Jansen, 2021). The sequences in this range were subjected to FIMO again to only find sequences matching M1 with the correct position relative to the secondary structures.

2.2.2.2 Motif for non-canonical membrane binding (M2), (K/R)X(W/Y/F)

Because many sequences matching M1 were found outside of the β1-L1-β2 range and the sequences had to be subjected to FIMO twice, the methodology was changed to find PH domains with a sequence matching the non-canonical binding motif. The S95 alignment from part 1 was cut in the middle of β1 and middle of β2, and subsequently ungapped. The S95 sequence excerpts were both subjected to MEME Suites FIMO and looked at manually to detect the non-canonical binding motif, (K/R)X(W/Y/F) (Grant, Bailey and Noble, 2011). FIMO was used with default parameters where the *P* value was less than 1e-4. The different sequence versions of M2 in the found PH domains were counted (KXW, KXY, KXF, RXW, RXY, RXF).

2.2.2.3 Lengths of the three binding loops L1, L2 and L3

The two types of membrane-binding uses the loops connecting β1/β2 and β3/β4 for canonical binding or β1/β2 and β5/β6 for non-canonical binding. It has been proposed that depending on the type of binding, the loops are more prominent and longer if they are used in binding, and

shorter if not (Hurley, 2006; Naughton, Kalli and Sansom, 2018). Because of this I wanted to look at the loop lengths in the different datasets obtained throughout this project. The lengths of the three binding loops were first looked at structurally and figures showing the three loops in the structures were made using PyMOL (v2.4.2) (Schrödinger, 2015). Datasets were also aligned using either PROMALS3D or T-Coffee Expresso (Armougom *et al.*, 2006; Pei, Kim and Grishin, 2008). The loop sections of the MSAs were cut out, ungapped and the length was calculated (find_loop_lengths.py, Jansen, 2021). Table 1 shows which alignment and the cut positions used for each dataset. Quartile calculations were done on the obtained loop length dataset, and a box-plot was made.

2.2.2.4 Quantifying variations in secondary structures in PH domains

Because of the diversity of the PH domains I also wanted to quantify the number of secondary structures (α and β) in the datasets. Most PH domains have seven β-strands and 1 α-helix, but I have not seen any statistics on how many have these numbers of secondary structures (Timm *et al.*, 1994; Cho and Stahelin, 2005; Lemmon, 2008). Therefore, the number of α-helices and β-strands in the datasets were counted and pie charts were made to visualise the obtained data (number_SS_DSSP.py, Jansen, 2021).

2.2.3 Part 3 - Amino acid composition of the PH domain

The third part of the project comprises the calculation and mapping of amino acids in the X parts of the sequence motifs, the three binding loops, and the full PH domain. Amino acid composition calculations were done using the sequences, MSAs corresponding to the datasets and python scripts (AAcomp_FullPH_from_Fasta.py, AAComp_loops.py, AAcomp_M1.py, AAComp_M2.py, Jansen, 2021). Column diagrams were made to display the percentages of each amino acids, and percentages of the different property groups of the AAs. The property groups are: hydrophobic (L, I, C, M, W, Y, F), aromatic (W,Y, F), negative (D, E), non-polar (V, A,G, P), polar (S, N, Q, T) and positive (H, K, R). The hydrophobic amino acids are grouped according to the Wimley and White hydrophobicity scale (1996).

2.2.4 Part 4 – Conserved amino acids in β1-L1-β2 region

The fourth part of this project addresses the level of conservation of specific amino acids in specific structural positions in the β1 - Loop 1 - β2 part of the PH domain structure. This region is a part of the peripheral membrane binding interface for both canonical and non-canonical membrane binding. The conservation of the amino acids L, (K/R), G, (K/R), (W/Y/F), (K/R) and (K/R) in their respective positions were investigated manually and by using MSAs and a python script (Conservation_AA.py, Jansen, 2021). L in the middle of β1, K/R in the end of β1, G in the beginning of Loop 1, K/R in the end of Loop 1, W/Y/F in the beginning of β2, K/ in the beginning of β2 and K/R in the middle of β2. Which MSAs that were used and the cut ranges can be seen in Table 1.

Table 1 Overview of MSA methods used for which dataset and corresponding alignment cut positions. *The datasets are presented in the results section on page 30 and 40, and an overview of how the datasets are related can be found in Figure 1A in Appendix I.*

Alignment/dataset	MSA method	What	Cut position
S95	PROMALS3D	Mid β1-mid β2	[135, 160]
S60*	Expresso	Loop 1	[196, 290]
		Loop 2	[333, 444]
		Loop 3	[494, 556]
		L	[186, 195]
		K/R	[189, 198]
		G	[194, 203]
		K/R	[267, 292]
		W/Y/F	[284, 302]
		K/R	[291, 302]
		K/R	[302, 307]
F56	PROMALS3D	β1 - Loop 1 - β2	[50, 130]
1M26	PROMALS3D	Loop 1	[50, 70]
		Loop 2	[95, 108]
		Loop 3	[123, 142]
2M59	Expresso	Loop 1	[89, 133]
		Loop 2	[171, 221]
		Loop 3	[250, 300]
		L	[83, 89]
		K/R	[85, 92]
		G	[88, 131]
		K/R	[120, 134]

		W/Y/F	[130, 140]
		K/R	[133, 138]
		K/R	[138, 141]
M1and2	Expresso	Loop 1	[58, 73]
		Loop 2	[103, 126]
		Loop 3	[151, 172]
OM1	Expresso	Loop 1	impossible
		Loop 2	impossible
		Loop 3	impossible
OM2	Expresso	Loop 1	[62, 105]
		Loop 2	[137, 174
		Loop 3	[213, 237]

2.2.5 Non-PH domains in the CATH 2.30.29.30 superfamily

At the end of the project it was discovered that the CATH superfamily 2.30.29.30 not only contained true PH domains, but also PH domain-like domains as Phosphotyrosine binding domains (PTB domains), Ran binding domains (RanBD), Enabled/ vasodilator-stimulated phosphoproteins (VASP) homology domain (EVH domains) and Decapping protein (Dcp), and other not yet determined domains with the PH domain fold. All the PTB domains in this CATH superfamily were found by finding one PTB domain in RSCB PDB, and in the annotation tab the Pfam accession number was found. By clicking the accession number one can find all PDB IDs in this protein family (date 23.08.21). The PDB IDs were checked against the CATH IDs in CATH superfamily 2.30.29.30. The same was done for other domains with the PH domain fold, RanBD, EVH and Dcp (date 22.06.21) (Lemmon and Ferguson, 2000; She *et al.*, 2004). The CATH IDs of the non-PH domains in this superfamily were checked against the other datasets obtained throughout the project. This was done using python scripts I wrote (non-PH_from_Pfam.py, non-PH_in_cluster_datasets.py, non-PH_in_motif_datasets.py, Jansen, 2021).

2.2.6 Python programming software

The programming language used throughout this project was Python (v3.8.5) (Van Rossum, G. & Drake, F.L., 2009). The Python package which was used was Biopython (v1.78) (Cock

et al., 2009) with the modules AlignIO, PDBparser. DSSP (v3.0.0) was used for secondary structure assignment (Kabsch and Sander, 1983; Touw *et al.*, 2015). All used scripts can be found on the GitHub project: https://github.com/KamillaOJ/Mapping_AA (Jansen, 2021).

3. *Results*

3.1 Part 1 – Datasets and MSA

3.1.1 Datasets directly from CATH superfamily 2.30.29.30

The main datasets used in this project were derived from the CATH superfamily 2.30.29.30 Pleckstrin homology domain (PH domain)/Phosphotyrosine binding domain (PTB domain) (Orengo et.al., 1997). The domains in CATH are clustered by sequence identities of 100, 95, 60 and 35 percent, and a representative from each cluster is chosen. This means that the number of PH domain structures in the derived datasets is the same as the number of clusters at this sequence identity level. Figure A1 in Appendix I shows how all datasets in this project are related.

- **S100:** Dataset with 328 domain structures with at most 100% sequence identity.
- **S95:** Dataset with 211 domain structures with at most 95% sequence identity.
- **S60:** Dataset with 162 domain structures with at most 60% sequence identity.
- **S60*:** T-Coffee Expresso has an upper sequence limit of 150 sequences. Because of this, 12 domains had to be removed from S60 to be able to align it by T-Coffee Expresso. 150 domain structures are in this dataset.
- **S35:** Dataset with 137 domain structures with at most 35% sequence identity.
- **I8:** Eight initial (Initial 8) domains used for determining the best multiple alignment method.

3.1.2 Evaluating the alignment methods

Five different alignment methods were tested using eight PH domains (dataset I8). Three sequence aligners were tested: Clustal Omega, T-Coffee and MUSCLE, and two aligners using both sequential and structural information: T-Coffee Expresso and PROMALS3D. The sequence identity of PH domains is quite low, around 30%, but generally structure is better conserved than sequence (Cho and Stahelin, 2005; Illergård, Ardell and Elofsson, 2009). Because of this, we decided to test both types of alignment methods.

3.1.2.1 Alignment without structural information

The result from the three initial alignments with the I8 dataset, using Clustal Omega, T-Coffee and MUSCLE can be seen in Figure 6. The lengths of the alignments are 231, 249 and 190 respectively. Thus, the MUSCLE alignment is much shorter than the other two, because there are less gaps introduced. The yellow β-strands are not well aligned in any of the alignments. The β-strands are better aligned in the MUSCLE alignment compared to the other two, where one can see five of the seven β-strands. In the Clustal Omega and T-Coffee alignments one can see two and three aligned β-strands out of the seven, respectively. The long red C-terminal α-helices are aligned in the T-Coffee and MUSCLE alignments, whereas the alignment of this α-helix, for half of the sequences, is shifted towards the C-terminal in the Clustal Omega alignment. Overall none of the alignments show a good conservation of secondary structure between the domains.

Figure 6 Multiple sequence alignment of PH domains in dataset I8 using sequence aligners Clustal Omega, T-Coffee and MUSCLE. *α-helices are red, β-sheets are yellow, and loops are green. ColorMSAwithSSE was used for visualization (Tubiana, 2021).*

3.1.2.2 Alignment with structural information

T-Coffee Expresso structural alignment method was used because of its ability to utilize both sequential and structural information to help the alignment. All three advanced options for pairwise structural methods for library computation, SAP, TMalign and Mustang were used, and the alignment lengths were 230, 236 and 230, respectively. There are more gaps in the TMalign alignment than when using the SAP or Mustang option. The three resulting alignments can be seen in Figure 7. T-Coffee provides an overall alignment score, ranging from 0-100, where a score lower than 50 is considered poor. The scores were 87 for SAP and TMalign, while Mustang had the score of 86. Thus, the alignments using SAP and TMalign pairwise structural methods are equally good, and Mustang almost as good. Expresso gave the same alignment result when chosen structures were uploaded and without providing the structures. As explained in the methodology, when structures are not provided, Expresso runs by default a PDB blast to find X-ray structures with over 60% sequence identity, to find suitable templates for the provided sequences (Di Tommaso *et al.*, 2011). One can see in the MSAs that all seven β-strands and the C-terminal α-helix are well aligned in all three alignments meaning that all three alignments showed high conservation of the secondary structures between the domains (Figure 7).

PROMALS3D also utilizes both sequence and structure to carry out the MSA, and the alignment made with this method can also be seen in Figure 7. The length of the alignment is 221, meaning that it contains less gaps than all three T-Coffee Expresso alignments. PROMALS3D alignment also showed high conservation of both red α-helices and yellow β-sheets (Figure 7), where one can see all seven β-strands and the C-terminal α-helix aligned very well to each other. PROMALS3D does not provide an alignment score.

Due to the result of T-Coffee Expresso and PROMALS3D were equally good, they were both used for all alignments throughout the project and were evaluated towards each other for each individual alignments of the datasets. The TMalign option used as structural aligner with T-Coffee Expresso was used exclusively. The criteria for choosing the optimal alignment method was how well the secondary structures were aligned. Which MSA method was used for which dataset can be seen in Table 1, in the materials and methods section.

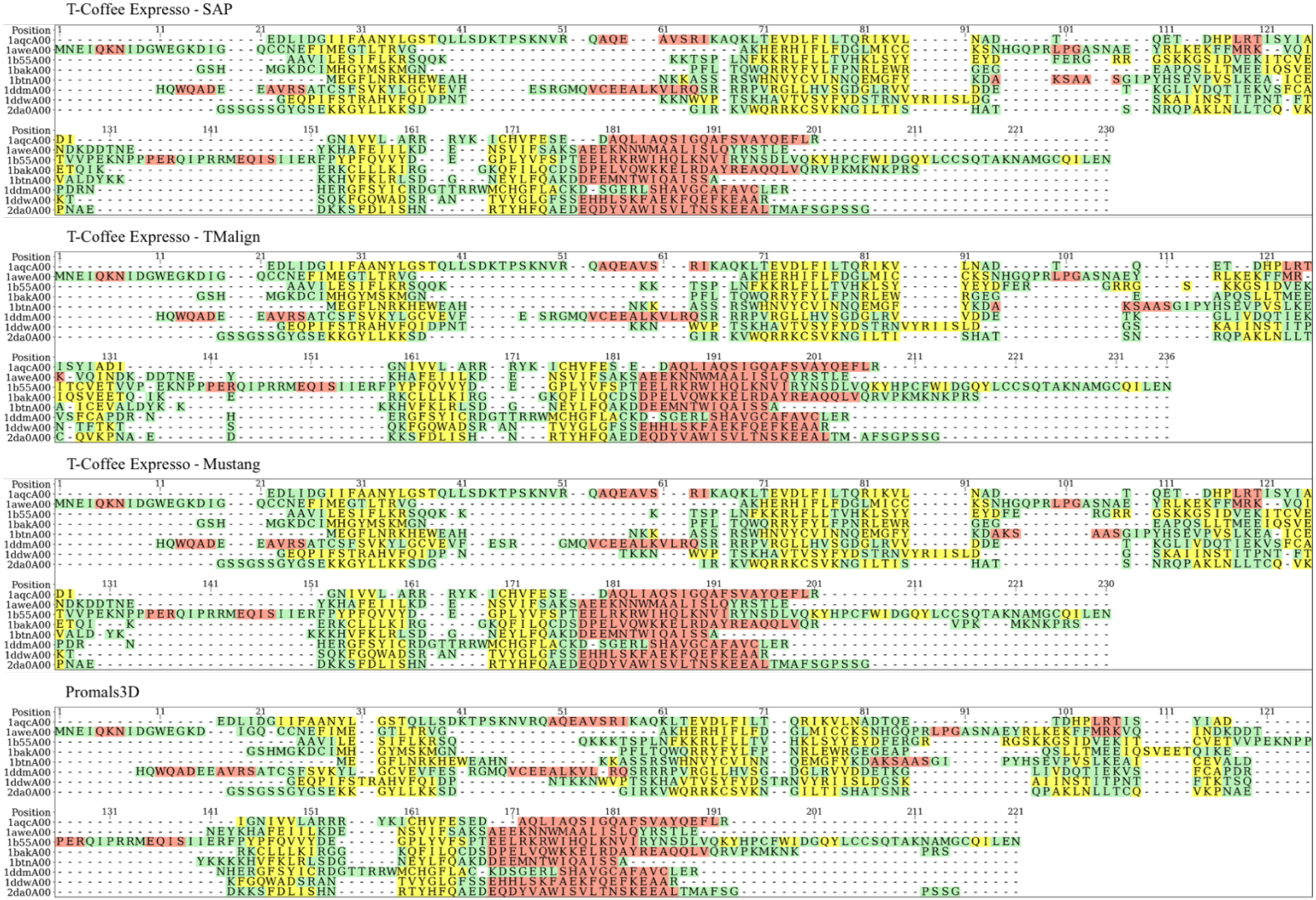

Figure 7 Multiple structural sequence alignment of PH domains in dataset I8 using T-Coffee Expresso with pairwise structural method for library computation set to SAP, TMalign and Mustang, and PROMALS3D. *α-helices are red, β-sheets are yellow, and loops are green. ColorMSAwithSSE was used for visualization (Tubiana, 2021).*

3.2 Part 2 – Analyzing the datasets and MSAs

As written in the introduction, there are two binding motifs which can be present in the sequence in L1 connecting β1/β2. These motifs are M1 and M2 for canonical and non-canonical membrane binding, respectively. M1 is KX(n)(K/R)XR and M2 is (K/R)X(W/Y/F) (Lemmon and Ferguson, 2000; Naughton, Kalli and Sansom, 2018). The structural position of the sequence motifs can be seen in Figure 3 in the introduction. We wanted to see how many of the 211 PH domains in CATH S95 contained sequences matching these motifs. We also wanted to find the lengths of the matching sequences, and the amino acids of the Xs in the motifs.

3.2.1 Search for binding motifs

3.2.1.1 Canonical membrane binding, motif M1

Lemmon (2007) suggests that ten percent of all PH domains binds to PIPs in a canonical way using the sequence matching M1 (KX(n)(K/R)XR) (Lemmon and Ferguson, 2000; Lemmon, 2008). Using the MEME Suite FIMO to identify sequence patterns in the β1 – L1 - β2 region, 35 domains were found to have at least one sequence matching M1. A new dataset called F35 containing these 35 PH domains was made and aligned for further analysis of the sequences. The alignment was made to see which of the FIMO detected sequences had the correct structural position for B1 binding. An excerpt of this alignment can be seen in Figure 8.

After analyzing the MSA the result was that 26 out of the 35 domains had a sequence matching M1 with the correct position relative to the secondary structures. This correct position is with K in the end of β1, (K/R) in the beginning of β2 and R in the middle of β2. This means that there are 26 domains with a sequence with a perfect match to all aspects of motif 1 for canonical binding out of the 211 domains in S95. This gives a percentage of 12% domains with a sequence fitting the canonical binding motif in the CATH S95 cluster. The resulting dataset consisting of these 26 domains is called 1M26. The length of the sequences matching the motif ranges from seven to 28 AAs. The most common length is 12 AA, and 15 of the 26 sequences have this length. The average length of the motif is 12,1 AA. The sequence motifs, lengths and sequential position corresponding to these 26 PH domains can be found in Table 2.

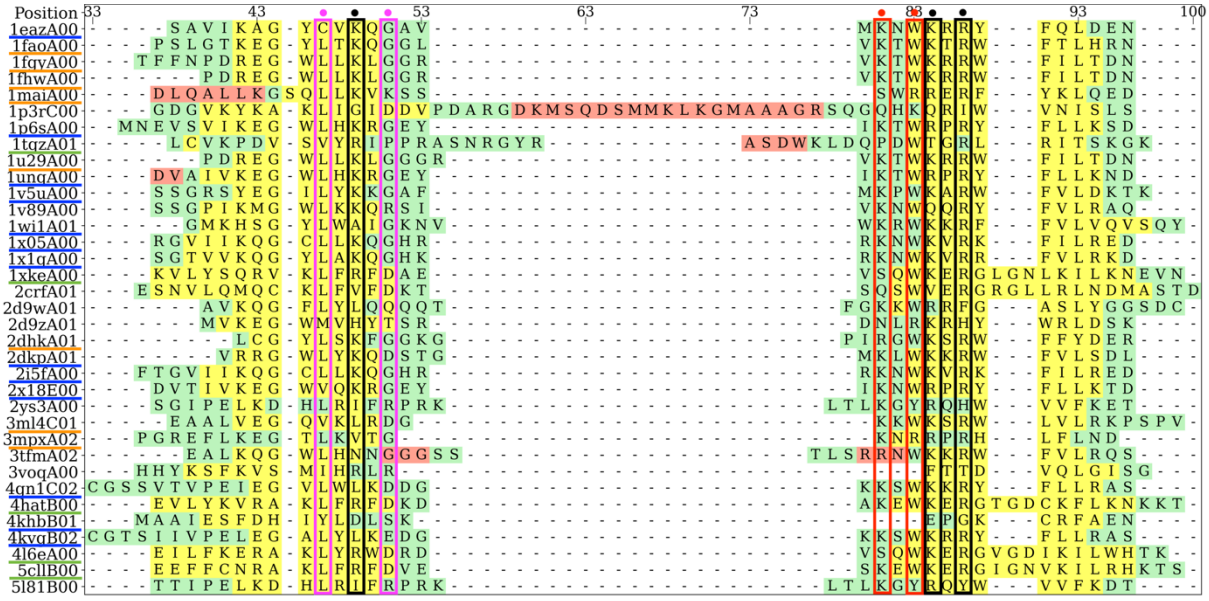

Figure 8 β1 – Loop 1 - β2 excerpt from MSA of dataset F35*. α-helices are red, β-strands are yellow and loops are green. The black boxes show the conservation of the K, (K/R) and R in the motif (KX(n)(K/R)XR) for canonical membrane binding discovered by Lemmon and Ferguson (2000). The red boxes show conserved amino acids (K/R), (W/Y/F) in non-canonical membrane binding discovered by Naughton, Kalli and Sansom (2018). The pink boxes show conserved amino acids (L, G) found in this alignment. Underlined domains are the domains in dataset 1M26. Blue underlined domains have M1 matching sequence detected by FIMO, orange underlined domains have sequences matching M1 found by analyzing this alignment, green underlined domains have sequences matching M1 with an R instead of K found by analyzing this MSA.*

Table 2 Overview of the PH domains in dataset 1M26 with sequences matching the canonical binding motif (KX(n)(K/R)XR), detected by FIMO and by analyzing MSA of dataset F35. *The table shows the CATH ID of the PH domains, the sequence matching M1, the length of the sequence and the sequential position in the domain. Blue: sequences with correct match to motif 1 found by FIMO. Orange: sequence matches to motif 1 found by analyzing the MSA of dataset F35. Green: sequences following the correct motif-pattern but with the K substituted for an R.*

CATH ID	Sequences fitting M1 (KX(n)(K/R)XR), found by FIMO	Length	Position
1eazA00	KQGAVMKNWKRR	12	11 → 22
1p6sA00	KRGEYIKTWRPR	12	14 → 25
1unqA00	KRGEYIKTWRPR	12	15 → 26
1wi1A01	KNVWKRWKKR	10	13 → 22
1x05A00	KQGHRRKNWKVR	12	26 → 37
1x1gA00	KQGHKRKNWKVR	12	26 → 37
2dkpA01	KQDSTGMKLWKKR	13	8 → 20
2i5fA00	KQGHRRKNWKVR	12	13 → 24
2x18E00	KRGEYIKKNWRPR	13	12 → 24

4gn1C02	KDDGKKSWKKR	11	42 → 52
4kvgB02	KEDGKKSWKRR	11	42 → 52
1v5uA00	KGAFMKPWKAR KKGAFMKPWKAR	11 12	17 → 27 16 → 27
CATH ID	**Sequences fitting M1 (KX(n)(K/R)XR), found by analyzing MSA**	**Length**	**Position**
4khbB01	KEPGKCR	7	16 → 22
1faoA00	KQGGLVKTWKTR	12	12 → 23
1fgyA00	KLGGRVKTWKRR	12	13 → 24
1fhwA00	KLGGRVKTWKRR	12	9 → 20
1maiA00	KVKSSSWRRER	7	21 → 29
1u29A00	KLGGGRVKTWKRR	13	9 → 21
2dhkA01	KFGGKGPIRGWKSR	14	7 → 20
3ml4C01	KLRDGKKWKSR	11	10 → 20
3mpxA02	KVTGKNRRPR	10	18 → 27
CATH ID	**Sequences almost fitting M1 (KX(n)(K/R)XR), with starting R, found by analyzing MSA**	**Length**	**Position**
1tqzA01	RIPPRASNRGYRASDWKLDQPDWTGRLR	28	11 → 38
1xkeA00	RFDAEVSQWKER	12	19 → 30
4hatB00	RFDKDAKEWKER	12	16 → 27
416eA00	RWDRDVSQWKER	12	18 → 29
5cllB00	RFDVESKEWKER	12	35 → 46

13 of the 26 domains had a sequence detected by FIMO which fit the description of M1 with the conserved lysine in the end of β1 and the conserved arginine in the middle of β2. One of the domains (1v5uA00) had two sequences that fit the description for M1, because it had lysine as the two first amino acids in the sequence. The 4khbB01 domain seems to have a β1 that is only 3 AAs long, but the sequence seems to be in the correct position relative to the M1 secondary structure position. These can be seen in Table 2, in the blue-highlighted rows.

Eight of the 26 domains had a sequence matching M1 detected by MEME Suite FIMO, but by looking at the alignment (Figure 8) they had the wrong placement relative to the secondary structures to be a perfect motif match for canonical PIP membrane binding. Instead they had a sequence matching the motif with the correct position that was not detected by FIMO in any of the attempts, but found by manually looking at the MSA. These correctly matched sequences can be seen in Table 2, in the orange highlighted rows.

Five out of the 26 domains had FIMO detected sequences, but the sequences did not have the correct starting position for canonical membrane binding where the motif starts in the end of β1. These domain sequences had the conserved (K/R)XR ending. These domains were found to have an arginine instead of lysine in the end of β1 which is aligned with the lysines from the other domains (Figure 8). These domains with their sequences can be seen in Table 2, in the green-highlighted rows.

Nine of the dataset F35 domains had sequences detected by FIMO which did not match the structural patterns known for M1. Three of the sequences had the starting lysine positioned at the beginning of β1 instead of at the end, making it not fit the structural M1 description. These domains are 2d9wA01, 2d9zA01 and 3tfmA01. One domain (2crfA01) had longer β–strands than the other domains in this dataset, and a shorter Loop 1. The lysine is positioned in the middle of β1 instead of in the end. The arginine is in the middle of β2, but since the β2 is longer than in the other domains, it leads to the arginine not being aligned with the other domains arginines in the MSA.

Five of the F35 domains all had varying oddities related to their detected sequence. In 1p3rC00 there is an α-helix in the loop area. The sequence in 1v89A00 is only positioned in the β1 and the beginning of the Loop 1. There is a presence of a sequence resembling the motif in 1v89A00 that is aligned with the other motifs (Figure 8), but it has a glutamine in the K/R position. It looks like the sequence of 2ys3A00 is not aligned correctly in the MSA in regard to the other domains, as it does not have its β1 aligned with the others. The detected sequence does not begin in a β-strand and end in another β-strand. For both 3voqA00 and 5l18B00 the FIMO detected sequence starts in β1 and ends in the beginning of Loop 1. There are no parts of the sequences present in β2.

3.2.1.2 Non-canonical membrane binding, motif M2

Sansom (2018) reports that some PH domains have another binding motif for binding PIPs in membranes than the canonical binding motif, M1. This non-canonical type of binding has the binding motif of (K/R)X(W/Y/F), and in the structure the aromatic amino acid should be positioned in the end of L1 or beginning of β2. The β1-Loop 1-β2 region was cut out of the S95 MSA and the sequences in this part were subjected to MEME Suite FIMO. The outcome from FIMO was only one PH domain with a sequence matching this non-canonical binding motif (M2). Because of this, another approach was used. The β1-L1-β2 sequence excerpt from the MSA was looked at manually, and 59 PH domains out of the 211 S95 PH domains were found to contain a sequence matching the non-canonical PIP membrane binding motif (M2). Thus, 28% of the 211 PH domains in the S95 dataset have a sequence fitting this motif description. The dataset containing these 59 PH domains is called 2M59, and they can be seen together with their motif sequence in Table 3. The position of the motif in the structure can be seen in Figure 3 in the introduction and the position of the motif sequence in the F35 MSA can be seen marked by red boxes in Figure 8.

Table 3 Overview of the 2M59 PH domains with their motif sequence for non-canonical membrane binding.

CATH ID	Motif sequence KXW	CATH ID	Motif sequence KXW	CATH ID	Motif sequence RXW
1ddvA00	KNW	2i5fA00	KNW	1btnA00	RSW
1ddwA00	KNW	2p0hA00	KNW	1droA00	RSW
1eazA00	KNW	2rloA00	KEW	1upqA00	RLW
1egxA00	KRW	2vszB02	KFW	1wjmA00	RSW
1faoA00	KTW	2x18E00	KNW	2cy5A00	RVW
1fgyA00	KTW	2yryA00	KQW	2d9vA01	RRW
1fhwA00	KTW	3aj4A00	KRW	2dhkA01	RGW
1p6sA00	KTW	3hk0B02	KSW	2p8vA00	RNW
1qc6A00	KKW	3ml4C01	KKW	2q13A02	RTW
1u29A00	KTW	3oanA00	KGY	3a8pB01	RKW
1unqA00	KTW	3pp2A00	KHW	3tfmA02	RNW
1v5uA00	KPW	4gn1C02	KSW	4k2pD01	RKW
1v89A00	KNW	4hatB00	KEW	4wsfA00	RQW
1wgqA00	KPW	4k81A02	KSW		
1wi1A01	KRW	4kvgB02	KSW		RXY

1x05A00	KNW	5cllB00	KEW	1x1fA00	REY
1x1gA00	KNW				
2d9wA01	KKW	**KXY**		**RXF**	
2d9yA00	KQW	1dynA00	KEY	2codA01	RMF
2da0A00	KVW	2ys1A00	KEY		
2dhiA00	KRW	3qbvB02	KLY		
2dkpA01	KLW	4f7hA00	KGY		
2dn6A00	KNW	5l81B00	KGY		

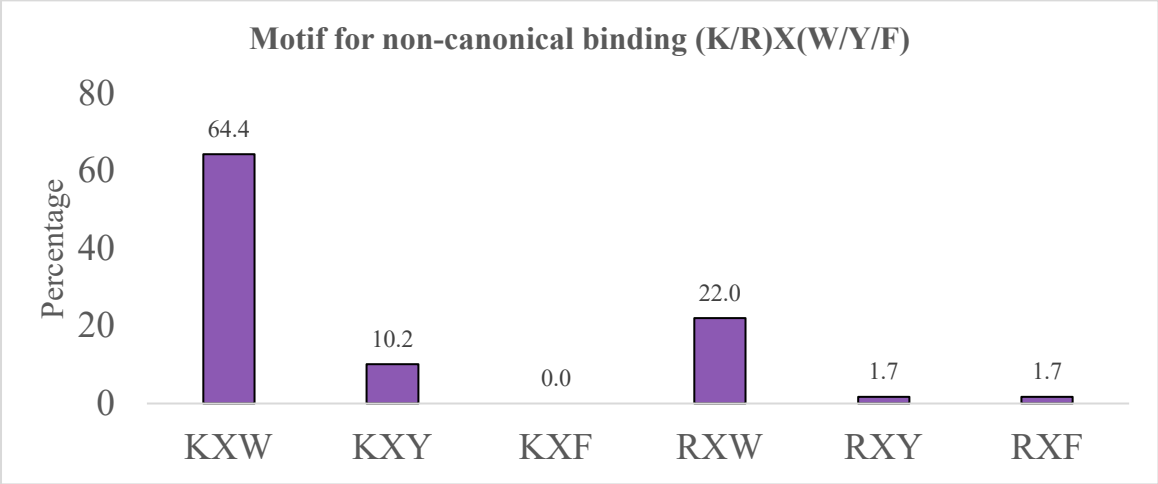

Figure 9 Percentage of the six different non-canonical binding motif variations.

The sequences matching the six different versions of the non-canonical binding motif were counted, and the result was that the most common motif version was KXW. 38 of the 59 PH domains in CATH S95 had this variation. The percentage of occurrences of the six motif variations can be seen on the plot in Figure 9. The second highest occurrence was RXW which was present in 13 of the domains. This makes tryptophan the ending aromatic residue in the motif sequences in 86% of the 2M59 domains. Six domains had the KXY motif version and there was only one occurrence of RXY and RXF. No KXF motif version was found.

3.2.1.3 Summary of resulting datasets

The analyses of the datasets and multiple sequence alignment reported above led to the formation of subsets of domain structures. We summarize here all subsets created with their name and origin. Figure A1 in Appendix I shows how all datasets are related.

Datasets concerning M1 for canonical membrane binding:

- **F56:** The S95 dataset was subjected to MEME Suite FIMO for detection of a canonical binding motif (KX(n)(K/R)XR). These are all the 56 domains with a sequence matching M1 at any position in the full sequence. The name means 56 domains found by FIMO (F56).

- **F35:** The F56 dataset was aligned using PROMALS3D and a β1/β2 excerpt was subjected to MEME Suite FIMO for detection of a canonical binding motif in this range. 35 domains were found to contain a sequence matching M1. The name means 35 domains found by FIMO (F35).

- **1M26:** The MSA of the F35 dataset was analyzed, resulting in this dataset which consists of 26 domains containing a sequence matching the binding motif for canonical binding (M1). These 26 domains have 95% sequence identity at most because they originated from the S95 dataset. The name means 26 domains with sequence matching M1 (1M26).

Datasets concerning M2 for non-canonical membrane binding:

- **2M59:** The S95 dataset was aligned and the β1-L1-β2 region was cut out. This region was looked at manually to find the motif for non-canonical binding ((K/R)X(W/Y/F), which gave this dataset with 59 domains containing this binding motif. The name means 59 domains with sequence matching M2 (2M59).

Datasets concerning both M1 for canonical membrane binding and M2 for non-canonical membrane binding:

- **M1and2:** This dataset consists of the 20 domains containing a sequence matching both the canonical and the non-canonical binding motif. It originates from both the 1M26 and the 2M59 datasets. The name means domains with sequence matching both M1 and M2 (M1and2)

- **OM1:** These six domains are the PH domains in 1M26 which only contain a sequence matching the canonical binding motif, and not the non-canonical motif. 1maiA00,

1tqzA01, 1xkeA00, 3mpxA02, 4khbB01 and 4l6eA00. The name means only M1 (OM1)

- **OM2:** 39 domains which only contain a sequence matching the the non-canonical binding domain, and not the canonical motif, originating from 2M59. The name means only M2 (OM2).

3.2.2 Lengths of the three membrane binding loops linking β1/β2, β3/β4 and β5/β6

PH domains bind to the membrane using Loop 1 between β1 and β2, and Loop 2 between β3 and β4 for canonical binding. For non-canonical binding they use Loop 1 and Loop 3 which links together β5 and β6 (Lemmon and Ferguson, 2000; Lemmon, 2008; Naughton, Kalli and Sansom, 2016, 2018). When a PH domain does not bind in the canonical way the L1 should be shorter than in a PH domain that does bind in the canonical way. When there is non-canonical binding between the PH domain and a membrane the L3 should be long (Hurley, 2006; Naughton, Kalli and Sansom, 2018). Because of this we wanted to find the lengths of the three binding loops in and see if there is a correlation between length and presence or absence of a binding motif. The lengths of the three loops were investigated in six datasets: S60*, 1M26, 2M59, M1and2, OM1 and in OM2.

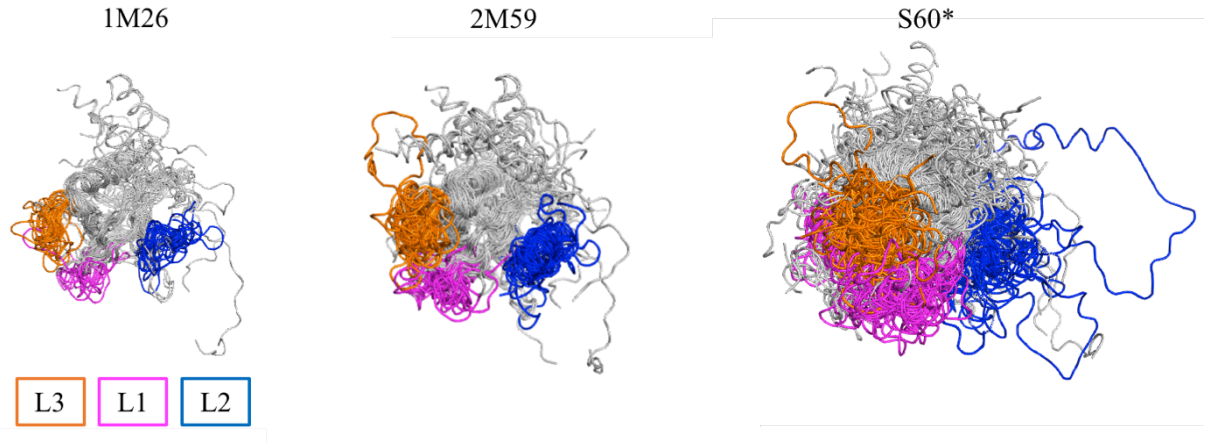

Figure 10 Visual representation of the three binding loops in the three datasets 1M26, 2M59 and S60*. *Loop 1 is magenta, Loop 2 is blue, and Loop 3 is orange, which links β1/β2, β3/β4 and β5/β6, respectively. The figures were made using PyMOL (v2.4.2) (Schrödinger, 2015).*

I had three hypotheses based on the literature recited above. The first was that L1 is shortest in S60*, and longest in 1M26 and OM1. The second was that datasets 1M26 and OM1 would have the longest L2 and shortest L3. The third hypothesis was that 2M59 and OM2 had the longest L2 and shortest L3. Figure 10 shows the three loops in the PH domain structure in datasets 1M26, 2M59 and S60*. In this figure one can see that the loops vary in length. Some look short, while others look much longer, especially in Loop 2 and 3.

The loop lengths for each loop were found using the MSAs corresponding to the datasets by cutting the loop part out of the alignment and ungapping it. Figure 11 shows a boxplot representing the length values found for each dataset. Table A1 in Appendix II shows the calculations associated to the boxplot (Figure 11).

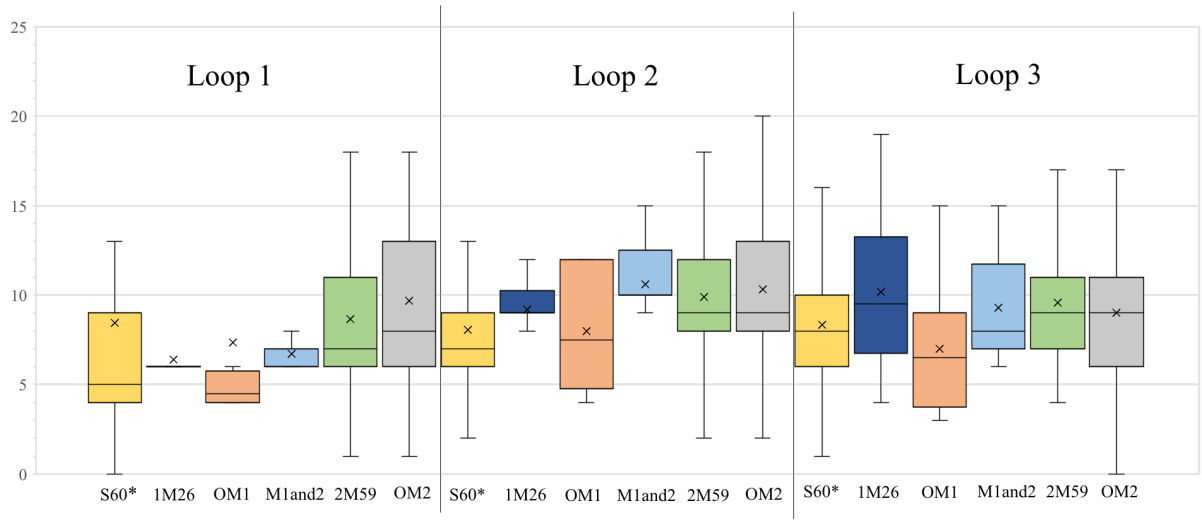

Figure 11 Boxplot showing the lengths of the three binding loops for the six datasets: S60*, 1M26, 2M59, OM1, OM2 and M1and2. *The boxes show the interquartile range, the vertical lines show the outlier line, the Xs are the averages and the horizontal lines inside the boxes are the medians. S60* is yellow, 1M26 is dark blue, OM1 is orange, M1and2 is light blue, 2M59 is green and OM2 is grey.*

From looking at the boxplot one can see that most of the PH domains in the six datasets have diverse lengths, by the sizes of the boxes. The biggest boxes, thus with the most diverse lengths are OM2 L1, OM1 L2, 1M26 L3 and S60* L1. The smallest boxes, where most loops have similar lengths are the 1M26 L1 and L2, and M1and2 L1 and L2. The ranges of the loop lengths vary a lot between the datasets and the loops from the lowest range of seven for L1 in M1and2 to the highest range of 56 for L2 in S60*. Overall the highest loop ranges are L2 for S60* and M1and2, and L1 has the highest range for the four other datasets. The shortest ranges are for L1 in M1and2, L3 for S60* and L2 for the other four datasets.

Overall Loop 1 is the shortest of the three binding loops in all datasets by looking at the median. S60* and OM1 had the shortest L1 median and Q1 median. S60* had a higher L1 mean than OM1. The dataset with the largest median and mean for L1 was OM2. Thus, the result did not confirm the first hypothesis.

For Loop 2 the datasets with the shortest loop were S60* and OM1. The datasets with the longest Loop 2 were M1and2 and OM2. For Loop 3 dataset 1M26 had the highest average loop length. 2M59 and OM2 had medians close to the median for 1M26, but with lower averages. OM1 had the shortest Loop 3 according to both the average and the mean. Hypothesis two and three were not confirmed by this result.

3.2.3 Quantifying variations in secondary structures in PH domains

PH domains usually have seven β-strands and one c-terminal α-helix, but this is not always the case. Therefore the number of the secondary structures were counted in seven datasets: S95, S60*, 1M26, 2M59, OM1, OM2 and M1and2. The number ranges from one to four α-helices and four to 11 β-strands. Figure 12 displays a pie chart representation of the percentages of PH domains with each number of secondary structures in the seven datasets. The number of occurrences and percentages can be found in Table A2 in Appendix II.

For β-strands there is only one dataset which contains a PH domain with 11 β-strands, and that is S95. The second highest number of β-strands is ten, and it is only present in S95 and S60*. The lowest number of β-strands is four, and it is also only present S95 and S60*. In S95, there are seven β-strands in 63% of domains, six are in nine percent and eight β-strands in 20% of the domains. The most common number of β–strands in all the datasets, is seven ranging from 62% in OM2 to 75% in M1and2. The second most common number of β-strands is eight, ranging from zero percent in M1and2 to 33% in OM2.

For α-helices the most common number is one, as mentioned above, and all seven datasets have PH domains with only one α-helix. The highest number of α-helices are four, and all datasets contain at least one PH domain with four α-helices, except OM1 and OM2. The range

for one α-helix goes from 58% in S95 to 70% in M1and2. The range for four α-helices goes from zero percent in OM1 and OM2 to ten percent in 2M59.

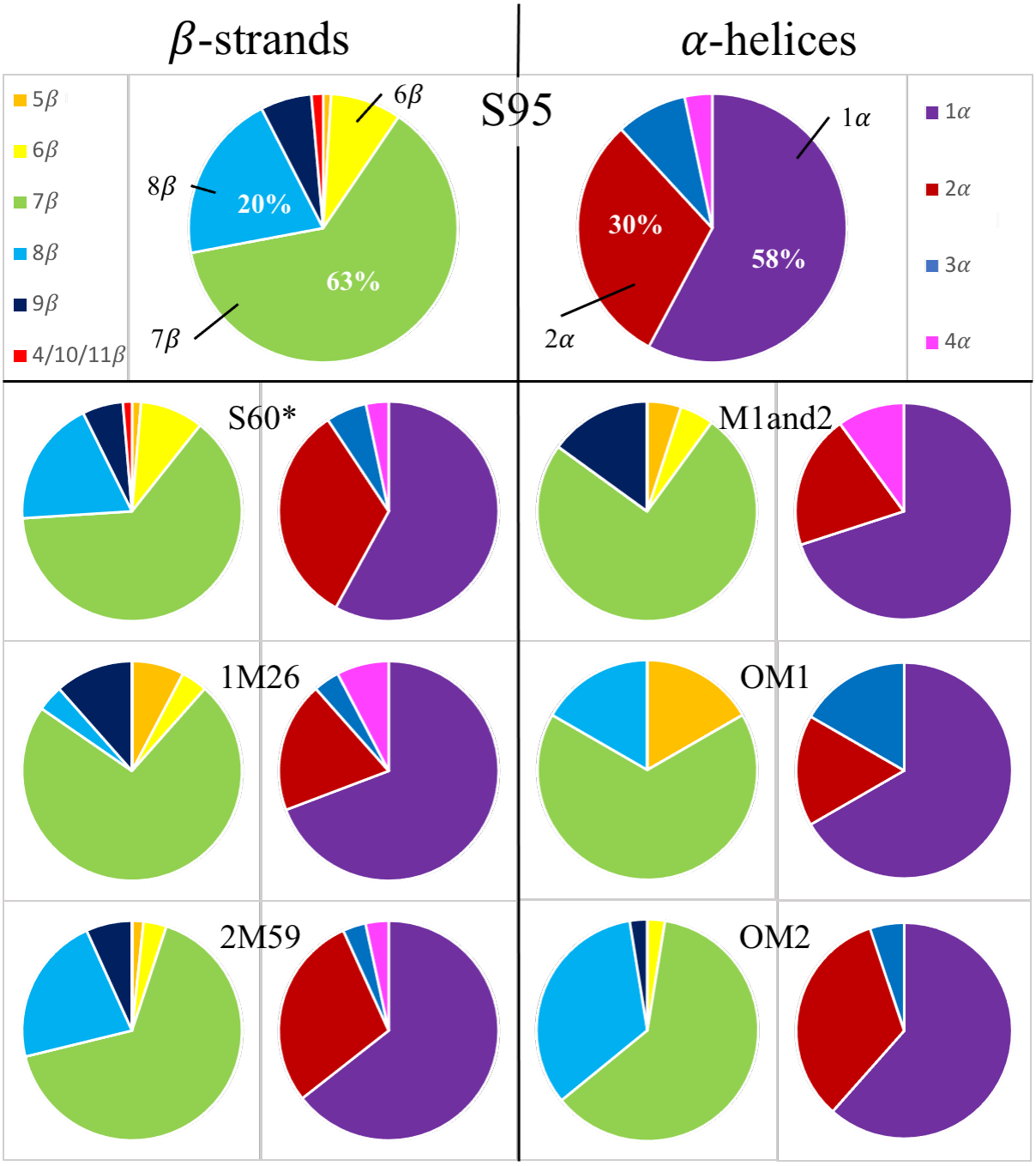

Figure 12 Pie charts showing the percentage of each number of secondary structures in the PH domains in seven datasets. *β-strands are displayed in the left columns and α-helices in the right columns. The different colors correspond to a number of a secondary structure. For β-strands orange is five, yellow is six, green is seven, light blue is eight, dark blue is nine, bright red is four, ten and 11. For α-helices purple is one, dark red is two, blue is three and magenta is four. The size of the colors correspond to the percentage of PH domains with this exact number of the secondary structure in the seven datasets.*

3.3 Part 3 - Amino acid composition of the PH domain

The amino acid composition of the PH domains in the datasets obtained throughout this project was calculated to reach the main goal of mapping the amino acid composition of the membrane binding interface of PH domains. We also wanted to see if some residues are more common than others at different places in the IBS. Datasets were aligned using either T-Coffee Expresso or PROMALS3D (Table 1 p. 28), and the alignments were used to extract different parts of interest, such as binding loops or sequences matching motifs, relative to the domain structure. The number of each amino acid was counted and percentages of each amino acid and amino acid property groups were calculated. The amino acids were grouped into six groups based on properties. These are: hydrophobic (L, I, C, M, W, Y, F), aromatic (W, Y, F), negative (D, E), non-polar (V, A,G, P), polar (S, N, Q, T) and positive (H, K, R).

3.3.1 Amino acid composition of the three binding loops L1, L2 and L3 and the full domain in S60(*)

The three PH domain binding loops are the loops linking β1/β2, β3/β4 and β5/β6 (Lemmon, 2008; Yamamoto *et al.*, 2020). The amino acid composition for these binding loops in S60* and for the whole PH domain in S60, were calculated and compared to each other. PH domains from CATH cluster S60 were used to avoid results where the AA composition from similar PH domains would be weighed too much. The result is visualized in Figure 13. An overview of the most and least common amino acids can be seen in Table 4. The number and percentages of each amino acid can be found in Appendix III.

When comparing all four categories (Loop 1, Loop 2, Loop 3, whole domain), there is a high occurrence of the two amino acids lysine and serine in all categories. There is also a clear trend of low occurring amino acids. All four categories are low in the three amino acids tryptophan, cysteine and methionine. Tyrosine is also low in L1 and L3. Isoleucine is low in both L2 and L3.

In Loop 1 there is a much higher amount of two of the tiny amino acids, alanine and glycine, compared to the other categories. Aspartic acid has a high level in both Loop 2 and 3, and the

level of glutamic acid in Loop 2 is very high, compared to all other amino acids and to the other three categories. Loop 3 has a much higher proline level than the other three categories.

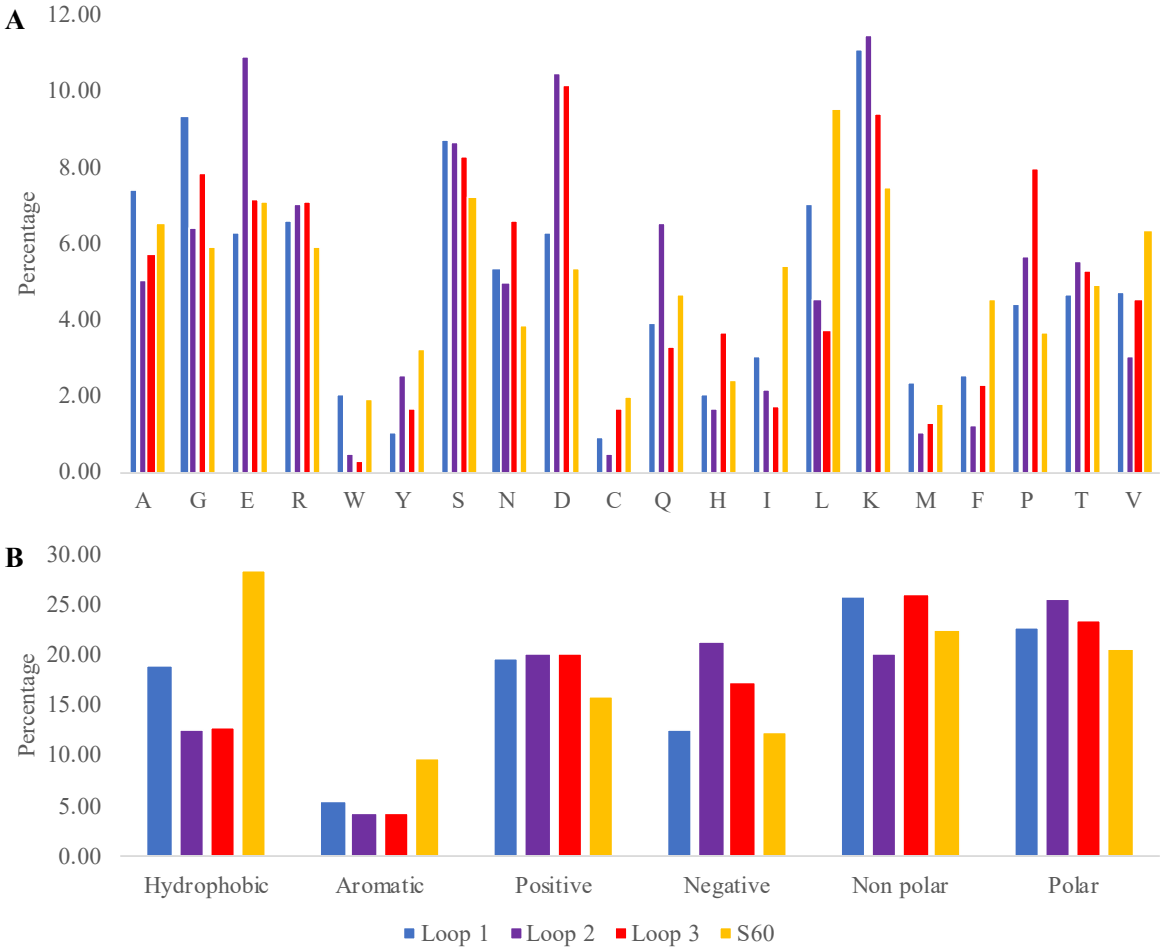

Figure 13 Amino acid composition of the three binding loops in S60* compared to the full PH domain in S60. *A: ratio (in %) of each of the twenty amino acids in the sequences of Loops 1, 2, 3 and the whole PH domain. B: same information for the amino acids grouped by properties. Blue is Loop 1, purple is Loop 2, red is Loop 3 and yellow is the whole domain. The amino acids in the property groups are hydrophobic: L, I, C, M, W, Y, F, aromatic: W, Y, F, positive: H, K, R, negative: D, E, non-polar: V, A, G, P, polar: S, N, Q, T.*

Table 4 Overview of the most and least common amino acids in the S60* binding loops and S60 full domain.

	Most common AAs	Least common AAs
Loop 1	A, G, S, L, K	W, Y, C, M
Loop 2	E, R, S, D, Q, K	W, C, I, M, F
Loop 3	S, D, K, P	W, Y, C, I, M
Full domain	E, S, L, K	W, C, M

For the categories compared by amino acid properties: Loop 2 and 3 have a similar composition of hydrophobic, aromatic and positive residues. For negative amino acids Loop 2 has a higher level because of the glutamic acid as mentioned above. For non-polar amino acids Loop 3 has a higher level because of the proline as mentioned above. The full domain has a much higher level of hydrophobic amino acids than the loops, but Loop 1 has more hydrophobics than the two other loops.

The amount of aromatics are much lower for all three binding loops compared to the full domain. Tryptophan has almost the same level in Loop 1 and in the full domain, but it is extremely low for Loop 2 and 3. Tyrosines are more common than tryptophan in all four categories, but it is highest Loop 2 of the loops, and very low in Loop 1 and 3. Phenylalanine is high in the full domain compared to in the binding loops. There are similar levels of phenylalanine in Loop 1 and 3, and even lower level in Loop 2.

Generally the three binding loops all have a similar and high level of positive amino acids, compared to the lower level in the full domain. Loop 2 has the highest level of lysine, followed by Loop 1. Loop 3 also has a high level of lysine, but it is lower than in the two other loops. Arginine is highest in Loop 3, but the Loops 2 and 3 levels are also high, just a little bit lower. The level of histidine is much higher in Loop 3 than in the other loops.

3.3.2 Amino acid composition of the 1M26 dataset domains and their sequence matching M1

The result from chapter 3.2.1.1 was that there are 26 domains with a sequence matching the canonical binding motif (KX(n)(K/R)XR) out of the 211 domains in S95. In this chapter the result from mapping the amino acid composition of these PH domains will be presented. The result is visualized in Figure 14.

The calculation of the amino acid composition of the sequences matching M1 show that in the X(n) position of the motif the four most common amino acids are glycine, arginine, tryptophan, and lysine. No cysteines were found in this position, and very few of tyrosine, histidine and

methionine. These four amino acids are also very rare in all categories compared (X(n), (X(-2), 1M26 L1, S60* L1 and full domain 1M26 and S60) (Figure 14).

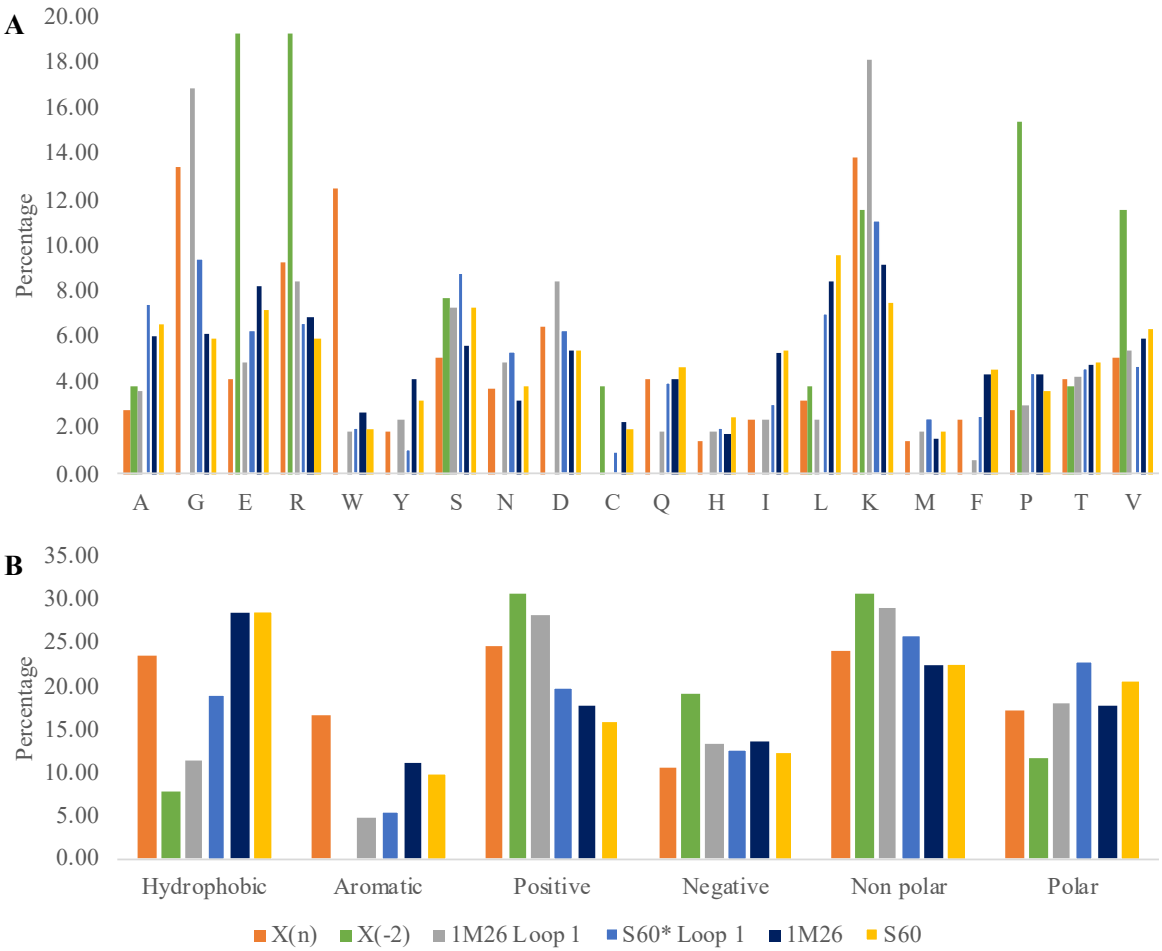

Figure 14 Amino acid composition of the two X positions in the canonical binding motif (M1), (KX(n)(K/R)XR), Loop 1 and the full PH domain in the 1M26 dataset, compared to S60* Loop 1 and S60 full domain . *A: ratio (in %) of each of the twenty amino acids in the sequences for X(n) and X(-2) position in the canonical binding motif (M1) in 1M26, the sequences of loops 1 in 1M26 and S60*, and the sequences for the full PH domain in 1M26 and S60. B: same information for the amino acids grouped by properties. Orange is X(n) in the motif, green is the X(-2) between (K/R) and R in the motif, grey is 1M26 Loop 1, blue is S60* Loop 1, dark blue is the full domain in 1M26 and yellow is the whole S60 domain The amino acids in the property groups are hydrophobic: L, I, C, M, W, Y, F, aromatic: W, Y, F, positive: H, K, R, negative: D, E, non-polar: V, A, G, P, polar: S, N, Q, T.*

X(-2) is the X in the end of M1, between (K/R) and R, and it got its name because it is the second amino acid counting backwards. The five amino acids which are most often in the X(-2) of the motif is glutamic acid, arginine, lysine, proline and valine. The calculation also showed that ten of the amino acids, including all of the aromatic AAs, were not found in the X(-2) position in any of the motifs in the 1M26 dataset. These ten amino acids are glycine,

tryptophan, tyrosine, asparagine, aspartic acid, glutamine, histidine, isoleucine, methionine and phenylalanine.

When comparing Loop 1 in 1M26 and S60* there are many similarities. They both have high levels of glycine and lysine, but while S60* Loop 1 has a glycine and lysine percentage of nine and eleven percent, 1M26 Loop 1 has 17% and 18%, respectively. Thus, the glycine and lysine levels are highest in 1M26 Loop 1. In S60* Loop 1 there are high levels of alanine and serine, while for 1M26 the next highest levels are for arginine and aspartic acid. Comparing Loop 1 in 1M26 and S60* they have similarities in the amino acids which are not so common. They both have low levels of cysteine, tryptophan, histidine, methionine and phenylalanine. In S60* Loop 1 tyrosine is also very low and glutamine is low in 1M26 Loop 1.

Looking at the amino acid composition of the full PH domains in 1M26 and S60 the same amino acids have high and low levels in both datasets. They both have high levels of leucine, lysine and glutamic acid, but the levels are highest in 1M26. S60 also has a high level of serine. The levels of cysteine, methionine and histidine are low in both datasets.

3.3.3 Amino acid composition of the 2M59 dataset domains and their sequence matching M2

The result from chapter 3.2.1.2 was that there are 59 PH domains with a sequence matching the non-canonical binding motif (K/R)X(W/Y/F) out of the 211 domains in S95. In this chapter the result from mapping the amino acid composition of these PH domains will be presented. The result is visualized in Figure 15.

Looking at the X position in M2 there are six amino acids which are not present in this position in the sequence for any of the domains in 2M59. These six amino acids are alanine, tryptophan, tyrosine, aspartic acid, cysteine and isoleucine. Three amino acids were only found for X in one domain each in the dataset. These three are histidine, methionine and phenylalanine. This means that the amino acid property group with the lowest frequency in this position are the aromatics. The five AAs with the highest level of occurrence are asparagine, threonine, glutamic acid, lysine and arginine. All of these five AAs are polar.

Comparing the amino acid composition of the full PH domains in 2M59 and S60, the same amino acids have high and low levels in both. This is the same result as the comparison of full domain in 1M26 and S60 in the previous chapter, 3.3.2. The 2M59 also have, in addition to 1M26 and S60, the highest levels of lysine, leucine, serine and glutamic acid, and the lowest levels of methionine, histidine and cysteine.

Figure 15 Amino acid composition of the X position in the non-canonical binding motif (M2), (K/R)X(W/Y/F) and the full PH domain in the 2M59 dataset, compared S60 full domain . *A: ratio (in %) of each of the twenty amino acids in the sequences for X position in the non-canonical binding motif (M2) in 2M59 and the sequences for the full PH domain in 2M59 and S60. B: same information for the amino acids grouped by properties. Yellow is the whole PH domain in S60, dark green is the whole PH domain in 2M59 and brown is the X position in the non-canonical binding motif. The amino acids in the property groups are hydrophobic: L, I, C, M, W, Y, F, aromatic: W, Y, F, positive: H, K, R, negative: D, E, non-polar: V, A, G, P, polar: S, N, Q, T.*

3.3.4 Amino acid composition of the three binding loops in OM1, OM2, M1and2 and S60*

It has been proposed that a binding loop used for binding contains more basic amino acids than a loop not used for binding (Lemmon, 2007, 2008; Naughton, Kalli and Sansom, 2018)

Because of this I had one amino acid composition hypothesis for each of the three binding loops. Loop 1 should have a high basic AA level in datasets OM1, OM2 and M1and2, and lower level in S60*. Loop 2 should have the highest basic AA level in OM1 and lowest in OM2, and Loop 3 should highest basic AA level in OM2 and lowest in OM1. The amino acid composition of each amino acid, and grouped by amino acid properties, in the four datasets OM1, OM2, M1and2 and S60* was calculated. The graphs showing the AA composition can be seen in Figure 16 (Loop 1), Figure 17 (Loop 2) and Figure 18 (Loop 3).

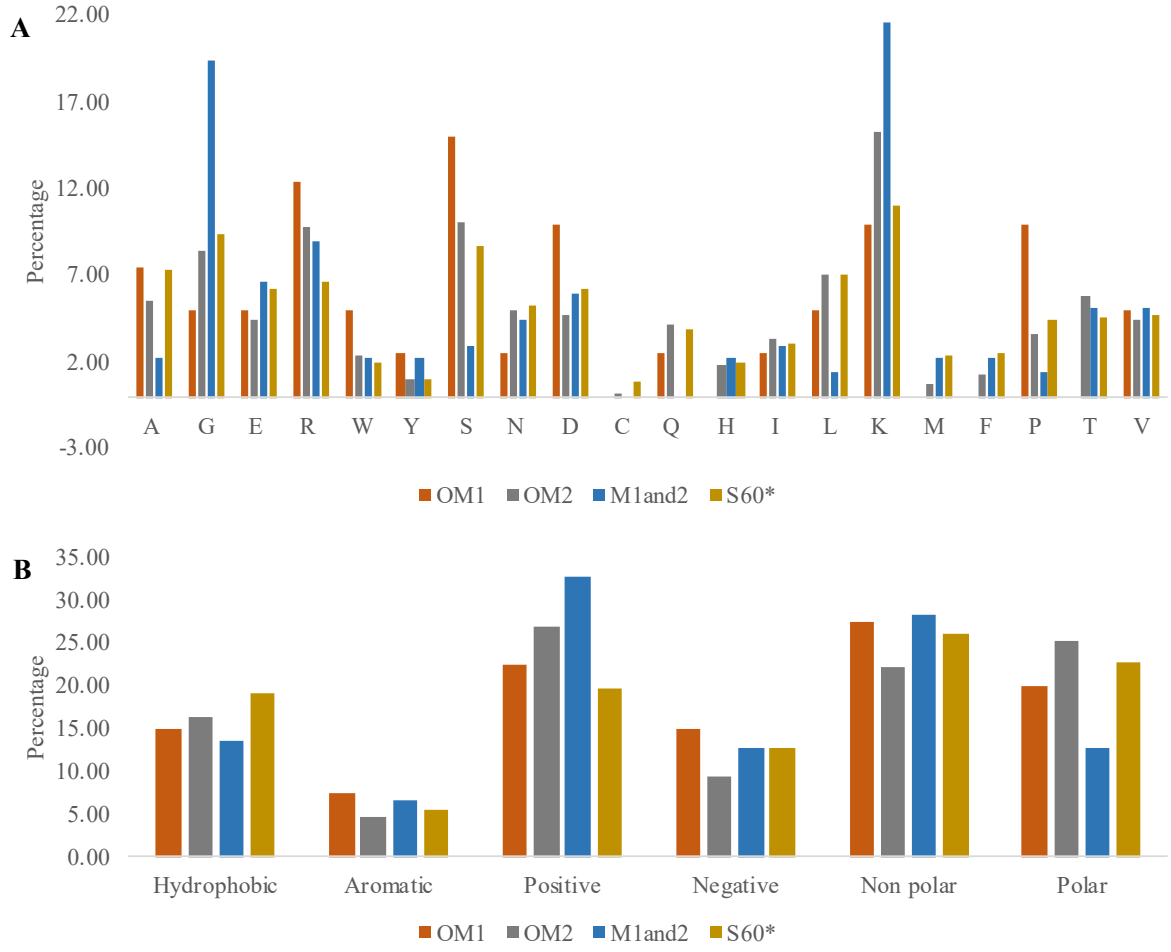

Figure 16 Loop 1 amino acid composition in OM1, OM2, M1and2 and S60*. *A: ratio (in %) of each of the twenty amino acids in the sequences for the binding loop linking β1 and β2 in the PH domains containing only a sequence matching M1 for canonical membrane binding (OM1), the PH domains containing only a sequence matching M2 for non-canonical membrane binding (OM2), the PH domains containing a sequence matching both M1 for canonical membrane binding and M2 for non-canonical membrane binding (M1and2), and for the S60* dataset. B: same information for the amino acids grouped by properties. Orange is OM1, grey is OM2, blue is M1and2 and mustard is S60*. The amino acids in the property groups are hydrophobic: L, I, C, M, W, Y, F, aromatic: W, Y, F, positive: H, K, R, negative: D, E, non-polar: V, A, G, P, polar: S, N, Q, T.*

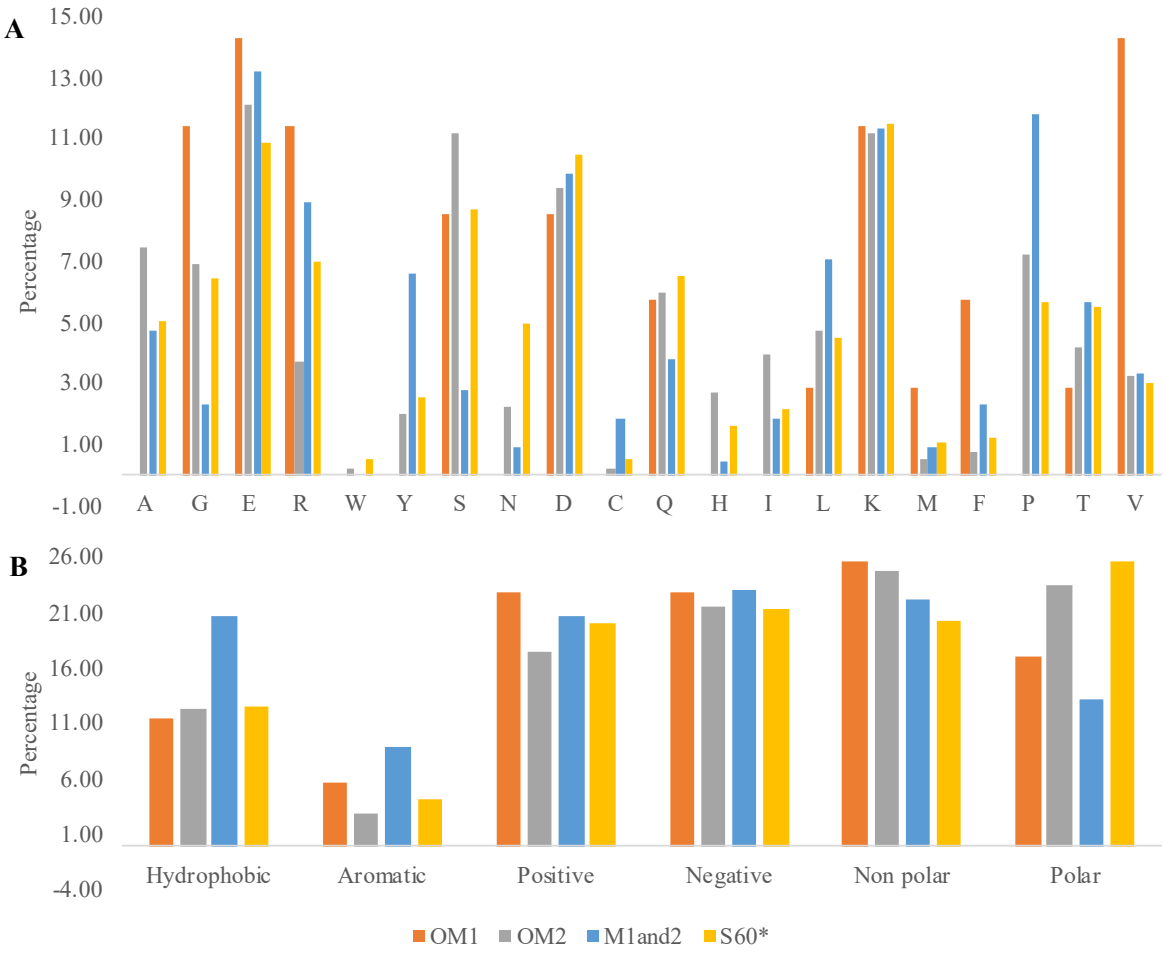

Figure 17 Loop 2 amino acid composition in OM1, OM2, M1and2 and S60*. *A: ratio (in %) of each of the twenty amino acids in the sequences for the binding loop linking β3 and β4 in the PH domains containing only motif 1 for canonical membrane binding (OM1), the PH domains containing only motif 2 for non-canonical membrane binding (OM2), the PH domains containing both motif 1 for canonical membrane binding and motif 2 for non-canonical membrane binding (M1and2), and for the S60* dataset. B: same information for the amino acids grouped by properties. Orange is OM1, grey is OM2, blue is M1and2 and yellow is S60*. The amino acids in the property groups are hydrophobic: L, I, C, M, W, Y, F, aromatic: W, Y, F, positive: H, K, R, negative: D, E, non-polar: V, A, G, P, polar: S, N, Q, T.*

In binding loop 1 all four datasets are high in arginine and lysine, but S60* has the lowest level of basic amino acids. Thus, the first hypothesis is confirmed. In binding loop 2, OM1 is overall highest in positive AAs when all datasets are compared based on AA properties. M1and2 had the second highest level of basic amino acids, followed by S60*. OM2 had the lowest level om basic amino acids in loop 2. This result confirms the second hypothesis. In Loop 3 OM2 has the highest level of basic amino acids, followed by S60*. M1and2 has the third highest level of basic AAs, and OM1 has the lowest level. This partly confirms the third hypothesis.

Another result was that the level of serine was high in OM2 in Loop 2 and in OM1 in loop 3, meaning that serine was one of the highest amino acids present in the binding loops not used for binding. It was also found that the level of tyrosine was high in L2 for dataset M1and2.

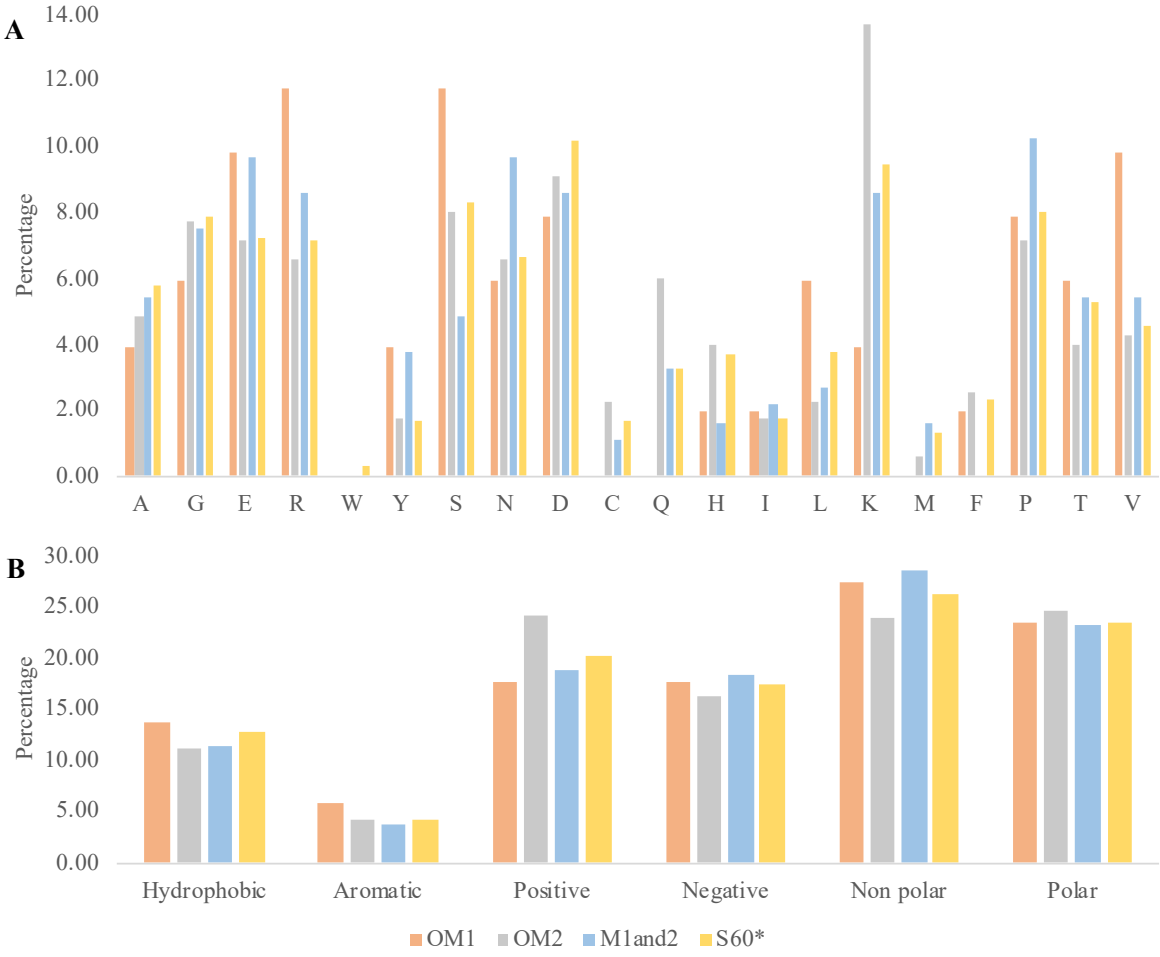

Figure 18 Loop 3 amino acid composition in OM1, OM2, M1and2 and S60*. *A: ratio (in %) of each of the twenty amino acids in the sequences for the binding loop linking β5 and β6 in the PH domains containing only motif 1 for canonical membrane binding (OM1), the PH domains containing only motif 2 for non-canonical membrane binding (OM2), the PH domains containing both motif 1 for canonical membrane binding and motif 2 for non-canonical membrane binding (M1and2), and for the S60* dataset. B: same information for the amino acids grouped by properties. Light orange is OM1, light grey is OM2 ,baby blue is M1and2 and yellow is S60*. The amino acids in the property groups are hydrophobic: L, I, C, M, W, Y, F, aromatic: W, Y, F, positive: H, K, R, negative: D, E, non-polar: V, A, G, P, polar: S, N, Q, T.*

3.4 Part 4 - Conservation of amino acids in the β1 – Loop 1 - β2 region

When a PH domain binds to membranes it uses Loop 1 for both canonical and non-canonical membrane binding. This loop is used for both types of specific binding to PIPs in the membrane. Because of this I wanted to look at the level of conservation of the amino acids in the two binding motifs (M1 and M2) independently in datasets representing PH domains with motif for canonical binding, non-canonical binding and all PH domains. While investigating the alignment of F35 (Figure 9) it was noticed that in addition to the amino acids present in the two binding motifs M1 and M2, two other amino acids were conserved too. These are leucine in the middle of β1 and glycine in the beginning of Loop 1. The conservation of these two amino acids and the presence of the AAs in the two motifs were investigated in the three datasets F35, 2M59 and S60*. Dataset F35 was used instead of dataset 1M26 because all domains in 1M26 are also in F35 and it was the alignment of F35 that revealed the conservation of leucine and glycine. Table 5 shows the result of the amino acid conservation calculation.

Table 5 Conservation of amino acids in the β1 -L1 - β2 region in the F35, S60* and 2M59 datasets. *The numbers in parenthesis are the percentage value of the amino acids in the column respectively, and the subtracted numbers are the percentage in which both amino acids were present in the sequence cut-out of the MSA.*

Amino acid	L	K / R	G	K / R	W / Y / F	K / R	K / R
Position in structure	**Middle of β1**	**End of β1**	**Beginning of Loop 1**	**End of Loop 1**	**Beginning of β2**	**Beginning of β2**	**Middle of β2**
% in F35	83	(57 / 23) = 80	57	(74 / 14) = 89	86 / 0 / 0 = 86	63 / 31 = 94	0 / 91 = 91.43
% in S60*	74	(52 / 33) - 17 = 68	35	(44 / 26) - 8 = 62	(38 / 4 / 8) - 1 = 49	(41 / 25) - 6 = 60	(12 / 42) - 3 = 51
% in 2M59	83	(75 / 34) - 25 = 84	71	(78 / 49) - 30 = 97	(73 / 8 / 12) - 7 = 86	(58 / 15) - 0 = 73	(27 / 53) - 17 = 63

The presence of leucine is high in all three datasets, but slightly lower in S60*. The same is for K/R at the end of β1. For this position lysine is more common than arginine in all three datasets.

The presence of glycine in the beginning of L1 is only 35% for PH domains in S60*, but in F35 and 2M56 the level is at 57% and 71%, respectively.

The level of K/R at the end of L1, the starting amino acid in M2, is present in 97% of the PH domains in 2M29 according to the sequence cut out of the corresponding MSA. F35 has 89% presence of this K/R as well. The level is slightly lower, but still quite high at 62% for S60*. The presence of an aromatic amino acid in the beginning of β2 is high in both F35 and 2M59 at 86% for both. In S60* this aromatic AA is not as common, with a presence in 49% of the PH domains in this dataset. Generally tryptophan is the most common aromatic AA, and in F35 it is the only aromatic amino acid in this position.

The (K/R) part of M1 in the beginning of β2 is present in 94% of the domains in the F35 dataset, and it is also quite high in S60* and 2M59 with 60% and 73%, respectively. In this position both lysine and arginine are common, but lysine is more common. For the ending of M1 there is a K/R located in the middle of β2. There is a 91% presence of arginine in F35, and no lysine. The levels for this K/R in S60* and 2M59 are 51% and 63%, respectively. Both lysines and arginines are found in this position, but the latter is most common.

3.5 Non-PH domains in the CATH 2.30.29.30 superfamily

It was found that the CATH superfamily for PH domains not only contain PH domains, but also four other PH domain-like domains: PTB, EVH, RanBD and Dcp. The S95 cluster contains 211 structures, and 41 of these are non-PH domains. There are 18 PTB, nine EVH, ten RanBD and four Dcp. The S60 cluster contains 162 structures and 29 are non-PH domains. There are 11 PTB, five EVH, nine RanBD and four Dcp.

When comparing the non-PH domain datasets to the other datasets obtained throughout this project it was found that all datasets contained non-PH domains. Table 6 shows which of the datasets contains non-PH domains, and their CATH IDs. All non-PH domains were found in datasets S60 and S60*. All motif datasets contained at least one RanBD domain, and datasets 2M59 and OM2 contained EVH domains as well. PTB and Dcp domains were not found in any of the motif datasets used for mapping the amino acid composition.

Table 6 Non-PH domains in the different datasets used in this project. *PTB is phosphotyrosine binding domain, EVH is enabled/vasodilator-stimulated phosphoprotein (VASP) homology domain, RanBD is Ran binding domain, and Dcp is decapping protein.*

Dataset	Number	CATH ID
S60	29	**PTB:** 1aqcB00, 1ntvA00, 1p3rC00, 1wguA01, 2ej8B00, 2m38A00, 2yt0A01, 3d8dA00, 3dxeA00, 3so6A00, 4xwxA00 **EVH:** 1ddwA00, 1egxA00, 1mkeA00, 1xodB00, 2p8vA00 **RanBD:** 1k5dB00, 1xkeA00, 2ec1A00, 2y8gB00, 3oanA00, 3wyfE00, 4hatB00, 4l6eA00, 5cllB00 **Dcp:** 1q67A01, 2lydA00, 4b6hA00, 5j3tA00
S60*	26	**PTB:** 1aqcB00, 1ntvA00, 1p3rC00, 1wguA01, 2ej8B00, 2m38A00, 2yt0A01, 3d8dA00, 3dxeA00, 3so6A00, 4xwxA00 **EVH:** 1egxA00, 1mkeA00, 1xodB00, 2p8vA00 **RanBD:** 1k5dB00, 1xkeA00, 2ec1A00, 2y8gB00, 3oanA00, 3wyfE00, 4hatB00, 4l6eA00 **Dcp:** 1q67A01, 2lydA00, 4b6hA00
F35	6	**PTB:** 1p3rC00, **RanBD:** 1xkeA00, 2crfA01, 4hatB00, 4l6eA00, 5cllB00
1M26	4	**RanBD:** 1xkeA00, 4hatB00, 4l6eA00, 5cllB00
2M59	8	**EVH:** 1ddvA00, 1ddwA00, 1egxA00, 1qc6A00, 2p8vA00 **RanBD:** 3oanA00, 4hatB00, 5cllB00
M1and2	2	**RanBD:** 4hatB00, 5cllB00
OM1	2	**RanBD:** 1xkeA00, 4l6eA00
OM2	6	**EVH:** 1ddvA00, 1ddwA00, 1egxA00, 1qc6A00, 2p8vA00 **RanBD:** 3oanA00

4. Discussion

4.1 Part 1 - Evaluating the alignment methods

Five alignment methods were tested using the I8 dataset. Clustal Omega, T-Coffee and MUSCLE only used sequence information, and T-Coffee Expresso and PROMALS3D used both sequence and structure information. The result from this was as predicted that T-Coffee Expresso and PROMALS3D were better than the aligners using only the sequential information. PH domains usually have less than 30% sequence identity, but they have a conserved structure (Cho and Stahelin, 2005). Generally structure is better conserved throughout evolution, which is the basis for CATH classification. When domains are in the same H-level in CATH there is good evidence that the domains are related by evolution (Orengo *et al.*, 1997; Illergård, Ardell and Elofsson, 2009). Because of this it is better to use aligners utilizing both sequence and structure for aligning PH domains. It is not sufficient to use an aligner based only on sequence when the sequence identity is so low and we already know there is a conserved structure. T-Coffee Expresso and PROMALS3D gave equally good results in the initial evaluation. Because of this it was decided to use both aligners for all MSAs done throughout the project, and compare the MSAs and see which alignment was best according to the alignment of the predicted secondary structures for each dataset. T-Coffee Expresso had an upper sequence limit of 150 sequences to be aligned which PROMALS3D did not have. Thus PROMALS3D had to be used for the larger datasets.

4.2 Part 2 – Analyzing PH domain MSAs and datasets

4.2.1 Search for binding motifs

The S95 dataset, from CATH superfamily 2.30.29.30 S95 cluster, was investigated for the presence of two motifs M1 (KX(n)(K/R)XR) and M2 (K/R)X(W/Y/F) for canonical and non-canonical membrane binding, respectively. Both motifs have their structural position in the β1-L1-β2 region (Figure 3 p. 13). S95 sequences were subjected to FIMO and looked at manually.

It was found that there are 26 domains containing a motif for canonical binding and 59 domains containing the non-canonical binding motif. These findings resulted in two datasets called 1M26 and 2M59, respectively. These datasets were further used to make the datasets OM1 (only motif 1), OM2 (only motif 2) and M1and2 (both M1 and M2).

4.2.1.1 M1 sequences starting with arginine

Five of the 1M26 domains containing a sequence matching M1 had an arginine instead of lysine as the beginning residue and it was decided to keep them in the dataset because of the similar properties of the two amino acids. It was later discovered that the CATH superfamily 2.30.29.30 not only contained PH domains, but also PTB, EVH, RanBD and Dcp domains. It was then found that four of these five domains containing the sequence matching M1 with a starting arginine were RanBDs (1xkeA00, 4hatB00, 416eA00, 5cllB00). RanBD binds to Ran which in turn binds to GTP (Guanosine TriPhosphate). RanBD binds to Ran-GTP using binding loops 1 and 2 (Vetter *et al.*, 1999; Lemmon, 2008). Because the binding strategy for RanBD corresponds to the PH domain canonical membrane binding, using L1 and L2, the results obtained throughout the project using these domains in dataset 1M26 are still relevant for PH domains.

4.2.1.2 Aromatic ending of M2 (K/R)X(W/Y/F)

The motif for non-canonical membrane binding ends with an aromatic amino acid positioned in the end of L1 or beginning of β2. The most common ending of M2 was with tryptophan, and only seven of 59 sequences matching M2 ended with tyrosine or phenylalanine. According to Naughton, Kalli and Sansom (2018) M2 has a phenylalanine or tyrosine ending if the PH domain can bind using both B1 and B2, but prefers canonical binding (B1). In this project none of these seven domains with phenylalanine or tyrosine ending also contained a motif for canonical binding (M1). A reason for tryptophan to be the most common aromatic amino acid can be that it is the most hydrophobic amino acid according to the Wimley-White hydrophobicity scale and likes to be at the membrane interfacial region (Wimley and White, 1996; Situ *et al.*, 2018). Many peripheral membrane proteins use aromatic amino acids in their IBS. Aromatics can bind to phosphatidyl choline (PC) lipids trough cation-pi interactions but this has not been mentioned for PH domains before (Grauffel *et al.*, 2013; Waheed *et al.*, 2019). It has on the other hand been suggested that hydrophobic residues in the PH domain in some

cases penetrate the interfacial region to make hydrophobic interactions (Cho and Stahelin, 2005; Lemmon, 2007).

4.2.1.3 Evaluating the search method for binding motifs

When the sequences were subjected to FIMO only one domain was found to match the M2 motif for non-canonical binding. Because of this all the sequences in the β1-L1-β2 region were investigated manually to detect this binding motif. From the manual investigation of the S95 domains 59 domains were found to have a sequence matching M2. This means that the result from FIMO was rather flawed. By default FIMO only shows motif matches with a *P* value above 1e-4 (Grant, Bailey and Noble, 2011). Because M2 is a short motif the result might had been better with the *P* value set to a higher threshold. The initial result when searching for M1 in the β1-L1-β2 region using FIMO was 35 domains, and it was later reduced to 26. Because of the large amount of undetected domains containing M2 and wrongly found M1, it might be possible that there are more domains in the S95 CATH cluster containing M1. They might just not have been detected. The result could be different if all 211 of the S95 were looked at manually or if a different FIMO *P* value threshold was used in the search for binding motifs.

4.2.2 Lengths of the three binding loops linking β1/β2, β3/β4 and β5/β6

PH domains binds in the canonical (B1) way using L1 and L2, and in the non-canonical (B2) way using L1 and L3, and it has been proposed that the loop lengths correspond to which type of binding is used (Hurley, 2006; Naughton, Kalli and Sansom, 2018). The loop lengths in six datasets (S60*, 2M59, 1M26, M1and2, OM1 and in OM2) were therefore looked at. The hypotheses were that PH domains with M1 and/or M2 had longer membrane binding loops corresponding to the presence of one or both motifs, than loops in PH domains without M1 and/or M2.

The visual lengths of the loops in datasets 1M26, 2M59 and S60* can be seen in Figure 10 (p. 42). One can see that some loops are much longer than others. What cannot be seen is that most of the PH domain structures were missing parts of their structure. Some were missing their loops, while others lacked bigger parts consisting of many β-strands and loops, making it hard

to give a proper result using this method for determining the loop lengths. If the structures with missing parts were removed from the datasets not many would be left for analyzation. It was therefore thought that the utilization of MSA would limit this flaw concerning missing structural parts, due to using sequence alignment with structural information. The six datasets were aligned and the loop regions of the alignments were cut out. The number of amino acids in the loops were counted, quartile calculations were done and a boxplot (Figure 11 p. 43) was made.

The results from this chapter did not confirm the hypotheses. Generally, all the loops had similar lengths with averages ranging from 6.38 to 10.60 AAs and with means ranging from 4.5 to 10. From the box plot it can be seen that there was a large variety of lengths within each of the loops and datasets as the boxes are quite large.

4.2.2.1 Evaluating the method used to find and calculate loop lengths

Reasons for this result not confirming the hypotheses can have multiple reasons. As mentioned above, many of the PH domains lacked different portions of their structure. The method using the MSAs and cutting out the loop regions did not account for missing parts, thus some of the loop sequences seem much shorter than they really are. With a whole loop missing it would count as a loop with the length of zero AA, instead of it being removed from the calculation as in the amino acid composition calculation in part 3. This together with the size of the datasets gives sources of error. The largest dataset, S60* has 150 domains, and the smallest, OM1, only has six domains. With small datasets each missing part is weighed much heavier than the same occurrence would be in a larger dataset. If the same number of loops were missing it would give a higher weighed value in dataset OM1 than S60*.

Another reason can be the methodology used. DSSP and Chimera secondary structure assigners were used on the MSAs, and the alignments were cut at the best possible position. In some sequences one AA may have been added or lost, which is impossible to avoid without doing it manually. Another reason for this result to not correspond to the hypothesis, can be that the loops lengths are not as important for membrane binding as it was previously thought, as the hypotheses were based on smaller studies done on a few PH domains.

The sequences used in this project were extracted from PDB files of experimentally determined structures obtained from CATH. As explained above, many of these structures had missing parts. To overcome this source of error the datasets could have been enriched with sequences from UniProt to replace the PH domains with missing parts (Bateman *et al.*, 2021). Unfortunately it was not possible to do this in the time of this master project, but it is what I would do if I was to do the experiment again.

4.2.3 Quantifying variations of secondary structures in PH domains

PH domains usually have seven β-strands and 1 c-terminal α-helix, but this is not always the case (Timm *et al.*, 1994; Cho and Stahelin, 2005; Lemmon, 2008). It was also found when looking at the loop lengths that many of the PH domains were missing parts of their structures. This was interesting and I wanted to check in the datasets used in this project, the number of secondary structures per domain. 61% to 75% had seven β-strands and 57% to 70% had only one α-helix, thus most PH domains in the datasets did in fact have the normal number of secondary structures. At the same time it was also shown that quite a few PH domains do not have the normal number of secondary structures, and the numbers for β-strands range between four to 11, and one to four for α-helixes. This is an important aspect of how diverse PH domains can be.

4.3 Part 3 - Amino acid composition of the PH domain

The main goal of this project was to map the amino acid composition of the peripheral membrane binding interface of PH domains. This was successfully accomplished, and the composition result was presented in chapter 3.3 (Part 3 - Amino acid composition of the PH domain). The main findings will be discussed in this section.

4.3.1 High level of basic amino acids in membrane binding loops

It has been reported in the literature that basic residues in the membrane binding loops, other than the ones present in the binding motifs, help with overall binding to membranes (Lemmon,

2007, 2008; Naughton, Kalli and Sansom, 2018). My hypotheses were therefore that there should be more basic residues in Loops 1 and 2 for PH domains with motif for canonical binding, and in Loops 1 and 3 for PH domains with motif for non-canonical binding. The result from chapter 3.3.4 (Amino acid composition of the three binding loops in OM1, OM2, M1and2 and S60*) confirmed these hypotheses.

4.3.1.1 Comparing the presence of the basic amino acids

Overall the level of basic amino acids followed what was expected, but it is interesting to look at the levels of each of the three amino acids in this property group. The positive amino acids are lysine, arginine and histidine. In all parts concerning the amino acid composition the level of histidine is much lower than arginine and lysine. A reason for this can be that histidine with its imidazole ring is bulkier than the two other basic amino acids, making it harder to bind PIPs. One would suspect that the levels of arginine and lysine are similar as they are in the same amino acid property group and are similar in both size and shape.

The result from this project is that lysines are much more common than arginines. In some cases the levels of arginine and lysine are the same, like for the X in M2 located at the end of L1. The only great exception here is in the X(-2) at the end of M1 between (K/R) and R in the motif, where arginine is much more common than lysine. This trend can also be seen in the result from chapter 3.4 Part 4 about conserved residues in the β1-L1-β2 region, where the two motif sequences are located (Table 5 p. 55). In the beginning part of the region lysines are almost double as common as arginines, but in the middle of β2 where M1 ends; arginines are double or more common than lysines, also in the PH domains without a sequence matching M1. The conclusion is therefore that lysines are more common than arginines except for in the middle of β2.

4.3.2 High glycine level in L1 in PH domains with both M1 and M2

The result from chapter 3.3.2 (Amino acid composition of the 1M26 dataset domains and their sequence matching M1) shows a high level of glycine in Loop 1 in dataset 1M26. An explanation of the high presence of glycine here can be to give the binding loop more length and flexibility to better bind to a PIP, which is not necessary when the PH domain does not bind to PIPs (Okoniewska, Tanaka and Yada, 2000). The result from part 4 about conservation

of AAs in the β1-L1-β2 region was that the presence of glycine in the beginning of L1 is much higher in the datasets containing domains with a sequence matching M1 or M2 than in S60*. This finding also substantiate the importance of the glycine for membrane binding.

4.3.3 High tyrosine level in L2 in PH domains with both M1 and M2:

It was found that the tyrosine level in L2 was almost as high as in the full domain, and it was found that the tyrosine level in L2 was very high in PH domains containing a sequence matching both M1 and M2. It is said that PH domains which are able to bind in the canonical and non-canonical way prefer canonical binding using L1 and L2 (Naughton, Kalli and Sansom, 2018). This tyrosine might help L2 to anchor in the membrane leading to the preferred canonical binding (Cho and Stahelin, 2005; Lemmon, 2007).

4.3.4 High serine level in loops not used for membrane binding

It was found in chapter 3.3.1 (Amino acid composition of the three binding loops L1, L2 and L3 and the full domain in S60(*)) that the most common amino acids in the S60* binding loops were serine and lysine. Another result from chapter 3.3.4 (Amino acid composition of the three binding loops in OM1, OM2, M1and2 and S60*) was that in Loop 2, the serine level was high for dataset OM2, and for Loop 3 the serine level was high for dataset OM1. The immediate conclusion could be that when the loop is not used in binding the serine level is high, and the high serine levels in the S60* loops is high because most PH domains do not bind to PIPs with canonical or non-canonical binding. But, the result from 3.3.4.1 concerning binding Loop 1 was a very high level of serine for OM1, when the expected outcome would be low for all dataset but S60*. Even though there can be no exact conclusion from this result I think it is worth mentioning in the discussion. Serine is a hydrophilic amino acid, thus the high serine level can help with avoidance of a membrane. A high level of hydrophilic amino acids can also lead to the secondary structure being a loop and not a part of a β-strand, to keep the structure conserved even if that specific loop is not used in binding.

4.4 Part 4 - Conservation of amino acids in the β1 – Loop 1 - β2 region

In part 4 of this project the level of conservation of amino acids in binding motifs in the β1-L1-β2 region was checked in three datasets: F35, S60* and 2M59 (Table 5 p. 55). When analyzing the MSA of dataset F35 in part 2 of the project two additional conserved amino acids were revealed. These were leucine in the middle of β1 and glycine in the beginning of L1, therefore the conservation of these two AAs was checked in the three datasets as well. The result was that all checked amino acids in this region were present in over 48% of the domains in all three dataset, with just one exception. Glycine at the beginning of L1 is only present in 35% of the domains in S60*, which is discussed above.

All the basic amino acids in this region are conserved with over 50% presence in the S60* domains, indicating importance outside of the two types of PH domain binding focused on in this project (canonical and non-canonical). The (K/R) at the end of β1 has the highest level of conservation of the basic AAs checked here, indicating this as the most important amino acid for PIP binding. Indeed, studies where this lysine was mutated to alanine resulted in a loss of binding function (Cronin *et al.*, 2004). The basic amino acids in this region when not in a motif, might be used in non-specific membrane binding with lower affinity and specificity. It has been suggested that PH domains with low PIP affinity can bind to multiple PIPs simultaneously to enable tighter binding to the membrane (Yamamoto *et al.*, 2020), and these basic amino acids might play a part in this PIP-clustering.

Leucine was found to be conserved in the middle of β1, and it is present in 83% of the domains in F35 and 2M59 and 74% in S60*, which indicates a possible importance of this specific amino acid. Leucine is a hydrophobic amino acid which prefers alpha helices over β-strands. Because the leucine is conserved in a β-strand here it probably has a function. This could be to help with the tightness of the membrane binding.

4.4.1 Evaluating the method for calculating level of conservation

The results show that there are only 86% of domains in 2M59 with an aromatic AA in the beginning of β2, but there should have been 100%. This is because all domains in 2M59 have

a sequence matching M2 ((K/R)X(W/Y/F)) in this region. A reason for this result of 86% domains with an aromatic amino acid can be the method. MSAs of datasets S60* and 2M59 were made, and the amino acid positions were cut out according to the structural predictions. The MSA methods used are heuristic, which means it does not guarantee the best possible result, but still a good outcome (Xiong, 2006). The results from this section can therefore only give an approximate impression of the overall level of conservation, not an exact percentage.

4.5 Non-PH domains in the CATH 2.30.29.30 superfamily

It was found that the CATH superfamily for PH domains not only contain PH domains, but also four other domains: PTB, EVH, RanBD and Dcp. These domains share a common structural fold, but that is not all. PTB domains generally bind to phosphotyrosine in small peptides, but some can bind to PIPs (DiNitto and Lambright, 2006). EVH domains also generally bind to small peptides, and EVH recognizes proline rich peptides using three conserved aromatic residues: Tyrosine in β1, tryptophan in the end of L1/beginning of β2 and phenylalanine in β6. This tryptophan has the same position as the tryptophan in M2 for PH domain non-canonical membrane binding. It has also been reported that some EVH domains can bind to membranes (Rottner *et al.*, 1999; Castellano *et al.*, 2001; Ball *et al.*, 2002; Renfranz and Beckerle, 2002). Dcp domains closely resemble EVH domains, and they also have the conserved tryptophan in β2, but it is not known if they can bind to membranes (She *et al.*, 2004). As written in chapter 4.2.1 (Search for binding motifs) RanBD does not bind to membranes, but it uses L1 and L2 in binding to Ran-GTP, which is the same as for PH domain canonical binding (Vetter *et al.*, 1999; Lemmon, 2008).

The result showed that many of the datasets with sequences matching M1 for canonical and/or M2 for non-canonical PIP binding contained some non-PH domains, especially EVH and RanBD domains. This indicates that these binding motifs might be important for binding generally, and not just for PIP binding in PH domains. It would be interesting to check either experimentally or with simulation if these non-PH domains containing sequences matching canonical and/or non-canonical binding motifs are able to bind to PIPs like PH domains can.

4.6 Future prospects

The structural position of where the conserved leucine was found in the middle of β1, raises a question about a bigger part of the domain used in binding. This project focuses on the loops connecting the β-strands, but it might be interesting in the future to look at the amino acid composition of the β-strand-half closest to a binding loop, considering this conserved leucine in the middle of β1 and that the motif for canonical membrane binding stretches from the end of β1 to the middle of β2. Mutation studies on the importance of this leucine would be interesting. It would also be interesting to look more closely at the importance of tyrosine in L2 in PH domains able to bind in both the canonical and non-canonical way. Another interesting finding that should be looked into is the possible high presence of serine in loops not used in specific membrane binding.

As mentioned above, another inspection for determining the loop lengths should be executed. In this study the domain sequences with missing parts from experimentally determined structures should be replaced with complete sequences fom UniProt.

Other future prospects could involve looking at bigger datasets of PH domains for example including artificial intelligence predicted 3D structures from Alphafold (Jumper *et al.*, 2021). Or looking at other types of PH domain binding like how PH domains bind to other proteins, where the opposite side of the PH domain binds using the c-terminal α-helix and β5, β6 and β7, and see if this part can bind to membranes as well. It would also be interesting to look more closely at the composition of the split PH domains (Lemmon, 2008; Okuda *et al.*, 2017). The similarities and differences between the split PH domain could be compared to what we now know about the composition of the full length PH domain.

4.7 Conclusions

As described in the introduction of my thesis, the characterization of the membrane binding sites in PH domains was rather incomplete, both in terms of amino acid content and structure. Most previous studies of PH domains involved only one or a few PH domains, and they were usually closely related in either function or sequence. My work using a bioinformatical

approach aimed to fill this knowledge gap by mapping the interfacial binding sites of a larger and more diverse dataset of PH domains. This project has revealed several new patterns.

Tryptophan is the most common ending of the motif for non-canonical membrane binding (M2). PH domains with this motif ending with tyrosine or phenylalanine do not have a sequence matching motif for canonical membrane binding (M1). The amino acid glycine structurally positioned at the beginning of Loop 1 connecting β1/β2 is more common in PH domains with a sequence matching M1 and/or M2.

The level of basic amino acids follows the pattern of which loop is used for PIP binding. PH domains with a sequence matching M1 and/or M2 have more basic amino acids than PH domains which do not have a sequence matching a binding motif. Domains with a sequence matching M1 have many basic AAs in L2 connecting β3/β4, and PH domains with a sequence matching M2 have many basic AAs in L3 connecting β5/β6. Lysines are more common than arginine, with one exception, which is in the middle of β2.

Even though the goals of this project were largely achieved using experimentally determined structures, the work can be taken further. In particular I would be interested in extending my datasets by including predicted structures for PH domains not represented in the Protein Data Bank. The Alphafold database contains computationally predicted 3D structures of high quality and it was released in July 2021 (Jumper *et al.*, 2021). This information could be utilized to continue this project.

References

Allen, K. N. *et al.* (2019) 'Monotopic Membrane Proteins Join the Fold', *Trends Biochem Sci.*, 44(1), pp. 7–20. doi: 10.1016/j.tibs.2018.09.013.Monotopic.

Andreeva, A. *et al.* (2020) 'The SCOP database in 2020: Expanded classification of representative family and superfamily domains of known protein structures', *Nucleic Acids Research*, 48(D1), pp. D376–D382. doi: 10.1093/nar/gkz1064.

Armougom, F. *et al.* (2006) 'Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee', *Nucleic acids research*, 34(Web Server issue), pp. W604–W608. doi: 10.1093/nar/gkl092.

Baldauf, S. L. (2003) 'Phylogeny for the faint of heart: A tutorial', *Trends in Genetics*, 19(6), pp. 345–351. doi: 10.1016/S0168-9525(03)00112-4.

Ball, L. J. *et al.* (2002) 'EVH1 domains: Structure, function and interactions', *FEBS Letters*, 513(1), pp. 45–52. doi: 10.1016/S0014-5793(01)03291-4.

Bateman, A. *et al.* (2021) 'UniProt: The universal protein knowledgebase in 2021', *Nucleic Acids Research*, 49(D1), pp. D480–D489. doi: 10.1093/nar/gkaa1100.

Blobel, G. (1980) 'Intracellular protein topogenesis.', *Proceedings of the National Academy of Sciences of the United States of America*, 77(3), pp. 1496–1500. doi: 10.1073/pnas.77.3.1496.

Castellano, F. *et al.* (2001) 'A WASp-VASP complex regulates actin polymerization at the plasma membrane', *EMBO Journal*, 20(20), pp. 5603–5614. doi: 10.1093/emboj/20.20.5603.

Ceccarelli, D. F. J. *et al.* (2007) 'Non-canonical interaction of phosphoinositides with pleckstrin homology domains of Tiam1 and ArhGAP9', *Journal of Biological Chemistry*, 282(18), pp. 13864–13874. doi: 10.1074/jbc.M700505200.

Cho, W. and Stahelin, R. V. (2005) 'Membrane-protein interactions in cell signaling and membrane trafficking', *Annual Review of Biophysics and Biomolecular Structure*, 34, pp. 119–151. doi: 10.1146/annurev.biophys.33.110502.133337.

Clayden, J., Greeves, N. and Warren, S. (2012) *Organic Chemistry*. Second Edi. Oxford: Oxford University Press.

Cock, P. J. A. *et al.* (2009) 'Biopython: Freely available Python tools for computational molecular biology and bioinformatics', *Bioinformatics*, 25(11), pp. 1422–1423. doi: 10.1093/bioinformatics/btp163.

Corey, R. A., Stansfeld, P. J. and Sansom, M. S. P. (2019) 'The energetics of protein–lipid

interactions as viewed by molecular simulations', *Biochemical Society Transactions*, 48(1), pp. 25–37. doi: 10.1042/BST20190149.
Cronin, T. C. *et al.* (2004) 'Structural determinants of phosphoinositide selectivity in splice variants of Grp1 family PH domains', *EMBO Journal*, 23(19), pp. 3711–3720. doi: 10.1038/sj.emboj.7600388.
Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978) 'A Model of Evolutionary Change in Proteins', *Atlas of Protein Sequence and Structure*, 5(3), pp. 345–352.
DiNitto, J. P. and Lambright, D. G. (2006) 'Membrane and juxtamembrane targeting by PH and PTB domains', *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids*, 1761(8), pp. 850–867. doi: 10.1016/j.bbalip.2006.04.008.
Do, C. B. *et al.* (2005) 'ProbCons : Probabilistic consistency-based multiple sequence alignment', pp. 330–340. doi: 10.1101/gr.2821705.1994.
Edgar, R. C. (2004) 'MUSCLE: A multiple sequence alignment method with reduced time and space complexity', *BMC Bioinformatics*, 5, pp. 1–19. doi: 10.1186/1471-2105-5-113.
Falkenburger, B. H. *et al.* (2010) 'Phosphoinositides: Lipid regulators of membrane proteins', *Journal of Physiology*, 588(17), pp. 3179–3185. doi: 10.1113/jphysiol.2010.192153.
Feng, D.-F. and Doolittle, R. F. (1987) 'Journal of Molecular Evolution Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees', *J Mol Evol*, 25, pp. 351–360.
Ferguson, K. M. *et al.* (2000) 'Structural basis for discrimination of 3-phosphoinositides by pleckstrin homology domains', *Molecular Cell*, 6(2), pp. 373–384. doi: 10.1016/S1097-2765(00)00037-X.
Floden, E. W. *et al.* (2016) 'PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases', *Nucleic acids research*, 44(W1), pp. W339–W343. doi: 10.1093/nar/gkw300.
Fuglebakk, E. and Reuter, N. (2018) 'A model for hydrophobic protrusions on peripheral membrane proteins', *PLoS Computational Biology*, 14(7), pp. 1–26. doi: 10.1371/journal.pcbi.1006325.
Grant, C. E., Bailey, T. L. and Noble, W. S. (2011) 'FIMO: Scanning for occurrences of a given motif', *Bioinformatics*, 27(7), pp. 1017–1018. doi: 10.1093/bioinformatics/btr064.
Grauffel, C. *et al.* (2013) 'Cation-π interactions as lipid-specific anchors for phosphatidylinositol-specific phospholipase C', *Journal of the American Chemical Society*, 135(15), pp. 5740–5750. doi: 10.1021/ja312656v.

Gurtovenko, A. A. and Vattulainen, I. (2007) ‘Molecular mechanism for lipid flip-flops’, *Journal of Physical Chemistry B*, 111(48), pp. 13554–13559. doi: 10.1021/jp077094k.
Haslam, R. J., Koide, H. B. and Hemmings, B. A. (1993) ‘Pleckstrin domain homology’, *Nature*, 363, pp. 309–310.
Henikoff, S. and Henikoff, J. G. (1992) ‘Amino acid substitution matrices from protein blocks’, *Proc. Natl Acad. Sci*, 89(November), pp. 10915–10919. doi: 10.1073/pnas.89.22.10915.
Hogeweg, P. and Hesper, B. (1984) ‘The alignment of sets of sequences and the construction of phyletic trees: An integrated method’, *Journal of Molecular Evolution*, 20(2), pp. 175–186. doi: 10.1007/BF02257378.
Holm, L. and Park, J. (2000) ‘DaliLite workbench for protein structure comparison’, *Bioinformatics*, 16(6), pp. 566–567. doi: 10.1093/bioinformatics/16.6.566.
Huang, X. and Miller, W. (1991) ‘A time-efficient, linear-space local similarity algorithm’, *Advances in Applied Mathematics*, 12(3), pp. 337–357. doi: 10.1016/0196-8858(91)90017-D.
Hurley, J. H. (2006) ‘Membrane binding domains’, *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids*, 1761(8), pp. 805–811. doi: 10.1016/j.bbalip.2006.02.020.
Hyvönen, M. *et al.* (1995) ‘Structure of the binding site for inositol phosphates in a PH domain’, *EMBO Journal*, 14(19), pp. 4676–4685. doi: 10.1002/j.1460-2075.1995.tb00149.x.
Illergård, K., Ardell, D. H. and Elofsson, A. (2009) ‘Structure is three to ten times more conserved than sequence - A study of structural response in protein cores’, *Proteins: Structure, Function and Bioinformatics*, 77(3), pp. 499–508. doi: 10.1002/prot.22458.
Jansen, K. O. (2021) ‘Mapping_AA’. GitHub. doi: 10.5281/zenodo.5647460.
Jian, X. *et al.* (2015) ‘Molecular Basis for Cooperative Binding of Anionic Phospholipids to the PH Domain of the Arf GAP ASAP1’, *Structure*, 23(11), pp. 1977–1988. doi: 10.1016/j.str.2015.08.008.
Jumper, J. *et al.* (2021) ‘Highly accurate protein structure prediction with AlphaFold’, *Nature*, 596(7873), pp. 583–589. doi: 10.1038/s41586-021-03819-2.
Kabsch, W. and Sander, C. (1983) ‘Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogren-Bonded and Geometrical Features’, *Biopolymers.*, 12(2), pp. 577–637. doi: 10.1002/bip.360221211.
Konagurthu, A. S. *et al.* (2006) ‘MUSTANG: A Multiple Structural Alignment Algorithm’, *PROTEINS: Structure, Function, and Bioinformatics*, 64(3), pp. 559–574. doi: 10.1002/prot.
Leinonen, R. *et al.* (2004) ‘UniProt archive’, *Bioinformatics*, 20(17), pp. 3236–3237. doi:

10.1093/bioinformatics/bth191.
Lemmon, M. A. (2007) ‘Pleckstrin homology (PH) domains and phosphoinositides’, *Biochemical Society Symposium*, 74(74), pp. 81–93. doi: 10.1042/BSS0740081.
Lemmon, M. A. (2008) ‘Membrane recognition by phospholipid-binding domains’, *Nature Reviews Molecular Cell Biology*, 9(2), pp. 99–111. doi: 10.1038/nrm2328.
Lemmon, M. A. and Ferguson, K. M. (2000) ‘Signal-dependent membrane targeting by pleckstrin homology (PH) domains’, *The Biochemical journal*, 350 Pt 1(Pt 1), pp. 1–18. Available at: https://pubmed.ncbi.nlm.nih.gov/10926821.
Lian, L. *et al.* (2009) ‘Loss of pleckstrin defines a novel pathway for PKC-mediated exocytosis’, *Blood*, 113(15), pp. 3577–3584. doi: 10.1182/blood-2008-09-178913.
Lomize, A. L. *et al.* (2007) ‘The role of hydrophobic interactions in positioning of peripheral proteins in membranes’, *BMC Structural Biology*, 7, pp. 1–30. doi: 10.1186/1472-6807-7-44.
Luckey, M. (2014) *Membrane Structural Biology*.
Luscombe, N. M., Greenbaum, D. and Gerstein, M. (2001) ‘What is Bioinformatics? A proposed Definition and Overview of the Field’, pp. 346–358.
Madeira, F. *et al.* (2019) ‘The EMBL-EBI search and sequence analysis tools APIs in 2019’, *Nucleic Acids Research*, 47(W1), pp. W636–W641. doi: 10.1093/nar/gkz268.
McGuffin, L. J., Bryson, K. and Jones, D. T. (2000) ‘The PSIPRED protein structure prediction server’, *Bioinformatics*, 16(4), pp. 404–405. doi: 10.1093/bioinformatics/16.4.404.
Naughton, F. B., Kalli, A. C. and Sansom, M. S. P. (2016) ‘Association of Peripheral Membrane Proteins with Membranes: Free Energy of Binding of GRP1 PH Domain with Phosphatidylinositol Phosphate-Containing Model Bilayers’, *Journal of Physical Chemistry Letters*, 7(7), pp. 1219–1224. doi: 10.1021/acs.jpclett.6b00153.
Naughton, F. B., Kalli, A. C. and Sansom, M. S. P. (2018) ‘Modes of Interaction of Pleckstrin Homology Domains with Membranes: Toward a Computational Biochemistry of Membrane Recognition’, *Journal of Molecular Biology*, 430(3), pp. 372–388. doi: 10.1016/j.jmb.2017.12.011.
Needleman, S. B. and Wunsch, C. D. (1970) ‘A general method applicable to the search for similarities in the amino acid sequence of two proteins’, *Journal of Molecular Biology*, 48(3), pp. 443–453. doi: 10.1016/0022-2836(70)90057-4.
Notredame, C., Higgins, D. G. and Heringa, J. (2000) ‘T-coffee: A novel method for fast and accurate multiple sequence alignment’, *Journal of Molecular Biology*, 302(1), pp. 205–217. doi: 10.1006/jmbi.2000.4042.
Okoniewska, M., Tanaka, T. and Yada, R. Y. (2000) ‘The pepsin residue glycine-76

contributes to active-site loop flexibility and participates in catalysis', *Biochemical Journal*, 349(1), pp. 169–177. doi: 10.1042/0264-6021:3490169.
Okuda, M. *et al.* (2017) 'Common TFIIH recruitment mechanism in global genome and transcription-coupled repair subpathways', *Nucleic Acids Research*, 45(22), pp. 13043–13055. doi: 10.1093/nar/gkx970.
Orengo, C. A. *et al.* (1997) 'CATH - A hierarchic classification of protein domain structures', *Structure*, 5(8), pp. 1093–1109. doi: 10.1016/s0969-2126(97)00260-8.
Pearson, W. R. and Lipman, D. J. (1988) 'Improved tools for biological sequence comparison.', *Proceedings of the National Academy of Sciences of the United States of America*, 85(8), pp. 2444–2448. doi: 10.1073/pnas.85.8.2444.
Pei, J., Kim, B.-H. and Grishin, N. V (2008) 'PROMALS3D: a tool for multiple protein sequence and structure alignments', *Nucleic acids research*. 2008/02/20, 36(7), pp. 2295–2300. doi: 10.1093/nar/gkn072.
Pettersen, E. F. *et al.* (2004) 'UCSF Chimera - A visualization system for exploratory research and analysis', *Journal of Computational Chemistry*, 25(13), pp. 1605–1612. doi: 10.1002/jcc.20084.
Renfranz, P. J. and Beckerle, M. C. (2002) 'Doing (F/L)PPPPs: EVH1 domains and their proline-rich partners in cell polarity and migration', *Current Opinion in Cell Biology*, 14(1), pp. 88–103. doi: 10.1016/S0955-0674(01)00299-X.
Rossum, V., Jr, G. & D. and L., F. (2009) 'Python 3 Reference Manual'. Scotts Valley, CA: CreateSpace.
Rottner, K. *et al.* (1999) 'VASP dynamics during lamellipodia protrusion', *Nature Cell Biology*, 1(5), pp. 321–322. doi: 10.1038/13040.
Schrödinger, L. (2015) 'The PyMOL Molecular Graphics System'.
She, M. *et al.* (2004) 'Crystal structure of Dcp1p and its functional implications in mRNA decapping', *Nature Structural Molecular Biology*, 11(3), pp. 249–256.
Sievers, F. *et al.* (2011) 'Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega', *Molecular Systems Biology*, 7(539). doi: 10.1038/msb.2011.75.
Sillitoe, I. *et al.* (2021) 'CATH: Increased structural coverage of functional space', *Nucleic Acids Research*, 49(D1), pp. D266–D273. doi: 10.1093/nar/gkaa1079.
Singer, S. J. and Nicolson, G. L. (1972) 'The fluid mosaic model of the structure of cell membranes', *Science*, 175(4023), pp. 720–731. doi: 10.1126/science.175.4023.720.
Situ, A. J. *et al.* (2018) 'Membrane Anchoring of α-Helical Proteins: Role of Tryptophan',

Journal of Physical Chemistry B, 122(3), pp. 1185–1194. doi: 10.1021/acs.jpcb.7b11227.
Smith, T. F. and Waterman, M. S. (1981) 'Identification of common molecular subsequences', *Journal of Molecular Biology*, 147(1), pp. 195–197. doi: https://doi.org/10.1016/0022-2836(81)90087-5.
Suzek, B. E. *et al.* (2015) 'UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches', *Bioinformatics*, 31(6), pp. 926–932. doi: 10.1093/bioinformatics/btu739.
Taly, J. F. *et al.* (2011) 'Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures', *Nature Protocols*, 6(11), pp. 1669–1682. doi: 10.1038/nprot.2011.393.
Taylor, W. R. (2008) 'Protein structure comparison using iterated double dynamic programming', *Protein Science*, 8(3), pp. 654–665. doi: 10.1110/ps.8.3.654.
Taylor, W. R. and Orengo, C. A. (1989) 'Protein structure alignment', *Journal of Molecular Biology*, 208(1), pp. 1–22. doi: 10.1016/0022-2836(89)90084-3.
Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) 'CLUSTAL W (improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice)', *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*, 22(22), pp. 4673–4680. doi: 10.1007/978-1-4020-6754-9_3188.
Timm, D. *et al.* (1994) 'Crystal structure of the pleckstrin homology domain from dynamin', *Nature Structural Biology*, 1(11), pp. 782–788. doi: 10.1038/nsb1194-782.
Di Tommaso, P. *et al.* (2011) 'T-Coffee: A web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension', *Nucleic Acids Research*, 39(SUPPL. 2), pp. 13–17. doi: 10.1093/nar/gkr245.
Touw, W. G. *et al.* (2015) 'A series of PDB-related databanks for everyday needs', *Nucleic Acids Research*, 43(D1), pp. D364–D368. doi: 10.1093/nar/gku1028.
Tran, Quoc-Nam, Wallinga, M. (2017) 'A Novel Method for Multiple Sequence Alignment Using Morphing Techniques', *Journal of Health Informatics and Management*, 1(2).
Tubiana, T. (2021) 'ColorMSAwithSSE'. GitHub. doi: 10.5281/zenodo.5575160.
Vetter, I. R. *et al.* (1999) 'Structure of a Ran-binding domain complexed with Ran bound to a GTP analogue: Implications for nuclear transport', *Nature*, 398(6722), pp. 39–46. doi: 10.1038/17969.
Waheed, Q. *et al.* (2019) 'Interfacial Aromatics Mediating Cation-πInteractions with Choline-Containing Lipids Can Contribute as Much to Peripheral Protein Affinity for Membranes as Aromatics Inserted below the Phosphates', *Journal of Physical Chemistry*

Letters, 10(14), pp. 3972–3977. doi: 10.1021/acs.jpclett.9b01639.
Waterhouse, A. M. *et al.* (2009) 'Jalview Version 2-A multiple sequence alignment editor and analysis workbench', *Bioinformatics*, 25(9), pp. 1189–1191. doi: 10.1093/bioinformatics/btp033.
Webb, B. L. J., Hirst, S. J. and Giembycz, M. A. (2000) 'Protein kinase C isoenzymes: A review of their structure, regulation and role in regulating airways smooth muscle tone and mitogenesis', *British Journal of Pharmacology*, 130(7), pp. 1433–1452. doi: 10.1038/sj.bjp.0703452.
Wimley, W. C. and White, S. H. (1996) 'Experimentally determined hydrophobicity scale for proteins at membrane interfaces', *Nature Structural Biology*, 3(10), pp. 842–848. doi: 10.1038/nsb1096-842.
Xiong, J. (2006) *Essential Bioinformatics*. New York: Cambridge University Press. Available at: http://marefateadyan.nashriyat.ir/node/150.
Yamamoto, E. *et al.* (2016) 'Interactions of Pleckstrin Homology Domains with Membranes: Adding Back the Bilayer via High-Throughput Molecular Dynamics', *Structure*, 24(8), pp. 1421–1431. doi: 10.1016/j.str.2016.06.002.
Yamamoto, E. *et al.* (2020) 'Multiple lipid binding sites determine the affinity of PH domains for phosphoinositide-containing membranes', *Science Advances*, 6(8). doi: 10.1126/sciadv.aay5736.
Zhang, Y. and Skolnick, J. (2005) 'TM-align: A protein structure alignment algorithm based on the TM-score', *Nucleic Acids Research*, 33(7), pp. 2302–2309. doi: 10.1093/nar/gki524.
Zhu, J. and Weng, Z. (2005) 'FAST: A novel protein structure alignment algorithm', *Proteins: Structure, Function and Genetics*, 58(3), pp. 618–627. doi: 10.1002/prot.20331.

Appendix I – Overview of the datasets in the project

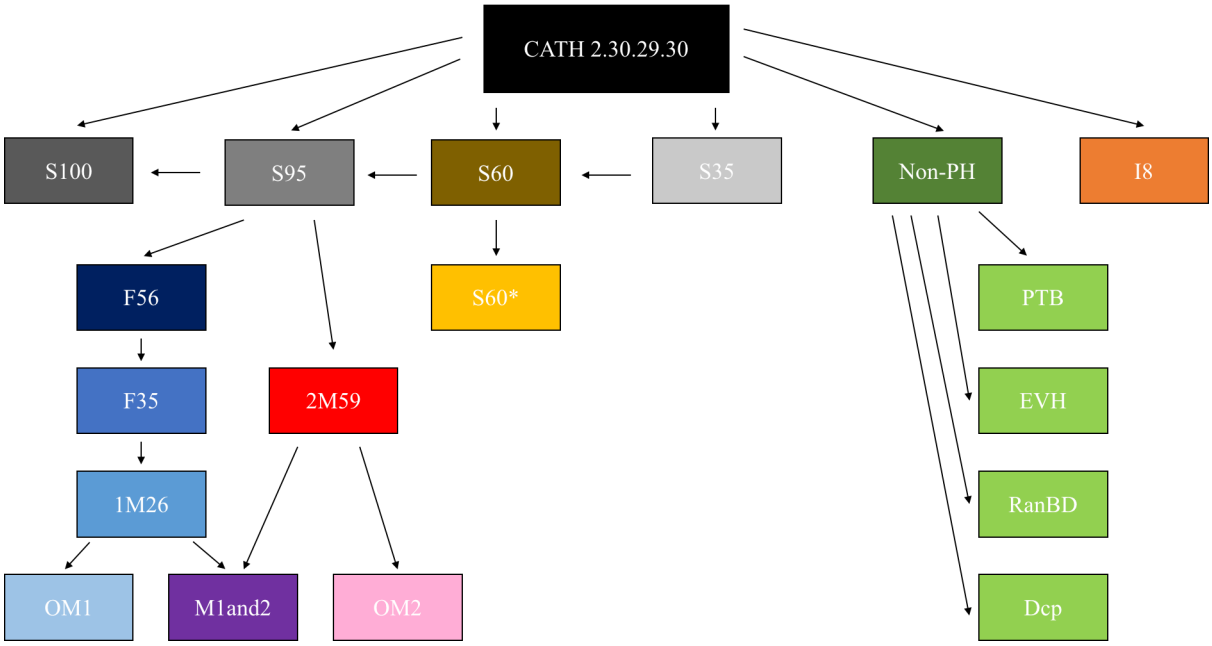

Figure A1 Overview of how the datasets used throughout the project are related, and where they originated from. *Datasets are colored for easier readability of the figure.*

Appendix II – Tables associated with loop length and secondary structure quantification

Table A1 Values associated with the loop length boxplot in Figure 12 in chapter 3.2.2 Lengths of the three binding loops linking β1/β2, β3/β4 and β5/β6.

	S60*			1M26			OM1		
	Loop 1	Loop 2	Loop 3	Loop 1	Loop 2	Loop 3	Loop 1	Loop 2	Loop 3
Minimum	0.00	0.00	1.00	2.00	2.00	4.00	4.00	4.00	3.00
Q1	4.00	6.00	6.00	6.00	9.00	7.25	4.00	5.25	4.50
Median	5.00	7.00	8.00	6.00	9.00	9.50	4.50	7.50	6.50
Q3	9.00	9.00	10.00	6.00	9.75	12.75	5.75	11.25	7.00
Maximum	53.00	56.00	28.00	20.00	13.00	19.00	21.00	12.00	15.00
Mean	8.46	8.07	8.34	6.38	9.19	10.19	7.33	8.00	7.00
Range	53.00	56.00	27.00	18.00	11.00	15.00	17.00	8.00	12.00
IQR	5.00	3.00	4.00	0.00	0.75	5.50	1.75	6.00	2.50
Outlier line	7.50	4.50	6.00	0.00	1.13	8.25	2.63	9.00	3.75
Lower outlier	-3.50	1.50	0.00	6.00	7.88	-1.00	1.38	-3.75	0.75
Upper outlier	16.50	13.50	16.00	6.00	10.88	21.00	8.38	20.25	10.75
	M1and2			2M59			OM2		
	Loop 1	Loop 2	Loop 3	Loop 1	Loop 2	Loop 3	Loop 1	Loop 2	Loop 3
Minimum	4.00	3.00	6.00	1.00	2.00	0.00	1.00	2.00	0.00
Q1	6.00	10.00	7.00	6.00	8.00	7.00	6.00	8.00	6.50
Median	6.00	10.00	8.00	7.00	9.00	9.00	8.00	9.00	9.00
Q3	7.00	11.50	11.25	11.00	12.00	11.00	13.00	12.50	10.50
Maximum	11.00	15.00	15.00	30.00	20.00	28.00	30.00	20.00	22.00
Mean	6.70	10.60	9.30	8.64	9.88	9.59	9.69	10.33	9.00
Range	7.00	12.00	9.00	29.00	18.00	28.00	29.00	18.00	22.00
IQR	1.00	1.50	4.25	5.00	4.00	4.00	7.00	4.50	4.00
Outlier line	1.50	2.25	6.38	7.50	6.00	6.00	10.50	6.75	6.00
Lower outlier	4.50	7.75	0.63	-1.50	2.00	1.00	-4.50	1.25	0.50
Upper outlier	8.50	13.75	17.63	18.50	18.00	17.00	23.50	19.25	16.50

Table A2 Percentages of occurrences PH domains with the exact number of secondary structures in the seven datasets: S95, S60*, 1M26, 2M59, OM1, OM2, M1and2. *This table is associated with the pie charts in Figure 13 in chapter 3.2.3 Number of secondary structures in PH domains.*

N of β	S95	S60*	1M26	2M59	OM1	OM2	M1and2
4	0.47	0.67	0.00	0.00	0.00	0.00	0.00
5	0.95	1.33	7.69	1.69	16.67	0.00	5.00
6	8.53	9.33	3.85	3.39	0.00	2.56	5.00
7	62.56	63.33	73.08	66.10	66.67	61.54	75.00
8	20.38	18.67	3.85	22.03	16.67	33.33	0.00
9	6.16	6.00	11.54	6.78	0.00	2.56	15.00
10	0.47	0.67	0.00	0.00	0.00	0.00	0.00
11	0.47	0.00	0.00	0.00	0.00	0.00	0.00
N of α							
1	57.82	58.00	69.23	64.41	66.67	61.54	70.00
2	30.33	32.67	19.23	28.81	16.67	33.33	20.00
3	8.53	6.00	3.85	3.39	16.67	5.13	0.00
4	3.32	3.33	7.69	3.39	0.00	0.00	10.00

Appendix III - Tables associated with amino acid composition

Table A3 Amino acid composition of the three binding loops in S60* compared to the full PH domain in S60*. Table associated to Figure 14 in chapter 3.3.1 Amino acid composition of the three binding loops L1, L2 and L3 and the full domain in S60(*). Table shows number of amino acids and their percentage. The amino acids in the property groups are hydrophobic: L, I, C, M, W, Y, F, aromatic: W, Y, F, positive: H, K, R, negative: D, E, non-polar: V, A, G, P, polar: S, N, Q, T.*

Dataset + position	S60	Full PH	S60*	Loop 1 [196, 290]	S60*	Loop 2 [333, 444]	S60*	Loop 3 [494, 556]
Total AA	19115		1269		1211		1251	
Amino acid	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
A	1249	6.53	94	7.41	61	5.04	72	5.76
G	1127	5.90	119	9.38	78	6.44	98	7.83
E	1361	7.12	80	6.30	132	10.90	90	7.19
R	1130	5.91	84	6.62	85	7.02	89	7.11
W	372	1.95	26	2.05	6	0.50	4	0.32
Y	620	3.24	13	1.02	31	2.56	21	1.68
S	1387	7.26	111	8.75	105	8.67	104	8.31
N	739	3.87	68	5.36	60	4.95	83	6.63
D	1024	5.36	80	6.30	127	10.49	127	10.15
C	379	1.98	12	0.95	6	0.50	21	1.68
Q	895	4.68	50	3.94	79	6.52	41	3.28
H	468	2.45	26	2.05	20	1.65	46	3.68
I	1036	5.42	39	3.07	26	2.15	22	1.76
L	1825	9.55	89	7.01	55	4.54	47	3.76
K	1430	7.48	141	11.11	139	11.48	118	9.43
M	346	1.81	30	2.36	13	1.07	16	1.28
F	874	4.57	32	2.52	15	1.24	29	2.32
P	700	3.66	56	4.41	69	5.70	100	7.99
T	935	4.89	59	4.65	67	5.53	66	5.28
V	1217	6.37	60	4.73	37	3.06	57	4.56
Hydrophobic	5452	28.52	241	18.99	152	12.55	160	12.79
Aromatic	1866	9.76	71	5.59	52	4.29	54	4.32
Positive	3028	15.84	251	19.78	244	20.15	253	20.22
Negative	2385	12.48	160	12.61	259	21.39	217	17.35
Non polar	4293	22.46	329	25.93	245	20.23	327	26.14
Polar	3955	20.69	288	22.70	311	25.68	294	23.50

Table A4 Amino acid composition of the two X positions in the canonical binding motif (M1), (KX(n)(K/R)XR), Loop 1 and the full PH domain in the 1M26 dataset*. Table associated to Figure 15 in chapter 3.3.2 Amino acid composition of the 1M26 dataset domains and their sequence matching M1. Table shows number of amino acids and their percentage. The amino acids in the property groups are hydrophobic: L, I, C, M, W, Y, F, aromatic: W, Y, F, positive: H, K, R, negative: D, E, non-polar: V, A, G, P, polar: S, N, Q, T.*

Dataset + position	1M26	Full domain	1M26	Motif 1 X(n)	1M26	Motif 1 X(-2)	1M26	Motif 1 Loop 1
Total AA	3045		216		26		166	
Amino acid	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
A	184	6.04	6	2.78	1	3.85	6	3.61
G	185	6.08	29	13.43	0	0.00	28	16.87
E	250	8.21	9	4.17	5	19.23	8	4.82
R	207	6.80	20	9.26	5	19.23	14	8.43
W	81	2.66	27	12.50	0	0.00	3	1.81
Y	126	4.14	4	1.85	0	0.00	4	2.41
S	172	5.65	11	5.09	2	7.69	12	7.23
N	98	3.22	8	3.70	0	0.00	8	4.82
D	165	5.42	14	6.48	0	0.00	14	8.43
C	69	2.27	0	0.00	1	3.85	0	0.00
Q	127	4.17	9	4.17	0	0.00	3	1.81
H	52	1.71	3	1.39	0	0.00	3	1.81
I	161	5.29	5	2.31	0	0.00	4	2.41
L	257	8.44	7	3.24	1	3.85	4	2.41
K	279	9.16	30	13.89	3	11.54	30	18.07
M	45	1.48	3	1.39	0	0.00	3	1.81
F	132	4.33	5	2.31	0	0.00	1	0.60
P	131	4.30	6	2.78	4	15.38	5	3.01
T	144	4.73	9	4.17	1	3.85	7	4.22
V	180	5.91	11	5.09	3	11.54	9	5.42
Hydrophobic	871	28.60	51	23.61	2	7.69	19	11.45
Aromatic	339	11.13	36	16.67	0	0.00	8	4.82
Positive	538	17.67	53	24.54	8	30.77	47	28.31
Negative	415	13.63	23	10.65	5	19.23	22	13.25
Non polar	680	22.33	52	24.07	8	30.77	48	28.92
Polar	541	17.77	37	17.13	3	11.54	30	18.07

Table A5 Amino acid composition of the X position in the non-canonical binding motif (M2), (K/R)X(W/Y/F) and the full PH domain in the 2M59 dataset. *Table associated to Figure 16 in chapter 3.3.3 Amino acid composition of the 2M59 dataset domains and their sequence matching M2. Table shows number of amino acids and their percentage. The amino acids in the property groups are hydrophobic: L, I, C, M, W, Y, F, aromatic: W, Y, F, positive: H, K, R, negative: D, E, non-polar: V, A, G, P, polar: S, N, Q, T.*

Dataset + position	2M59	full domain	2M59	Motif2 X
Total AA	6935		59	
Amino acid	Number	Percentage	Number	Percentage
A	455	6.56	0	0.00
G	444	6.40	4	6.78
E	489	7.05	6	10.17
R	420	6.06	5	8.47
W	181	2.61	0	0.00
Y	261	3.76	0	0.00
S	544	7.84	7	11.86
N	252	3.63	12	20.34
D	347	5.00	0	0.00
C	139	2.00	0	0.00
Q	308	4.44	3	5.08
H	157	2.26	1	1.69
I	339	4.89	0	0.00
L	585	8.44	3	5.08
K	592	8.54	5	8.47
M	108	1.56	1	1.69
F	308	4.44	1	1.69
P	260	3.75	2	3.39
T	340	4.90	7	11.86
V	406	5.85	2	3.39
Hydrophobic	1921	27.70	5	8.47
Aromatic	750	10.81	1	1.69
Positive	1169	16.86	11	18.64
Negative	836	12.05	6	10.17
Non polar	1565	22.57	8	13.56
Polar	1444	20.82	29	49.15

Table A6 Amino acid composition of the OM1 dataset domains. *Table shows the number of amino acids and their percentage for the full domain, Loop 1, Loop 2 and Loop 3. The table is associated to chapter 3.3.4 Amino acid composition of the three binding loops in OM1, OM2, M1and2 and S60*. The amino acids in the property groups are hydrophobic: L, I, C, M, W, Y, F, aromatic: W, Y, F, positive: H, K, R, negative: D, E, non-polar: V, A, G, P, polar: S, N, Q, T.*

Dataset + position	OM1	Full domain	OM1	Loop 1	OM1	Loop 2	OM1	Loop 3
Total AA	676		40		35		51	
Amino acid	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
A	41	6.07	3	7.50	0	0.00	2	3.92
G	43	6.36	2	5.00	4	11.43	3	5.88
E	53	7.84	2	5.00	5	14.29	5	9.80
R	49	7.25	5	12.50	4	11.43	6	11.76
W	18	2.66	2	5.00	0	0.00	0	0.00
Y	20	2.96	1	2.50	0	0.00	2	3.92
S	43	6.36	6	15.00	3	8.57	6	11.76
N	20	2.96	1	2.50	0	0.00	3	5.88
D	45	6.66	4	10.00	3	8.57	4	7.84
C	12	1.78	0	0.00	0	0.00	0	0.00
Q	37	5.47	1	2.50	2	5.71	0	0.00
H	9	1.33	0	0.00	0	0.00	1	1.96
I	32	4.73	1	2.50	0	0.00	1	1.96
L	67	9.91	2	5.00	1	2.86	3	5.88
K	59	8.73	4	10.00	4	11.43	2	3.92
M	7	1.04	0	0.00	1	2.86	0	0.00
F	30	4.44	0	0.00	2	5.71	1	1.96
P	23	3.40	4	10.00	0	0.00	4	7.84
T	28	4.14	0	0.00	1	2.86	3	5.88
V	40	5.92	2	5.00	5	14.29	5	9.80
Hydrophobic	186	27.51	6	15.00	4	11.43	7	13.73
Aromatic	68	10.06	3	7.50	2	5.71	3	5.88
Positive	117	17.31	9	22.50	8	22.86	9	17.65
Negative	98	14.50	6	15.00	8	22.86	9	17.65
Non polar	147	21.75	11	27.50	9	25.71	14	27.45
Polar	128	18.93	8	20.00	6	17.14	12	23.53

Table A7 Amino acid composition of the OM2 dataset domains. *Table shows the number of amino acids and their percentage for the full domain, Loop 1, Loop 2 and Loop 3. The table is associated to chapter 3.3.4 Amino acid composition of the three binding loops in OM1, OM2, M1and2 and S60*. The amino acids in the property groups are hydrophobic: L, I, C, M, W, Y, F, aromatic: W, Y, F, positive: H, K, R, negative: D, E, non-polar: V, A, G, P, polar: S, N, Q, T.*

Dataset + position	OM2	Full domain	OM2	Loop 1	OM2	Loop 2	OM2	Loop 3
Total AA	4590		378		403		351	
Amino acid	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
A	314	6.84	21	5.56	30	7.44	17	4.84
G	304	6.62	32	8.47	28	6.95	27	7.69
E	290	6.32	17	4.50	49	12.16	25	7.12
R	265	5.77	37	9.79	15	3.72	23	6.55
W	116	2.53	9	2.38	1	0.25	0	0.00
Y	157	3.42	4	1.06	8	1.99	6	1.71
S	417	9.08	38	10.05	45	11.17	28	7.98
N	170	3.70	19	5.03	9	2.23	23	6.55
D	226	4.92	18	4.76	38	9.43	32	9.12
C	85	1.85	1	0.26	1	0.25	8	2.28
Q	225	4.90	16	4.23	24	5.96	21	5.98
H	115	2.51	7	1.85	11	2.73	14	3.99
I	210	4.58	13	3.44	16	3.97	6	1.71
L	399	8.69	27	7.14	19	4.71	8	2.28
K	370	8.06	58	15.34	45	11.17	48	13.68
M	70	1.53	3	0.79	2	0.50	2	0.57
F	205	4.47	5	1.32	3	0.74	9	2.56
P	158	3.44	14	3.70	29	7.20	25	7.12
T	227	4.95	22	5.82	17	4.22	14	3.99
V	267	5.82	17	4.50	13	3.23	15	4.27
				0.00				
Hydrophobic	1242	27.06	62	16.40	50	12.41	39	11.11
Aromatic	478	10.41	18	4.76	12	2.98	15	4.27
Positive	750	16.34	102	26.98	71	17.62	85	24.22
Negative	516	11.24	35	9.26	87	21.59	57	16.24
Non polar	1043	22.72	84	22.22	100	24.81	84	23.93
Polar	1039	22.64	95	25.13	95	23.57	86	24.50

Table A8 Amino acid composition of the M1and2 dataset domains. *Table shows the number of amino acids and their percentage for the full domain, Loop 1, Loop 2 and Loop 3. The table is associated to chapter 3.3.4 Amino acid composition of the three binding loops in OM1, OM2, M1and2 and S60*. The amino acids in the property groups are hydrophobic: L, I, C, M, W, Y, F, aromatic: W, Y, F, positive: H, K, R, negative: D, E, non-polar: V, A, G, P, polar: S, N, Q, T.*

Dataset + position	M1and 2	Full domain	M1and 2	Loop 1	M1and 2	Loop 2	M1and 2	Loop 3
Total AA	2369		134		212		186	
Amino acid	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
A	143	6.04	3	2.24	10	4.72	10	5.38
G	142	5.99	26	19.40	5	2.36	14	7.53
E	197	8.32	9	6.72	28	13.21	18	9.68
R	158	6.67	12	8.96	19	8.96	16	8.60
W	63	2.66	3	2.24	0	0.00	0	0.00
Y	106	4.47	3	2.24	14	6.60	7	3.76
S	129	5.45	4	2.99	6	2.83	9	4.84
N	78	3.29	6	4.48	2	0.94	18	9.68
D	120	5.07	8	5.97	21	9.91	16	8.60
C	57	2.41	0	0.00	4	1.89	2	1.08
Q	90	3.80	0	0.00	8	3.77	6	3.23
H	43	1.82	3	2.24	1	0.47	3	1.61
I	129	5.45	4	2.99	4	1.89	4	2.15
L	190	8.02	2	1.49	15	7.08	5	2.69
K	220	9.29	29	21.64	24	11.32	16	8.60
M	38	1.60	3	2.24	2	0.94	3	1.61
F	102	4.31	3	2.24	5	2.36	0	0.00
P	108	4.56	2	1.49	25	11.79	19	10.22
T	116	4.90	7	5.22	12	5.66	10	5.38
V	140	5.91	7	5.22	7	3.30	10	5.38
Hydrophobic	685	28.92	18	13.43	44	20.75	21	11.29
Aromatic	271	11.44	9	6.72	19	8.96	7	3.76
Positive	421	17.77	44	32.84	44	20.75	35	18.82
Negative	317	13.38	17	12.69	49	23.11	34	18.28
Non polar	533	22.50	38	28.36	47	22.17	53	28.49
Polar	413	17.43	17	12.69	28	13.21	43	23.12