Using peptide-phage display to capture conditional motif-based interactions

GUSTAV SUNDELL
Abstract


This thesis explores the world of conditional protein-protein interactions using combinatorial peptide-phage display and proteomic peptide-phage display (ProP-PD). Large parts of proteins in the human proteome do not fold into well-defined structures instead they are intrinsically disordered. The disordered parts are enriched in linear binding-motifs that participate in protein-protein interaction. These motifs are 3-12 residue long stretches of proteins where post-translational modifications, like protein phosphorylation, can occur changing the binding preference of the motif. Allosteric changes in a protein or domain due to phosphorylation or binding to second messenger molecules like Ca$^{2+}$ can also lead conditional interactions. Finding phosphorylation regulated motif-based interactions on a proteome-wide scale has been a challenge for the scientific community.

Through combinatorial peptide phage display selections against the calcium binding protein Sorcin we discovered the Φ-Φ-X-P motif that preferentially interacts with calcium bound Sorcin. Through C-terminal ProP-PD we discovered that the engineered L391F, K329M Ø9 mutant of the second PDZ domain of DLG1 has a reduced specificity and increased affinity towards proteomic targets as compared to the wild type DLG1 PDZ2 domain.

Finally a novel application of ProP-PD that allows us to capture phospho-regulated interactions by combining it with glutamic phosphomimetic substitution is presented. We created a phosphomimetic ProP-PD library of C-terminal sequences of known and putative serine/threonine phospho-sites, and their phosphomimetic counterparts. The library was used in selection against PDZ 1, 2 and 3 of Scribble and DLG1. We found site-specific phosphorylation events that either enable or disable interactions. Selectivity towards phosphopeptide binding was confirmed for Scribble PDZ1 through biophysical experiments, supporting the viability of the phosphomimetic ProP-PD. Through structural and mutational analysis we found that the gate-keeping residue for phosphopeptide binding in Scribble PDZ1 is arginine 762. With this proof-of-concept study we have introduced phosphomimetic ProP-PD as a viable method to discover phospho-regulated protein-peptide interactions.

Taken together, this work has contributed with methods to capture selectivity differences in motif-based interactions of proteins depending on allosteric binding of calcium, protein engineering and on ligand modification by phosphorylation.

Keywords: Protein-protein interaction, PDZ domain, phosphorylation, peptide-phage display, ProP-PD, phospho-peptide, phosphomimetic substitution, Scribble, NMR, DLG1, PTM, affinity

Gustav Sundell, Department of Chemistry - BMC, Biochemistry, Box 576, Uppsala University, SE-75123 Uppsala, Sweden.

© Gustav Sundell 2018

ISSN 1651-6214
urn:nbn:se:uu:diva-359434 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-359434)
Till mamma
12 februari 1960 - 7 december 2011
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Reprints were made with permission from the respective publishers.


## Contents

Introduction ................................................................................................... 11  
Proteins and intrinsic protein disorder.......................................................... 13  
Protein-protein interactions ........................................................................ 14  
Posttranslational modifications and phosphorylation ..................................... 16  
Scaffold proteins and peptide binding protein domains ................................... 18  
PDZ domains .............................................................................................. 19  
  PDZ domains bind C-terminal motifs in proteins ........................................ 21  
  Phosphorylation of PDZ binding motifs ............................................... 22  
Discovering novel protein-protein interactions ............................................... 23  
  Proteome-wide protein-protein interaction discovery ............................ 23  
  Experimental methods for finding PPIs based on linear binding-motifs ...... 24  
  Phage display ....................................................................................... 27  
  Discovering interactions regulated by phosphorylation ........................ 27  
  Next-generation sequencing in relation to large-scale protein-protein interaction discovery ........................................................... 30  
  Biophysical validations of protein peptide interactions ............................... 32  
  Isothermal titration calorimetry .............................................................. 34  
  Protein NMR for structure determination .............................................. 35  
  Defining specificity and selectivity in terms of differences in affinity and binding-motifs ................................................................. 37  

Present investigations .................................................................................... 39  
  Paper I Finding binding motifs of the calcium activated protein Sorcin using combinatorial peptide-phage display .............................................. 41  
  Paper II Exploring the specificity of a PDZ domain engineered for high affinity interaction with the viral oncoprotein E6 ............................... 45  
  Paper III & IV Proteome-wide analysis of phospho-regulated PDZ domain interactions ................................................................. 49  
    Part A Method development and validation ........................................... 49  
    Part B structural details of PDZ phosphopeptide binding paper III and IV ................................................................. 63  
    Part C Expanding the phosphomimetic ProP-PD to more PDZ domains ................................................................. 66  

Concluding remarks ...................................................................................... 71
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDZ</td>
<td>Postsynaptic density protein-95/Disc large tumor suppressor/Zonula occludence-1</td>
</tr>
<tr>
<td>PPI</td>
<td>Protein-Protein Interaction</td>
</tr>
<tr>
<td>PTM</td>
<td>Post-Translational Modification</td>
</tr>
<tr>
<td>PDZbm</td>
<td>PDZ binding motif</td>
</tr>
<tr>
<td>ELM</td>
<td>Eukaryotic Linear Motif database</td>
</tr>
<tr>
<td>ProP-PD</td>
<td>Proteomic Peptide-Phage Display</td>
</tr>
<tr>
<td>NGS</td>
<td>Next generation sequencing</td>
</tr>
<tr>
<td>Y2H</td>
<td>Yeast two hybrid</td>
</tr>
<tr>
<td>MS</td>
<td>Mass spectrometry</td>
</tr>
<tr>
<td>HEK</td>
<td>Human Embryonic kidney cells</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear magnetic resonance</td>
</tr>
<tr>
<td>HSQC</td>
<td>Heteronuclear single quantum correlation</td>
</tr>
<tr>
<td>NOE</td>
<td>Nuclear Overhauser effect</td>
</tr>
<tr>
<td>FACS</td>
<td>Fluorescence-activated cell sorting</td>
</tr>
<tr>
<td>SH2</td>
<td>Src Homology 2</td>
</tr>
<tr>
<td>SH3</td>
<td>Src Homology 3</td>
</tr>
<tr>
<td>ITC</td>
<td>Isothermal titration calorimetry</td>
</tr>
<tr>
<td>MST</td>
<td>Microscale thermophoresis</td>
</tr>
<tr>
<td>WT</td>
<td>Wild type</td>
</tr>
<tr>
<td>Mim</td>
<td>Mimetic</td>
</tr>
<tr>
<td>PWM</td>
<td>Position weighed matrix</td>
</tr>
<tr>
<td>ELISA</td>
<td>Enzyme-linked immunosorbent assay</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein data bank</td>
</tr>
<tr>
<td>HPV</td>
<td>Human papillomavirus</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
</tbody>
</table>
Introduction

Imagine a cell.

In a cell there are three major types of molecules, which make it in to the cell that we imagine. There are nucleic acids in the form of DNA and RNA, there are lipids that form membranes surrounding the cell and parts inside the cell called organelles, and there are proteins. Proteins are what get things done in a cell, they are the key players in everything think of as a task that the cell does. The central dogma of molecular biology (Figure 1) describes that genes in the DNA gets transcribed in to RNA, and RNA in turn gets translated into proteins. DNA can also copy itself, and to a lesser extent RNA can copy itself and reversely transcribe itself in to DNA, but all of these tasks needs proteins to work. There are roughly 20 000 genes in the human genome giving rise to about 70 000 different protein isoforms \(^1\). All of these proteins have a function in the cell. Many of them catalyze chemical reactions i.e. putting molecules together or breaking them apart. This can be for the purpose of making energy, building up the structures and molecules that the cell needs, modifying proteins to convey a signal or physically moving thing. These proteins are enzymes, and they might be the stars of the show, but they will not be the stars of my thesis. Instead, I will focus on proteins that have until the last 20 years been somewhat overlooked. The proteins of this thesis perform no catalysis; instead they bring other proteins together by forming specific interactions with the proteins that need to be in close proximity to each other to perform their desired function \(^1\). These interactions are called protein-protein interactions or PPIs, and the proteins are called scaffold and adaptor proteins \(^2\).

\(^1\) There are of course other proteins in the cell that perform no catalysis e.g. passive transporters, proteins that act as reservoirs for certain molecules, and a lot of cytoskeleton proteins etc.
Figure 1. The central dogma of molecular biology. The genes in the DNA are transcribed into a certain type of RNA namely mRNA. mRNA in turn gets translated into proteins by ribosomes. DNA can also be replicated to pass on the genetic information to daughter cells. Some viruses can also replicate RNA and transcribe RNA to DNA, in a process called reversed transcription.

My thesis is about finding PPIs that are based on interactions between one protein and a short linear stretch of amino acid residues in a target protein, called motifs. The thesis is further about exploring the relationship between affinity and selectivity in the context of interactions involving these short linear motifs. The majority of the research is focused on finding such interactions, which are conditional. To better understand the research presented in this thesis I will first introduce some key concepts in biochemistry, molecular and systems biology, and techniques that have been utilized in producing these findings.
Proteins and intrinsic protein disorder

The genetic code of DNA codes for 20 distinct amino acids (Figure 2). In the process of making a protein (translation) the amino acids get coupled together by peptide bonds between the carboxylic and the amine part of the amino acid in to polypeptide chains. Proteins are translated from the N-terminal to the C-terminal and when a protein sequence is depicted it is always written in this direction. Amino acid residues can be written with three-letter abbreviation, Ser for Serine, or with a one-letter code, S for Serine (Figure 2).

**Small**

- Glycine (Gly) G
- Serine (Ser) S
- Threonine (Thr) T
- Glutamine (Gln) Q
- Asparagine (Asn) N

**Polar**

- Histidine (His) H
- Lysine (Lys) K
- Arginine (Arg) R
- Glutamate (Glu) E
- Aspartate (Asp) D

**Basic**

- Proline (Pro) P
- Cysteine (Cys) C
- Tryptophan (Trp) W
- Phenylalanine (Phe) F
- Tyrosine (Tyr) Y

**Hydrophobic**

- Alanine (Ala) A
- Valine (Val) V
- Leucine (Leu) L
- Isoleucine (Ile) I
- Methionine (Met) M

Figure 2. The 20 canonical amino acids, with their names indicated and the three-letter and one-letter code. The amino acids are divided into groups of similar chemical properties.
While being translated a protein chain fold into secondary structures (α-helical and β-sheets) and tertiary structures. Since the discovery of protein structure in 1958 the consensus was that a protein has to fold in to a defined structure to perform its functions. However, in the 1990th researchers started to challenge this structure-function dogma as more and more evidences were gathered that large parts of proteins or whole proteins can be intrinsically disordered, meaning that a protein does not necessarily fold in to a defined three dimensional structure to function but instead may sample many different conformations. These regions that are intrinsically disordered have special properties compared to proteins that fold in to globular proteins. Disordered regions tend to be composed of more charged and polar amino acid residues, where P, E, S, Q and K have a strong correlation with intrinsic disorder but L, F, Y, I, W and C are negatively correlated with disorder. If you measure the net charge and mean hydrophobicity of proteins and plot those properties perpendicular to each other, the disordered proteins and regions will fall in to the area where the hydrophobicity is lower than 0.4 for uncharged polypeptides up to a mean hydrophobicity of closer to 0.5 for polypeptides with a higher mean net charge. Today, more than 30 % of the human proteome is estimated to be disordered, and disordered regions are found in a wide variety of proteins. The evolutionary rate of disordered regions is generally faster than for structured domains. Short stretches of evolutionary conservation in disordered regions may be indicative of some kind of function, as will be further elaborated below. The disordered parts of proteins can be situated in between parts of proteins that fold into globular structures called protein domains, these domains can in many cases be isolated and fold and function independently of its parent protein. These domains are usually between 35 and 250 amino acid residues long. The domains can have different functions e.g. catalytic, protein binding, regulating activity and they can associate with membranes etc.

Protein-protein interactions

Proteins interact with each other in an ordered, specific manner. One of the first protein structures reported, hemoglobin made this clear. The structure revealed that the α and β chains of hemoglobin interact in a specific manner to form a tetrameric structure. However, protein interactions were known even before the advent of protein structures. Already in 1922 an interaction between the protease trypsin and an inhibitor from the albumin fraction were shown to interact “to form an inactive but dissociable compound”. Even before that, in the late 19th century the discovery of neutralizing antibodies should be noted. However, it was with the revolution of x-ray crystallography that scientists could start to investigate protein-protein interactions on
a molecular level. The first structures of PPIs were of globular structures that interact with each other through large surface interface. There are many examples of PPIs that form stable complexes of high affinity that are unlikely to break apart during the lifespan of the protein. These interactions may form during translation of one or both proteins\(^\text{20,21}\). Many of these complexes are homomers meaning that the same protein interacts with one or more copies of itself. A study from 2006 found that 41% of structures involving PPIs, in the protein data bank (PDB), were structures of homomeric proteins complexes\(^\text{22}\) and most of the homomers are dimers\(^\text{23}\). Estimations of the numbers of large protein complexes land around a thousand instances\(^\text{24}\). Even though protein complexes and globular domain interactions are important in the cell, they are not nearly enough to carry out the dynamical aspect of many cellular processes. Such aspects require interactions that are more transient in nature and that are conditional.

Cell signaling is the cell’s reaction upon stimuli. These stimuli can for instance be a signal from outside the cell in the form of a ligand binding to a receptor that spans the cell membrane. The signal is transferred through the receptor and will be passed down in the cell through signaling pathways leading for instance to motility, exocytosis, endocytosis and differential transcription of genes. In cell signaling it is important that PPIs are dynamic, so they can form and break apart upon signal propagation. One way of achieving dynamic interactions is by utilizing linear motifs. As stated above, there are islands of evolutionary conservation in disordered parts of proteins that are likely to serve specific functions, as otherwise they would not be conserved\(^\text{25,26}\). Among the functions are the linear binding motifs, where a stretch of 3-12 amino acid residues part severs as a partner for a protein-protein interaction\(^\text{27}\). These binding motifs consist of specific combinations of amino acid residues in certain positions and positions that are less restricted in what residue they contains. The fact that these linear motifs are so short and can be flexible in composition in some of the positions makes them abundant in the proteome, which makes them hard to identify bioinformatically\(^\text{28}\). The fact that disordered regions have faster evolution also makes it possible to create new functional binding motifs ex-nihilo\(^\text{29}\). There are many examples residue stretches that contain what looks like a binding motif that is not functional. This is due to its location in the protein, what flanking residues it has, or which protein it is in. These random occurrences of motif-like regions tend to not be evolutionary conserved\(^\text{30}\).

Motif-based PPIs are usually of moderate affinity, with \(K_D\) values in the range of 1-150 \(\mu\text{M}\)\(^\text{31}\). The moderate affinities can allow the interactions to be temporary, associate slower and dissociate more rapidly than interactions with higher affinity, allowing for other proteins to bind or posttranslational modifications to occur. This is important for cell signaling where multi-protein complexes rapidly form and break apart\(^\text{27}\). Motif-based PPIs can readily be evolved experimentally towards high affinity (1-100 nM)\(^\text{32}\).
In many instances two or more repeats of the same linear motif is present in the same protein and many proteins have more than one type of linear motif present in their sequences, meaning it can bind to several different peptide binding domains. In signaling complexes, linear motifs often interact with modular domains of multi-domain containing proteins such as scaffold and adaptor proteins (Figure 4). This gives rise to the possibility of multivalent interactions between proteins in a protein complex, which may keep a complex together.

How common are these motifs? Tompa et al. estimated in a review paper in 2014 that the number of linear binding motifs in the human proteome should be in the order of 100,000 to 130,000 based on its occurrence in well studied proteins. The eukaryotic linear motif resource (ELM) is a database of linear motifs that are manually curated to include motifs and instances of high quality. It gathers linear binding motifs as well as motifs for post-translational modifications (PTMs). In its latest update it lists 275 types of linear motifs (called in ELM motif classes). Compared to the estimated 130,000 instances of linear binding motifs (and even more PTMs see below) only 3091 curated instances is yet annotated in the database. Since ELM is manually curated and ‘ELM instances’ requires external published experimental evidence there is a bottleneck between experimentally verified motifs and a motif becoming an ELM instance. But the number of instances that are curated tell us something about the work there is yet to be done in terms of finding both novel motifs and functional instances of those linear motifs. Since linear motifs can easily evolve in to or out of existence this also happens in normal cells in the human body when DNA gets randomly mutated. Therefore linear motifs are important in disease, 22% of all disease mutations occur in disordered regions. Hijacking of peptide-protein interactions by infectious diseases especially viruses is also prevalent, and when this happens the viral motif is usually of higher affinity than the normal host interactions of that peptide binding protein.

Posttranslational modifications and phosphorylation
A way to regulate binding, and thereby signaling is by PTMs. PTMs are chemical modifications on specific amino acid residues performed by enzymes. There are more than 300 different posttranslational modifications serving a vast array of purposes. PTMs can for instance be small modifications like adding a methyl, acetyl or phosphate on an amino acid; the modification can be a sugar, which can be subsequently expanded it to larger carbohydrates; and it can even be whole proteins, like ubiquitin or sumo-domains. Some PTMs found will serve no apparent purpose because of the low specificity for some of the enzymes catalyzing PTMs, but many of them
serve distinct functions. Some aid in protein stability and structure, some open up proteins for new binding interfaces or activation of catalysis and others will create or modify existing protein binding interfaces, both between structured domains and between linear binding motifs and their binding partners. The PTM can either be directly required for specificity like serine/threonine phosphorylation of peptides binding to 14-3-3 proteins, or it can enhance or decrease the affinity between a linear motif and its binding partner although it is not required for specificity. Tompa et al. made a conservative estimation using the PTM information of well-studied proteins. This resulted in an estimation that PTMs as common as to appear on average on every 10th residue in the proteome, bringing the number of potential PTM sites up to a million. The most abundant PTM in terms of found experimentally verified instances is phosphorylation. Phosphositeplus.org lists as of 2 July 2018 as many as 235 407 phosphorylation sites in human gathered from experimental data. In eukaryotic cells canonical protein phosphorylation can be carried out on serine, threonine or tyrosine residues by protein kinases and dephosphorylation by protein phosphatases (Figure 3). Kinases either phosphorylate serine and threonine residues (serine/threonine kinase) or tyrosine residues (tyrosine kinase). There are more than 518 kinases in the human proteome that perform protein phosphorylation, and the vast majority of them perform serine/threonine phosphorylation (428) and the resulting 90 perform tyrosine phosphorylation. In contrast there are only approximately 200 protein phosphatases, where the majority dephosphorylates tyrosine residues. Kinases are more or less specific in what sequences they phosphorylate where Cycline dependent kinases favor proline at the first position after the serine/threonine residue and RPS6KA kinases for instance prefer an arginine three amino acid residues before the serine/threonine. Most kinases require upstream phosphorylation to be activated. Already in 1956 glycogen phosphorylase kinase was found to be regulated by phosphorylation and in 1991 it was discovered that this activation was due to protein-protein interaction forming a tetramer of the protein upon phosphorylation. Many driver mutations in cancer are affecting kinases upstream in signaling pathways like receptor tyrosine kinases and Ras, which leads to hyperactive kinases and downstream activation of secondary kinases and therefore increased protein phosphorylation.
Figure 3. Protein phosphorylation in eukaryotes is predominantly O-linked and therefore occurs on serine, threonine and tyrosine residues.

Scaffold proteins and peptide binding protein domains

Scaffold proteins, anchoring proteins and adaptor proteins are abundant in countless processes in the eukaryotic cell. In signaling, scaffolding proteins are building up a physical network bringing together enzymes and targets involved in signal transduction. These proteins can both bring together proteins that propagate a signal e.g. kinases, acetyltransferases et cetera, and their targets, or to cancel signals by bringing together phosphatases or deacetylases, and their targets. Scaffold proteins achieve this by utilizing protein and peptide binding domains and linear binding motifs. As you can see in figure 4 scaffold proteins contains several different domains and these domains are common for several of the proteins. Many of the proteins also have several copies of the same domain in the protein. Scaffold proteins also have large areas of intrinsically disordered regions, as discussed above these regions are rich in both PTM sites and linear binding motifs. In the ELM resource you can search for predicted instances and classes of linear motifs. If we employ that for the scaffold protein Disk large 1 (DLG1) we find that it contains 6 structured domains which binds to peptides or proteins, and
in terms of binding motifs in the protein we find 55 motif classes and 142 instances.

The first peptide-binding domains discovered was the SH2 domain, which stands for Src homolog 2, and SH3 domain. In 1988 Tony Pawson was studying cytosolic protein tyrosine kinases from human and from viral pathogens. He found sequence similarities between these proteins in three different regions of the sequences. The regions were named SH1, SH2 and SH3 domains. The catalytic kinase domain was SH1, but the SH2 and SH3 domains were non-catalytic and therefore had other roles for the protein. It was later determined that deleting the SH2 domain made the protein lose their ability to phosphorylate targets. 1993 it was proven that SH2 domains recognize specific phosphotyrosine containing peptides making it the first domain reported to bind linear binding motifs, although several motifs were known before that but not what they bound to. The SH2 domain family contains 120 domains in 110 proteins and for a detailed representation of the binding motifs of different SH2 domains see Tinti et al. 2013. In 1994 it was discovered that SH3 domains bind proline-rich peptide sequences. There are 296 SH3 domains distributes over 217 proteins in the human proteome, for detailed description of the specificities of SH3 domains see Xin et al. 2013. Since the early nineties several other peptide binding domains like WW, PTB and 14-3-3 domains have been discovered binding different linear binding motifs.

PDZ domains

The main focus of this thesis is a special class of peptide binding domains called PDZ domains. The name is derived from three proteins that contains the domain namely Postsynaptic density protein-95, Disc large tumor suppressor, and Zonula occludence-1. The domain was first described as a GLGF repeat in 1992 based on the sequence GLGF conserved in PDZ domains of rat PSD-95, subsequently renamed DHR domain to end up being named PDZ domains in 1995. These domains are common features of scaffold proteins (Figure 4). In the human proteome there are 269 PDZ domains in 154 proteins. PDZ domains commonly consists of 80 to 90 amino acid residues with a common fold of five to six β sheets and 2 α helices (Figure 4B). Additionally, the domains contain loop regions that can be of longer or shorter lengths, especially the loop between the second and third β-sheet can vary making some domains to become over 110 amino acid residues long. Almost all PDZ domains in eukaryotes have the normal domain architecture of PDZ domains that are β1-β2-β3-α1-β4- (β5)-α2-β6 (Figure 4B) except for some special instances, the Golgi reassembly stacking protein 1 and 2 have a special tandem PDZ domain that has the secondary structure of β3-α1-β4-β5-α2-β6-β1-β2.
Figure 4. A Common adaptor and scaffold proteins containing PDZ domains along side other protein and peptide binding domains. B Structure of the common fold of a PDZ domain where the binding pocket is situated between $\beta2$ and $\alpha2$. 
PDZ domains bind C-terminal motifs in proteins

The canonical binding motifs for PDZ domains (PDZbm) are C-terminal peptides. When describing the sequence of C-terminal binding, the convention is to assign the C-terminal residue position 0 and the immediate residue before that -1 and so on. The binding of PDZ domains to specific C-terminal peptides was conclusively proven in a study by Kim et al. 1995 where they showed by selectively mutating each residue in the C-terminal of K+ channel Kv1.4 (current gene name KCNA4), one by one to alanine (a so-called Ala scan) that alanine substitutions at p0 and p-2 of the peptide AVETDV COO- broke the interaction. In this study they showed the domain binding to other C-terminal peptides with an S/T-X-V COO- consensus motif (where X is any amino acid residue). The first structure of a PDZ domain bound to a C-terminal peptide was reported in 1996. In this study the third PDZ domain of DLG4 (also known as PSD-95) was crystalized with a nine residue long peptide TKNYKQTSV COO-, which revealed the site of the peptide binding-pocket to be situated in the groove between β2 and α2, and that the peptide binds as an antiparallel β-sheet to the β2 β-sheet. The carboxylic part of the terminal valine interacts with the backbone of the domain and, via a water molecule, the R chain of an arginine in the carboxylate-binding loop situated between β1 and β2. This loop contains the sequence GLGF for which PDZ domains were originally named. The consensus sequence in PDZ domains of the carboxylate binding loop is R/K-XXX-G-Φ-G-Φ, where Φ is a hydrophobic residue and X is any amino acid 68,70. The R part of the valine is buried in a hydrophobic pocket. The threonine in the consensus sequence S/T-X-V COO- is hydrogen bonding to a histidine in the beginning of the α2 helix, which is conserved in PDZ domains binding this consensus sequence.

The early evidence for C-terminal binding was reviewed by Saras and Heldin from Ludwig institute of cancer research Uppsala in 1996 summarizing C-terminal binding of seven PDZ containing proteins by eleven C-terminal peptides. It was by then shown that the C-terminal is not exclusively selective for valine but tolerates leucine and isoleucine shifting the consensus sequence to X-S/T-X-Φ COO-. This binding motif is conventionally referred to the class I PDZ binding motif (PDZbm). The class II binding motif X-Φ-X Φ COO- was established in 1997. Among PDZ domains accommodating class II peptides there is no conservation of the histidine in the α2, instead that pocket is lined with hydrophobic residues. The PDZ domain of Cask was the first PDZ domain binding to class II peptides and it has a valine in the histidine position of class I binding domains. There are class II binding domains that still have the histidine though, and some domains can recognize both class I and class II motifs. An additional class of recognition class III consists of consensus motif X-D/E-X-Φ COO-, which was found by the determining of the binding preferences for the PDZ domain of
neuronal NOS where the preference for aspartic acid is determined by a tyrosine situated in the α2 helix. Subsequent larger studies of several PDZ domains have revealed that the consensus sequences of the classes can be further divided in the more specific motifs for the residues at p-1, p-3 down to p-7 for some domains. PDZ domains can interact with other PDZ domains in a non C-terminal fashion and it can also bind to internal peptide sequences, but those questions are outside the scope of this thesis.

**Phosphorylation of PDZ binding motifs**

At the position -2 in the class I PDZbm resides a serine or threonine. Phosphorylation of this residue in endogenous targets of PDZ domains been shown to disrupt the PDZ-peptide interaction. For example phosphorylation of p-2 or p-1 of the PDZbm of the serotonin 5-HT<sub>c2</sub> disrupts the interaction with the MPDZ PDZ10. P-2 phosphorylation has also been shown to disrupt the interaction between PDZ domains of DLG4 and the NMDA receptor. Using synthetic combinatorial peptides in an peptide array Boisguerin et al. showed that PDZ domains of Erbin, SNTA1 and Afadin were disrupted by phosphorylation at p-1 and p-2 in the PDZbm.

It is not however always true that phosphorylation decreases the interaction with a PDZ domain. Pangon et al. showed that phosphorylation at p-1 of the PDZbm of the protein MCC was favored in its interaction with PDZ domains of Scribble. Furthermore p-3 phosphorylation of the PDZbm of ABCC2 increased its association with NHERF1. However neither of these two studies proved that the increased association in their cell based assay was due to increased affinity towards the phosphorylated PDZbm. In 2010 Shepherd et al. showed that p-1 phosphorylation of a tyrosine residue in the PDZbm of Syndecan moderately increased the affinity with TIAM-1 PDZ domain. The PDZ domain was later co-crystalized with the phosphorylated peptide finding the phosphate group interacts with a lysine in the loop between β3 and α1. Recently Clairfeuille et al. showed in a detailed biophysical study relying on affinity measurements and co-crystallization that the PDZ domain of sorting nexin 27 (SNX27) interacts stronger after phosphorylation of a number of C-terminal peptides. Phosphorylation of p-6 serine in the PDZbm of β2-adrenergic receptor, at p-5 serine in the PDZbm of 5-HT<sub>4</sub> receptor, and at p-3 serine of LLC3B all increased the affinity between SNX27 and the ligands. From the crystal structures it was determined that the phosphate groups of the p-5 and p-3 interact with an arginine in the lower end of the β2 sheet in the PDZ domain. Finally, Gógl et al. recently published a paper where they solved a crystal structure of the RPS6KA1 C-terminal peptide with serine phosphorylation at p-3 in complex with Magi-1 PDZ2 and show that phosphorylation have miniscule weakening effect on the affinity.
Discovering novel protein-protein interactions

Now I have described what protein-protein interactions and linear binding motif-based interactions are in relation to scaffolding proteins, cell signaling and modular peptide binding domains, especially PDZ domains. But how are PPIs discovered? For a long time most PPIs were discovered in “low-throughput”. This means that one or a couple of interactions to a protein was reported at a time when cell and molecular biologists found a favorite protein and investigated it in detail. Typical methods employed were, pull-downs, co-immuno precipitation, co-localization in cells and through bio-physical and structural studies of the interactions. This is very valuable, high quality research and it should still continue to be performed. However as a first line of discovery research for finding interactions it is not the best option. Not just because of the time required, but also because the approach created biases on the level of which proteins that are being investigated. In 2014 Rolland et al. published a comprehensive literature review of binary protein-protein interactions that had been reported in databases and in literature before 2013. By that time ~33 000 such interactions were found. The majority of interactions had just been reported using a single method, and only 1/3 of the interactions were supported by multiple pieces of evidence.

Proteome-wide protein-protein interaction discovery

One way of finding PPIs on a larger scale is through affinity purification coupled to mass spectrometry (AP-MS) where you grow a human cell-line or use tissue and then lyse the cells and do a pull-down using a specific antibody against a bait protein, proteins associated with the bait proteins will if the interaction is strong enough be pulled down with the bait protein. After pull-down the product is digested enzymatically and analyzed by MS to determine the constitution of the pull-down. If you have a way to tag a substantial amount of proteins in the cell or have specific antibodies for a large number of baits there is a possibility of capturing many interactions. This was done by two studies that came out in the fall of 2015; BioPlex where they used 7668 Flag-tagged bait-proteins and found 23 744 interactions in HEK 293T cells; and QUBIC where they used GFP tagged bait-library, analyzing 5457 bait proteins finding 28 780 interactions in HeLa cells. Another way of utilizing MS for protein-protein interactions is to fractionate cell lysate, by for example ion exchange chromatography, and then analyze what proteins are in which fraction via MS, this was done by Wan et al. in 2015 where they found 16 468 interactions. What is important to note is that protein-protein interactions captured these ways are not to be considered binary interactions but interactions in a complex. There will be proteins that are pulled down or fractionated with a bait that is not physically interacting.
with the bait but with another protein in the complex. They can also just be a frequent flyer, of which there are many as evident by the CRAPome 95.

Another way to study protein-protein interactions on a large scale is yeast-two-hybrid (Y2H). Y2H is a method that captures binary PPIs in the host of yeast cells. The method works by splitting a transcription factor, necessary for transcribing an essential gene for cell survival, between the DNA binding domain and the activation domain. The split domains are then fused to whole proteins, this can be done in a library fashion where the domain is fused to several proteins in different vectors 96. The proteins containing the DNA binding fusion and the protein that contain the activation domain are then pairwise combined in yeast cells. If the bait proteins interact with each other, the domains of the transcription factor are able to interact with each other and transcribe the gene and the cell survives. Colonies of surviving yeast can then be sequenced to see what proteins were retained and therefore must have interacted binary. When utilizing larger libraries this can be pooled and sequenced by next generation sequencing, this was done by Rolland et al. 2014 in the same publication described above. They studied 4297 proteins and found 13 867 pairs 90.

In a review comparing the four studies described above, a very limited overlap was found in terms of interaction pairs 97, indeed only 16 interactions were found in all studies and the majority of the interactions found for each study was only found in that particular study. All these studies of course have various systems of validation both experimentally and bioinformatically, which suggests that even if the overlap is small many of the interactions found in only one study will still, be correct. Neither Y2H nor AP-MS captures linear binding motif mediated interactions very well, due to the low affinity of the interactions. Even if an interaction found with these methods were the result of a protein binding to a linear binding motif, the method itself can not determine that 89. AP-MS can be extended to capture more low affinity interactions either by chemically crosslinking the proteins 91 or by a proximity biotenylation approach 98. This however does not circumvent the problem of finding the motifs; such information can sometimes be extracted bioinformatically.

Experimental methods for finding PPIs based on linear binding-motifs

**Peptide micro array**

Peptide microarrays Involve the synthesis of multiple copies of peptides on a solid support, followed by incubation of the array with a bait protein and thereafter detect if the protein bound by antibodies or fusion proteins that are radioactive or fluorescently labeled (Figure 5A) 89. Normal peptide microarrays have a couple of thousand peptides/chip but method development in
arrays have yielded ultra-dense microarrays with sizes up to $10^6$ unique peptides\textsuperscript{99}. With this one can either incorporate proteomic sequences in the array or combinatorial sequences. To map binding specificities of peptide binding domains combinatorial approach is the most straightforward and you can systematically vary each position for all amino acids, and you can readily do alanine scanning as described above. Not only do you get information of binding but also what does not bind (an advantage compared to phage display described below). An additional advantage of peptide microarray is that you can by synthesis incorporate modified amino acids such as phospho-Ser/Thr/Tyr\textsuperscript{82} or acetylated or methylated lysines\textsuperscript{100}. The main disadvantage of peptide arrays is the cost of peptide synthesis, which limits the scope of either peptides or domains tested. A additional disadvantage is that is not simple to get high quality peptide incorporation over the whole array\textsuperscript{89}.

**Y2H**

There are examples where peptides, rather than full proteins, have been used in yeast two hybrid. The method has been used to map peptide protein interactions either by randomized libraries, e.g. where it discovered the motif L-X-C-X-E motif of retinoblastoma protein\textsuperscript{101}, or by proteomic sequences, e.g. where a large portion of human PDZ domains were mapped against the c-termini of HPV E6 oncoviral protein\textsuperscript{102}. The major drawback of this method is a large false negatives due to the assay having problems of finding interactions in this affinity range\textsuperscript{103}. Another drawback is the lack of possibility to incorporate PTMs.

**Yeast surface display**

There is a possibility to utilize yeast to find PPIs in a different way with yeast surface display. Yeast surface display utilizes a library incorporated in yeast that displays on the yeast surface rendering up to 50 000 copies on the surface. Incubating the yeast with fluorescently labeled proteins one can use fluorescence-activated cell sorting FACS to distinguish yeast that has bound proteins (Figure 5B). An advantage of this technique compared to phage display is the ability to sort cells of different intensity giving the ability to distinguish yeast with high occupancy (strong binding) and moderate occupancy (weaker binding) of fluorescent protein\textsuperscript{104}. A limitation compared to phage display is the yeast library size where it can generate a diversity of $10^7$ can be generated compared to phage display with $10^{10}$ or higher\textsuperscript{89}. 

---

25
Figure 5. A Peptide micro array. Peptides are synthesized covalently to a surface. Protein in solution is allowed to bind unbound protein is washed away and by using antibodies specific to the protein you detect which peptides the protein has bound to. B Yeast surface display. Display of proteins or peptides on the surface of yeast cells. Proteins free in solution can bind to the displayed peptides and the proteins are then detected via antibodies attached to fluorophores. The yeast is then sorted using FACS. C and D Phage display. Proteins or peptides are monovalently or multivalently displayed on the phage surface, the phages are then panned over proteins or peptides immobilized on a surface. Unbound phages are washed away and bound phages are eluted, amplified and sequenced.
Phage display

Phage display is a method where a DNA library coding for proteins or peptides is inserted into the DNA of a phage or a phagemide so that the encoded sequences gets displayed on the surface of bacteriophages. This is done by fusing the DNA of the desired displayed product to one of the ends of the coat proteins of the phage. Phage or bacteriophages are viruses that infect and amplify in bacteria. In terms of exit from the bacteria there are two distinct ways a phage achieves this, for instance the termophilic phages T4 and T7 along with phage λ exit the cell by lysis of the bacteria, in contrast the filamentous phages like M13 phage exit the bacteria via simultaneous assembly and exit through the cell wall of the bacteria 105. There are a number of phages that have been used in biochemical research historically for different purposes, for example enzymes derived from T4 and T7 phage is commonly used in many biochemical labs for the purpose of DNA cloning. All of those phages have been used to greater or lesser extent in phage display 105.

The research in this thesis was made using M13 phage display. GP Smith proved in 1985 that proteins can be displayed on the surface of an M13 phage as a fusion of minor coat protein pIII, and proved that it could be amplified compared to empty phages by panning it on an antibody specific for the fusion protein, showing that the protein is displayed and folded on the surface 106. The M13 bacteriophage looks like a cylinder of protein that surrounds the single stranded DNA genome. The phage consists of a protein coat composed of 5 coat-proteins pIII, pVI, pVII, pVIII and pIX 105. The two proteins of M13 phage most commonly used in phage display are pIII and pVIII. pVIII is the major coat protein that covers the side of the cylinder, there are approximately 2700 individual copies of pVIII per phage. Fusion protein to pVIII thereby results in multivalent display. The pIII protein is situated in the end of the phage and has five copies per phage particle so display on pIII renders a display of single or low valance of the displayed peptide or protein 107. The pIII protein is also responsible for the initiation of infection since it recognizes the F-pilus and TolA epitopes of the E-coli bacteria.

Display on the M13 phage is often accomplished by a hybrid system of library phage and helper phage. The library phage genome consists of a phagemide containing an origin of replication a gene coding for antibiotic resistance and the fusion protein. To produce phages in the bacteria one needs to utilize helper phage that carries genes required for phage production and export through the membrane, it also contains the standard version of the fusion protein. The helper phage origin of replication is mutated so that the production of helper phage genome is slower than library genome. This makes it faster and easier to produce phages containing a smaller phagemide and it brings down the number of fusion proteins on the surface of a phage.
The reason for having a mixture of fusion protein and the normal coat protein is stability of the phage itself and the access of the fusion protein to the bait is better if the peptide is not completely packed with each other. Also if the fusion protein is pIII having only fusion protein might reduce the ability to infect bacteria. With the aid of the helper phage system the number of fusion proteins displayed on pVIII protein is up to 1000 copies per phage (around 1/3 of the coat-proteins).

An advantage of M13 phage display is the way the library is cloned in to the DNA of the phage. Using the fact that phages under the right conditions have single stranded DNA genome, incorporation of large libraries of different sequences is easily done by having sequences complementary to the phagemide on each side of the varied sequence and annealing 5' phosphorylated oligonucleotides to the single stranded DNA by a single cycle of denaturation and annealing, followed by complementary strand synthesis. By electroporation of the library in to E-coli superinfected with helper phages this can yield libraries above 10^10 unique sequences. When the phage library is created it can be used in hundreds of selections and through amplification of the library the number of selections possible goes up by a factor of 10. Even though one should be aware that amplification of phage libraries might skew the composition of the library and introduce more mutations.

**Phage display selections**

In our phage display system we immobilize a bait protein, usually fused to GST or other larger protein tags, on the surface of a well in a high adsorbent 96 well-plate, and a pre-selection protein, usually the fusion protein, in a separate well. The immobilization can be done in different ways for example on magnetic beads, and in some systems pre-selection is not used. Then the peptide-phage library is first added to the pre-selection wells and after binding of phages to unspecific things (to the bait protein), the phage pool is moved to the bait protein wells. After incubation with the bait protein the phage pool is removed and the wells washed with mild detergent and buffer to remove phages that are not bound to the bait protein. Thereafter the phages are eluted by addition of log phase bacteria, for initial infection of the phagemide and thereafter the bacteria are infected with helper phage for production of new phage particles over night. There are other ways of eluting the phages from the bait protein e.g. by low pH and there after infecting the eluted phages in a log phase culture. The phages that have been amplified over night are used as the in-phages in the next round. This is then repeated in 3-5 rounds of selection (Figure 5). Monitoring the selection efficiency can be done in two ways. One way is infect bacteria with the phages used as in phages and out phages and plate them on plates with selection pressure for the library phagemide, there after count colonies. Through a selection you should enrich binders to your bait therefore the ratio between out phages/in phages should trend towards one, since you retain more phages.
in later selection rounds. Another way to monitor the success of selection, which is the main way selections have been monitored in this thesis, is by enzyme linked immunosorbent assay ELISA. After selection bait protein and pre-selection protein are immobilized in wells on a 96 well plate, one for each round of selection. Then phages from the different days are added to the protein wells and pre-selection wells, the phages that are specifically selected for the bait protein will bind and will not be washed away. After that an anti-M13 antibody coupled to horseradish peroxidase is added and will bind to the phages in the wells. The detection is then made by adding TMB substrate to all the wells simultaneously starting a reaction that creates a blue color. The reaction is then stopped by addition of an acid, and then the absorbance in the wells is read by a plate reader. From a successful selection you will see high signal in the wells with the bait protein and low signal in the pre-selection wells. If you take the ratio between the signal in the bait protein well and the pre-selection well and get an increase in signal throughout the rounds with a ratio of 2-15 the selection has been successful and by sequencing of the phage pools you will find DNA coding for peptides that binds specifically to your bait protein.

**Protein-phage display**

One common use of phage display system is protein phage display where whole proteins or protein domains are used as the fusion protein and the library consists of varied residues in the binding pocket of the protein. The most common way of protein phage display is monovalent protein phage display on pIII which has successfully been used in protein engineering creating proteins that have greater affinity towards a specific ligand protein \(^{112}\), peptide \(^{113}\), or DNA \(^{114}\). To the greatest extent protein phage display has been utilized to produce light chain antibody libraries for the creation of synthetic antibodies \(^{110,115}\).

**Peptide-phage display**

Peptide phage display is a method where the display on the phage is shorter stretches of amino acid residues where they usually are so short that they take no secondary structure (7-20 aa). This is often done as a multivalent display on pVIII. Wild type pVIII is oriented with its N-terminal to the outside of the phage and the C-terminal towards the center, so conventional peptide phage display can not rely on interactions that require free C-terminal such as PDZ domains. In 2004 through extensive mutational analysis a recombinant pVIII protein was created that displayed C-terminal peptides on its surface \(^{116}\). Using combinatorial peptide phage display one can find binding motifs of a domain and then couple that with bioinformatics approaches to find putative novel protein-peptide interactions (as done in paper I). This has been done to find the binding motifs of SH3 domains \(^{60}\) and PDZ domains\(^{77}\).
cDNA and ORF phage display
Phage display has also been utilized to display proteomic proteins or peptides in a couple of different ways in order to directly map interaction between the bait and the proteomic interaction partner. cDNA or open reading frame libraries, where libraries consisting of different full length proteins or proteins fractionated in to shorter segments of various lengths around 200 bp, have been reported both for pIII and pVIII display. The problem with these display system is the fact that the peptides are of varying lengths and therefore expressed differently well. Furthermore it also displays fragments regardless of domain boundaries, which might make the cut right in the middle of a domain, and therefore display peptides that in a cell will be buried in a folded domain. Such libraries may also contain a high proportion of long frame shifted sequences, and that also produces a lot of premature stop codons.

Proteomic peptide phage display
A novel method that utilizes proteomic sequences for peptide phage display is Proteomic peptide-phage display (ProP-PD). In ProP-PD we combine bioinformatics with the creation of synthetic oligonucleotide libraries, making it possible to incorporate proteomic sequences of defined and equal lengths in to the phages. This technique, as used in M13 phage display, was first described by Ivarsson et.al in 2014 where they used two heptametrical C-terminal libraries, one containing all C-terminal sequences of protein-isoforms and one library containing C-terminal sequences of known infectious viruses. They used these libraries in phage display against PDZ domains of Scribble, DLG1, Erbin and densin-180. They found several new putative protein-peptide interactions and also confirmed previously reported interactions. This phage library is used in paper II. The ProP-PD method has been further developed to display 16-mer peptides of internal disordered regions of the human proteome, and used to find protein-protein interactions of a number of peptide binding domains, and the phosphatase holoenzyme of PP2A-B'. The main disadvantage of this method, as it shares with other methods except peptide array, is the fact that it currently only works with the canonical amino acids. Interactions based on posttranslational modifications, can thereby not be discovered in a straightforward manner.

Discovering interactions regulated by phosphorylation
There is a significant challenge in large-scale studies to capture interactions that are affected by phosphorylation, since neither yeast nor bacteria can perform phosphorylation on the level of higher eukaryotes like human. This means that Y2H, yeast surface display or phage display have not been useful for these types of interactions. The most straightforward method for mapping...
phosphorylation-dependent interactions is to use peptide arrays. Tinti et al. mapped the specificities of the SH2 domains by synthesizing all known (at the time) human peptides containing phospho-tyrosines. A randomized phosphopeptide array library was employed to map yeast 14-3-3 by Panni et al. Phosphopeptide arrays were also employed to find PDZ domain interactions negatively regulated by phosphorylation at the -2 position. Another system to map binding based on phosphorylated peptides is to use a phosphopeptide library free in solution to bind to a bait protein immobilized in a column and after that elute the peptides and employ peptide sequencing. This was done Yaffe et al. to map binding specificities of most human 14-3-3 isoforms. Since the number of confirmed or bioinformatically generated phosphorylation sites in the human proteome is so vast, it is very costly to synthesize peptide arrays large enough to cover the phosphoproteome. In addition, phosphorylation outside the core motif may modulate affinity rather than being a yes/no switches. Finding these interactions will be extremely costly if solely employing peptide micro arrays. In paper III and IV I will introduce a novel method that is an extension of the proteomic peptide phage display called phosphomimetic phage display that are a solution to some of these problems.

Next-generation sequencing in relation to large-scale protein-protein interaction discovery

Next-generation sequencing (NGS) has opened novel possibilities in terms of scale for methods such as Y2H, yeast surface display and phage display in the pursuit of finding PPIs. NGS is a way to massively parallel sequence a lot of different DNA sequences. This allows the researcher to sequence pools of various DNA sequences. I will here describe the Illumina system but other systems, like Ion torrent, works in similar fashion up to the sequencing step where it uses different detection methods. If you want to investigate the DNA content of several different pools at the same time you need to distinguish sequences resulting from the pools by adding unique flanking base pairs to the end of the sample sequence, these are called barcodes. This is done via PCR preferably with a high-fidelity polymerase. During the PCR you also incorporate an adaptor region per side after the barcode sequence. The adaptor region is a sequence that will base pair to covalently attached DNA on the illumina chip. After PCR you can pool all your samples preferably in a way that normalizes product between the different PCR:s. In the Illumina machine single stranded DNA will bind to the covalently attached sequence complementary to the adaptor sequence spread out over the chip. The first amplification takes place rendering your sequences attached to the surface. The original template is washed away. Then ”bridge amplification” is started where a bridge is formed by your sequences to a complementary
adaptor region on ship and amplification is performed. This is repeated until you have a spot with a lot of copies of the original sequence. After that, sequencing starts where the TGAC bases fluorescently labeled to different fluorophores compete for incorporation. When the base that is complementary to the sequence gets incorporated the amplification is stopped and a fluorescent read is made determining which base got incorporated. There are now several spots on the chip that all will flash in the color of the fluorophore. After this the fluorophore is cleavage off and next fluorescent base can be incorporated. This goes on until you have read the whole sequence including the barcodes. As the readout you get a FASTQ file containing all sequences and an ASCII score for how confident the read was for each position. With that information you can sort the sequences in to pools based on the barcodes and quality of the reads 126,127.

Biophysical validations of protein peptide interactions

The interaction data generated by a large scale study needs to be independently verified, this can be done either by testing the validity of the majority of the interactions via an analogous high-throughput method, or by validating some representative interactions via more extensive low throughput methods. Interactions found via methods optimized for finding interactions of low affinity need to prove that interactions found are of sufficient affinity to be biologically relevant.

Binding affinity as represented by the dissociation constant

Proteins that interact specifically with each other without forming covalent disulfide bonds can be described as interacting through a normal equilibrium reaction.

$$A + B \rightleftharpoons AB$$

By knowing the concentrations of $[A]$, $[B]$, and $[AB]$ one can determine the equilibrium constant for the association reaction $K_a$ (the reaction going to the right) and the dissociation $K_D$ (the reaction going to the left). Under pseudo first order conditions, the concentration of $A$ is considerably smaller than $B$. Under such conditions, the formation of AB complex reduces the concentration of free $B$ insignificantly. If you increase the concentration of $B$ at a constant concentration of $A$, then eventually all the molecules of $A$ will be bound to a $B$ meaning that $A$ is saturated. Under those conditions the reaction can be described as follows:

$$K_a = \frac{[AB]}{[A][B]} = \frac{1}{K_D} \Rightarrow K_D = \frac{[A][B]}{[AB]}$$

Where $K_a$ is the association constant and $K_D$ the dissociation constant.
Under conditions where the formation of AB significantly reduces the concentration of B the $K_D$ can be described by:

$$K_D = \frac{([A_0]-[AB])([B_0]-[AB])}{[AB]}$$

$K_D$ has the unit of concentration, [Molar], and stronger binding results in a lower $K_D$.

Titrations of binding as described above results in a hyperbolic function with the fraction of A that is bound to B on the Y-axis and the concentration of B on the X-axis. $K_D$ is where half of the binding sites are occupied. If the scale of the concentration of $[B]$ is converted from a linear scale to a $\log_{10}$ scale the curve becomes sigmoidal and the $K_D$ is at the inflection point (Figure 6).

![Figure 6. Binding curves for affinity determination](image)

There are several methods for determining the $K_D$ value. In this thesis two main methods have been employed, i.e. microscale thermophoresis and isothermal titration calorimetry.

**Microscale thermophoresis**

Microscale thermophoresis (MST) utilizes the thermophoretic effect. Creating temperature gradients in a liquid makes the molecules move in the solution. The speed and the direction of movement is largely due to the hydration shell around a molecule. Going back to the binding scenario described, the species A, B and AB will each have a different thermophoretic profile. Assuming that you can monitor the movement of A as it gradually gets saturated by B you will see a change in speed and possibly in direction of the movement (to or from a heated area). Plotting this change in movement against the concentration of B will result in a binding curve. The MST system provided by Nanotemper technologies uses glass micro capillaries, one per titration point, that once filled with a small volume of the titration, it is loaded in to the machine and an infrared laser to heat up a spot in the capillaries. The motion is monitored via fluorescents of the A compo-
ponent until the reaction has reached a steady state (Figure 7A)\textsuperscript{130}. In this thesis this technique is used to measure $K_D$ values of interactions between peptides fluorescently labeled with fluorescein isothiocyanate and varied concentrations of PDZ domains.

**Isothermal titration calorimetry**

Binding events between proteins can be described with thermodynamic parameters. $H$ is enthalpy, which is the heat, spent or produced during binding, $S$ is the entropic effect describing the randomness of a system and $\Delta G$, which is the Gibbs free energy. The decrease in Gibbs free energy in a closed system equals the work performed by the reaction. The thermodynamic parameters relate to each other according to the following equation, where $T$ is the absolute temperature:

$$\Delta G = \Delta H - T \Delta S$$

The free energy of a binding reaction can also be described in terms relation to the equilibrium constant and therefore $K_D$ in the following way:

$$\Delta G = -RT \ln K_D \Rightarrow K_D = e^{-\frac{\Delta G}{RT}}$$

Isothermal titration calorimetry uses the fact that a binding reaction has an enthalpic component to determine $K_D$. By stepwise titration, via a stirring syringe, of one component of A or B into a constant concentration of the other in a cell one can measure the energy in terms of heat released or consumed by the binding-reaction. This is practically done by measuring the temperature between the reaction cell and a reference cell and monitoring the energy it takes for the machine to get the reaction cell back to baseline temperature. When more and more of the components of A and B have bound to each other the heat released or consumed by additional titration steps decreases until saturation has occurred (Figure 7B)\textsuperscript{131}. From an ITC experiment you get $\Delta H$ and $\Delta S$ And $K_A$, and the system is set to a constant predetermined temperature $T$, which makes it possible to calculate $\Delta G$ for a reaction.
To understand the specificity and selectivity of an interaction it is important to study it on a structural level. There are two main ways of probing PPIs on a structural level, namely X-ray crystallography and protein nuclear magnetic resonance spectroscopy (NMR). There are advantages and disadvantages of both methods, both require large amount of protein. For crystallization finding the right condition can be very hard, the construct might have large parts that is unstructured which does not crystalize well and therefore the boundaries of the domain might need to be optimized. The structure generated by X-ray crystallography might also contain artifacts generated by crystal packing. One other problem for studying protein peptide interactions is that it can be tricky to crystalize interactions that are of low-moderate affinity. To
make protein structures in NMR one needs to incorporate $^{15}$nitrogen and $^{13}$carbon in to the amino acid residues of the protein. This is done by growing the protein in *E.coli* in minimal media and add labeled glucose and labeled ammonium. This can be a tricky process if you have low or sensitive expression of your protein. Some proteins are not well expressed in *E.coli* at all for instance proteins that require PTM to fold and function properly. In this thesis we have used NMR for structural investigation. The main advantage of NMR over X-ray crystallography is that it is made in solution, which makes it more native like.

Protein NMR has been developed since the 1970s towards *de novo* structured determinations $^{132}$. The first protein structure determined solely by NMR was the proteinase inhibitor IIA from bull seminal plasma published in 1985 $^{133}$. The main principle is to put NMR sensitive nuclei ($^{15}$N $^{13}$C have active nuclei) in to a large magnet (the NMR machine) and apply radiofrequencies of various strengths and amplitudes to the sample. The strength of a NMR magnet varies in MHz and it is usually denoted in the resonance frequency of protons between 500 MHz and 600 MHz. Each NMR active nuclei has a distinct chemical environment and therefore gives a unique signal called a chemical shift, represented as a peak in a spectrum. NMR used for less complex molecules than proteins can often be assigned using a one-dimensional spectra, because fewer atoms reduces the chance of having overlapping shifts. But protein molecules have so many atoms in it that two (and possibly three-) dimensional spectra of $^1$H-$^{15}$N-$^1$H-$^{13}$C heteronuclear single quantum correlation (HSQC)(Figure 8a) is a way to circumvent this problem. In a spectrum where you plot the $^1$H (ppm) against the $^{15}$N (ppm) each hydrogen coupled to a nitrogen atom should give a peak in this spectrum. This includes the amide bond for all amino acid residues except proline because it has a secondary amide. Usually a series of experiments are done and the entire protein backbone and side chains are assigned. Subsequently, experiments are done to determined distant restraints between pairs of nuclei in space (NOE). Together this information is fed into a software where the structure is then calculated. Experiments can also be done to determine the flexibility of different regions such as loops (Figure 8b). Adding a peptide that binds to the protein will cause a change in the chemical shifts due to a change in the structure and/or a change in the local chemical environment of the binding residues. One can use this information of the change in the shifts of the protein to determine affinity by sequential titration in increasing amounts of peptide as well as the change in the structure upon binding.
Defining specificity and selectivity in terms of differences in affinity and binding-motifs

Binding specificity and selectivity are words that might be used interchangeably and might not have clear distinct definitions from each other. This might in part be due to the words being more used in catalysis where enzyme specificity is better defined. In this thesis both terms will be used describing different things. When a protein or a peptide-binding domain binds to peptides they do this with linear motifs as described above. If we use the class I binding motif of PDZ domains as an example X-S/T-X-Φ COO-, this shows the specificity of the domain is hydrophobic residue in the terminal position and serine/threonine in the p-2 position. There are usually several peptides that bind to a PDZ domain that exhibit this motif, but might bind with different affinity. These affinity differences results from the residues that are in the other positions in the peptide. If a domain has the affinity of $K_D^A$ 2 µM and $K_D^B$ 6 µM to two different peptides you can show the difference as $K_D^B/K_D^A=3$ which makes the domain 3 x selective towards peptide A. In a cell small differences in affinity might define if the domain bind to peptide A or peptide B due to local concentrations of the peptides and the domain. Local events in the cell might affect the affinity towards one or both of the peptides. This can be a change in pH, a PTM or allosteric changes in the protein due to binding of something else to another part of the protein. For example the $K_D^B$ can change from 6 µM to $K_D^{B*}$ 0.5 µM, which makes the domain 12 times more selective binding to “B*” over condition “B“ and 4 times more selective to “B*” over “A”. A PTM that gives A* an immeasurable $K_D^{A*}$ towards the domain means that the domain has no specificity for A*, and that is highly selective for A versus A*.
Having introduced the basic concepts and methods of my work it is now time to turn to the present investigations.
Present investigations

The common theme of the research presented in this thesis is the investigation of protein-protein interactions on a domain-motif level through the application and development of the peptide-phage display method. In the present investigations I will explore the differences in affinity that is the basis for selectivity and specificity and how this is connected to conditional events like calcium binding or ligand phosphorylation. For this research I made use of four phage display libraries to answer three different questions about what peptides binds to what bait protein and under what condition:

I Paper I is about Sorcin which is a calcium binding protein. The hypothesis was that calcium binding creates a conformational change in the protein, which opens up a new binding sites\(^{134}\). This is investigated combinatorial peptide-phage library displaying 16 amino acid residues long peptides N-terminally of the pVIII protein.

II In paper II we investigate the specificity of a PDZ domain engineered to bind the C-terminal residues of the viral oncoprotein E6 of high risk HPV 18\(^{135}\), as compared to the wild type domain. For this we made use of a phage library that displays C-terminal peptides (7 residues) of the human proteome\(^{120}\).

III Paper III, which represents my main PhD project, presents a proof-of-concept study expanding the use of peptide phage display to find phospho-regulated interactions, by utilizing the phospho-serine/threonine mimicking ability of glutamate and comparing it to pre-phosphorylation of the phage library using a serine/threonine kinase. The method is used in selections against PDZ domains, a set of peptides was selected for detailed structure-function analysis as detailed in paper III and IV.
All studies presented involve collaborations with different research groups. My contributions to the studies are:

Paper I  I performed the selections and the colony phage ELISA and subsequent sequencing of positive clones, alignment and analysis of the peptides found in the selection.

Paper II  I participated in the experimental design and the experiments of the phage selection. Together with analysis of phage display results and writing of the manuscript.

Paper III  I contributed to the design of the study. I created the phage libraries and performed all selections and data analysis. I also performed the MST affinity determination of Scribble PDZ1 and DLG1 PDZ2, ITC affinity determination of all variants of Scribble PDZ1. I performed all cloning and cell-based experiments. I produced the protein for NMR spectroscopy. I participated in analysis of the results and in writing of the manuscript.

Paper IV  I produced the protein for NMR spectroscopy, and participated in interpretation of the results and writing of the manuscript.
Soluble resistance-related calcium binding protein (Sorcin) is a penta EF-hand protein, EF-hands are common helix-loop-helix fold that usually binds one Ca\(^{2+}\) ion per hand. The protein is a 22 kDa homo dimer where the EF hand 4 and 5 participate in the dimerization and EF hand 1-3 are free to bind calcium. Sorcin bind calcium in the \(\mu\)M range and does therefore not meaningfully bind calcium in the cytosol under normal conditions where the Ca\(^{2+}\) concentration is 10-100 nM. There are several pieces of evidence that sorcin could bind to different proteins in a calcium dependent manner so we sough out to explore the mechanism of this and if the interaction was based on linear motifs and in that case if the motifs can be discerned by peptide-phage display.

The group led by Dr. Gianni Colotti solved the crystal structure of Sorcin in presence and absence of Ca\(^{2+}\) (Figure 9). The protein consists of 8 \(\alpha\)-helices, of which A, B, C, E, F and H are short and D and G are long. When the protein is binding Ca\(^{2+}\) the long D \(\alpha\)-helix makes a large conformational shift (ca 21\(^\circ\)), which opens up the center core of the protein (see red arrow Figure 9A). The conformational shift exposes a hydrophobic patch in the protein where a potential peptide binding-site emerges (Figure 9C and D). Indeed, the calcium bound structure was actually co-crystalized with a (GYYPGG) peptide, from the flexible N-terminal the of other Sorcin dimer, bound in the opened binding site. There are a number of hydrophobic residues aligning this binding-site in helix D there is a tryptophan (W105) that makes a stacking interaction with the proline in the peptide. Based on these observations we decided to explore the motif-mediated interactions of Sorcin, in absence and presence of Ca\(^{2+}\).
I performed phage display selections against immobilized Sorcin with a 16-mer highly diverse combinatorial peptide-phage library (4x10^10). The library is N-terminal of the pVIII protein, and to avoid effects of the free amine in the N-terminal, and to get the peptide free in solution from the phage, linkers of SSSG-peptide-GGSGG flank the displayed peptide. Selections were performed in parallel, using either 1 mM Ca^{2+} or 1 mM or the chelating agent EDTA supplemented in the buffers. The selections were successful as determined by pooled phage ELISA. Individual phage clones confirmed by clonal phage ELISA were subjected to Sanger sequencing. From the selection 38 sequences from Ca^{2+} selection was obtained and 20 from the EDTA selection. The sequences were manually aligned and two motifs emerged (Figure 10, Table 1), a Φ-Φ-X-P motif and one with D-Φ. Both motifs were found in both presence and absence of calcium. The Φ-Φ-X-P peptide share the most similarities to the GYYPGG peptide found in the crystal structure and likely bind in the same fashion.
The results of sequencing the clones positive in ELISA and aligning their peptides resulted in two motifs. A the $\Phi$-$\Phi$-$X$-$P$ motif that was most abundant ($n=34$) and which is mostly matching the peptide crystalized in the binding-pocket of the structure binding calcium. And B the $D$-$\Phi$ motif ($n=18$), which is mainly found in the presence of calcium (16/18). Over (3/18)

The fact that the $\Phi$-$\Phi$-$X$-$P$ containing peptides are found in both presence and absence of $Ca^{2+}$ indicate that peptides of this motif might bind to and stabilize a open conformation even in the absence of calcium. Indeed, in SPR experiments Sorcin was found to bind the $\Phi$-$\Phi$-$X$-$P$ motif of PDC6 both in presence ($K_D$ 3.5 $\mu$M) and absence of calcium 12 $\mu$M, which gives a selectivity change of $K_D/K_D^{Ca^{2+}}=3.5$.

The $D$-$\Phi$ motif was found in 16/2 in the calcium selection/ EDTA selection but the number of peptides displaying this motif was lower than the number of peptides that had the $\Phi$-$\Phi$-$X$-$P$ motif. Out of the peptides found for the calcium selection 20 had the $\Phi$-$\Phi$-$X$-$P$ motif and 18 had the $D$-$\Phi$. Where the $D$-$\Phi$ motif would bind is not determined, and in fact the motif could be an artifact of the combinatorial phage display where such DW containing peptides have been observed to be enriched in unrelated selections of diverse domains (unpublished results).

Another serious drawback with combinatorial peptide phage display is that tryptophan seems to be over represented in these selections. This can be why the peptide bound in the structure is less hydrophobic than the peptides selected. The disordered regions of proteins will have a low content of tryptophans since it is the residue with highest negative correlation with disorder.
Table 1. Alignment of the peptides found to bind the sorcin protein in ELISA and which were found in what quantities in the different selections.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>#1 mM Ca²⁺</th>
<th>#1 mM EDTA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group 1:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DTSYWVVWVGDGYST</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PAYSMMFWMPAISYV</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GWWWFQPQNHILGNMVS</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>QSYLWWQFPVSIISA</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SWWWFMDNFWTFYVT</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>WWWQWWWQFQEPMVM</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SVAWWTWERTWQSITM</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>AYSRWWFQMYEVVG</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>VEEMTWTEYIQYMQT</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PEWHELSTREMMWEILR</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MWWWWAESPSEVRHLS</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GNAGFFWHHWFRLES</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>VYMMWWSDFMQWQWQI</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>YDFSWMWQHMGWQLSLV</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GSSEHWFSEWTAWVTE</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>GSWWGLATTFQQWFFV</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PWLAEPAEHEVMQYYF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>VWTGYDFMNITTWFIM</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>HQWEYHWQWQASGM</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TWMGTQWWWVGGVQ</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>WPFWWVQGQYAAAS</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>PSEWWWWRELGMTMVE</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>DTSYWVWVEQGDYST</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>SIGAWWWGECQDYVHV</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>WPVYWWGELAMQLDYV</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FMLNWNATWEWTTPHQ</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>WSRSEWWQWAIIFSFWTNY</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>AWYRFNVNWSWFVEYVPT</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MQWQWGVEHNMWLDIE1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>YTGFMSMWRELWQPVG</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FYFWMWMAQQGWDLVLE</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MYWMDLEYGWGWEMYQ</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>YWWHEFQEFENMDILE</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>WYGFDFWNFWIQTSWS</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>Group 2:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QHDMYSQWYTLVSTMV</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>PHELWLTTPFDMDMQ</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>VVSTDNYMWSWDFMMW</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MEVITWETFYRMYH</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>TGDAFMFDLYQWQNV</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FAEQEGEPDWPwaldVV</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>EMWADWYTLMAEMENV</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>WTERDFMPMLNLSMQM</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>WQDDFMMMAWIKHA</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>WTELDWLMQNASMT</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>VCDMDTWDDNWMTLWE</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ELDWFNYMVSQWRTMV</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>YNdEYWQNLWETKTY</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MPPLSDWDWTVYMTT</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IEQWYERMMSLQADHI</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>AWDCWGTDNDWPDPVF</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>VDHDWVDWQGEMARVYQ</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>WMNFVDWLYWKRHTIE</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Paper II Exploring the specificity of a PDZ domain engineered for high affinity interaction with the viral oncoprotein E6

Human papillomavirus (HPV) family is a common virus infecting epithelial cells in the body. Some viruses infect mucosal tissue and the infection is in some cases not cleared by the immune system. In those cases it can cause several types of cancer, of which the major concern is cervical cancer. HPV 16 and 18 are the cause of 70% of all cervical cancers. This is mainly due to sustained expression of the oncoprotein E6 and E7. E6 of the high-risk HPV strains targets PDZ containing proteins for degradation by binding to the PDZ domains with its C-terminal PDZbm. The HPV16 C-terminal sequence is TRRETQL^COO^- and the HPV 18 sequence is RRRETQV^COO-. The sequence differences lead to discrepancies between which PDZ domains they bind to, as some are selective for a peptide with a V at position 0 and some prefer an L at position 0. For example HPV 18 binds with high affinity to all DLG PDZ domains while HPV 16 binds tighter to PDZ domains of Scribble. The E6 protein of both strains bind with high affinity to the second PDZ domain of the MAGI proteins. Targeting of E6 protein is an important step in preventing cancer progression and the fact that it binds with its C-terminal to PDZ domains is something that can be utilized.

To create an optimized binder of the HPV 18 E6 protein my collaborators performed a protein phage-display using the second PDZ domain of DLG1 (Figure 11). In the study they used a pseudo wild type DLG1 PDZ2 domain (pWT) C347A I342W optimized for fluorescence based affinity measurements. They made a combinatorial phage library where they varied 5 positions in the α2 helix of the pWT DLG1 PDZ2 domain (His384, Glu385, Val388, Leu391, Lys392) and used it in selections against an immobilized peptide (residue 52-158) of the E6 protein from HPV 18 (Figure 11B). They measured the kinetic parameters of binding to the E6 construct of nine phage derived domains. The pWT PDZ domain had an affinity towards the E6 construct of 0.53 μM K_D. The L391F, K329M Ø9 (Figure 11D) showed a substantial increase in affinity with a K_D 0.093 μM, (K_D^WT/K_D^Ø9 = 6). The assumption was that the domain also would be more specific for the target peptide that it was engineered to bind. In paper II we explore if high affinity is coupled to high specificity.
Karlsson et al. performed protein engineering of the second PDZ domain of DLG1 through combinatorial protein-phage display (A-B), where they varied residues in the α2 helix to find domains that bound better to the C-terminal portion of the oncogenic protein E6 from high-risk HPV18. The domain from the selection that bound best to E6 called Ø9, bound 6 times stronger than wild type (C) due to substitutions L391F and K392M (D). C binding between HPV 18 E6 and DLG1 PDZ2 (PDB code 2I0L). D Is a model done by swiss-model based on the structure in C. (Part of the domain is hidden to better show the binding site).

The focus of paper II was to investigate what the L391F and K329M mutations of Ø9 do to the selectivity of the domain for peptides present in the human proteome. This was done by simultaneous selection against the Ø9 and the pWT domain using a heptameric C-terminal proteomic peptide-phage library that displays the C-terminal regions of the human proteome. Both selections were successful as determined by phage pool ELISA. Clonal phage ELISA was performed until 96 positive clones were found for each selection. Sanger sequencing of the binding clones found 71 peptides for pWT and 76 for the Ø9 mutant in total. Out of the total number of peptides the pWT domain had 9 unique peptides selected and the Ø9 had 17 unique peptides (Table 2).
Table 2. The peptides sequenced from the different selections. The number of sequences per peptide is indicated for the different domains. The number of unique peptides found for the wild type domain was 9 and the number of peptides found for the Ø9 domain was 17.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Prevalence</th>
<th>Uniprot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pWT PDZ2 (71)</td>
<td>PDZØ9 (76)</td>
</tr>
<tr>
<td>KRKETLV</td>
<td>41</td>
<td>31</td>
</tr>
<tr>
<td>RSISTDV</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>IKETTTV</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>WKHETTV</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>NSKETVV</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>KIKEETTV</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>WKSETTV</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AGRETTV</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>WKNETTV</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>EKKHTDL</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>VSKETPL</td>
<td>5</td>
<td>MK12_HUMAN</td>
</tr>
<tr>
<td>SARSTDV</td>
<td>4</td>
<td>ANO9_HUMAN</td>
</tr>
<tr>
<td>RAISTDV</td>
<td>2</td>
<td>F163A_HUMAN</td>
</tr>
<tr>
<td>TSRETDL</td>
<td>1</td>
<td>KCNA5_HUMAN</td>
</tr>
<tr>
<td>YRRESAI</td>
<td>1</td>
<td>KCN4_HUMAN</td>
</tr>
<tr>
<td>PGKETQL</td>
<td>1</td>
<td>SO1C1_HUMAN</td>
</tr>
<tr>
<td>YKKTPL</td>
<td>1</td>
<td>ANR50_HUMAN</td>
</tr>
<tr>
<td>KGTETTL</td>
<td>1</td>
<td>S4A4_HUMAN isoform 5</td>
</tr>
<tr>
<td>AGKTTTL</td>
<td>1</td>
<td>F8VP99_HUMAN</td>
</tr>
</tbody>
</table>

I produced two different position weight matrices using weblogo 3 where I used all the sequences as input and one where I used only the unique peptides (published) (Figure 12A and B). In terms of sequence specificity both domains still show a preference for the classic class I PDZbm but the L391F, K392M mutations of Ø9 makes the domain less selective for the residues in p0. In the PWM comprised of all sequences it seems like the domain also is less selective in the p-3 position, something that is not obvious if only unique peptides are used in the PWM. Based on the results of the selection the pWT domain is specific towards valine at position 0 but the Ø9 domain has roughly the same preference for leucine in that position. This difference in specificity can be explained by the L391F mutation. In figure 11 you see that this residue makes contact with the terminal residues R chain. Phenylalanine is bigger than leucine, however the second carbon in phenylalanine is part of the phenyl ring and therefore planar, while the second carbon in leu-
cine is bound to two methyl groups and a hydrogen which makes the angles between the carbons different, and one of the methyl groups of leucine might protrude into the binding pocket making the pocket smaller and therefore less suited for a leucine in the terminal position of the peptide. The more relaxed specificity for residues at p-3 of Ø9 domain compared to the pWT domain is less understandable since that residue is pointing in towards the β-sheets and not the α helix.

Through affinity measurements it was found that Ø9 bound with higher affinity to all of the peptides tested and that the domain clearly is more promiscuous. Similar observations have been made for other engineered PDZ domains \(^{144}\). Thus, we conclude that affinity and specificity for motif-based interactions does not necessarily go hand in hand. To achieve increased affinity and specificity a setup purely based on affinity might not be the best solution. Also if one thinks about the structure of PDZ domain peptide interactions, the residues that are specific in the motif S/T and the hydrophobic terminal residue are the ones oriented towards the helix, which carried the varied positions. It might have been more successful if the original library would have had their varied positions in the β2 and β3 sheets.

Figure 12. A and B Position weight matrices produced with the peptides sequenced from the different selections. A PWN is generated by using all peptides generated. B PWMs generated using unique peptides. At position 0 it is clear that there is a loss of specificity from exclusively V to a mixture of V and L. In the PWM comprised of all selected peptides there is a distinct difference in p-3 between the Ø9 and the pWT domain, which is not shown in the PWM from unique peptides.
Paper III & IV Proteome-wide analysis of phospho-regulated PDZ domain interactions

Background
As discussed in the introduction protein phosphorylation is the most common PTM and plays a critical role in cell signaling and function in health and disease. The majority of phosphosites are within or in close proximity to binding motifs within intrinsically disordered regions of the proteome phosphorylation can lead to increased or decreased affinities of interactions between the motifs, and binding domains or proteins, and consequently regulate PPIs. There is a need for methods to find phospho-regulated interactions at a large-scale with resolution on the level of binding motifs and the sites of phosphorylation, as most large-scale methods are ill suited for the task (see introduction). For example, ProP-PD is not equipped to find these interactions since bacteria do not perform eukaryotic phosphorylation. This project is aimed at bridging this problem.

In cell experiments it has been common, to use substitutions to aspartate or glutamate to mimic the phosphorylation of serine and threonine residues. Of these phosphomimetic substitutions, glutamate is sterically more similar to phospho-serine/threonine than aspartate (Figure 13A), although its hydrated shell is smaller than a phosphorylated residue. Even nature makes phosphomimetic mutations, as revealed by a study that investigated the conservation and evolution of phosphorylation sites. The authors determined at what serines known to be phosphorylated in one species of yeast, were substituted to other amino acids in other species. They found that substitution to phosphorylatable threonine was most common, followed by glutamate and aspartate.

Based on this we decided to explore the use of phosphomimetic substitution in combination with proteomic peptide-phage display, phosphomimetic ProP-PD as a way to identify phospho-serine/threonine regulated interactions.

Part A Method development and validation

Bait selection, library design, construction and quality validation
As bait proteins for the phosphomimetic ProP-PD we selected PDZ domains. As PDZ domains preferentially bind C-terminal peptides, the bait selection limits the size of the phage library to be constructed since there are only so many C-terminals in the proteome. At the start of this study there where only sporadic reports of PDZ domains that could bind phosphorylated peptides, but several reports of PDZ mediated interactions disabled by phosphorylation (see introduction and recent review by Liu et al.). PDZ domain
interactions can thus be both disabled and enabled by phosphorylation, which made them ideal targets for this proof-of-concept study.

To be able to directly evaluate the effects of the phosphomimetic substitution on binding, we determined to include both the glutamic phosphoserine/threonine substitution and the wild-type sequence in the library design. We designed a phage library displaying nine residue long peptides representing all C-terminal sequences of the human proteome that contains known or putative serine/threonine phosphorylation site, and the phosphomimetic version thereof. The sequence data was gathered from SwissProt/Uniprot and the phosphosites were gathered from PhosphoSitePlus \textsuperscript{45}, Phospho ELM \textsuperscript{151} and NetPhos 3.1 \textsuperscript{152}, in the fall of 2013. For peptides that contained more than one phosphorylation site in the same sequence both phosphosites were added individually and in combination, depending on the evidence for multiple phosphorylation. This created a library of 4827 wild type sequences and 7627 mimetic sequences (Figure 13B). The library was synthesized as costume designed oligonucleotide array with codons optimized for \textit{E.coli} expression and incorporated in to a phage library as described in the introduction.
Figure 13. A The basis for selecting glutamate as the phosphomimetic substitution is that it sterically more resembles serine and threonine. B-C We made a C-terminal library containing all known and putative serine threonine phosphorylation sites (4827). The library design contains the wild type sequences and the phosphomimetic variant of it and the total library size is 12454 sequences. The custom oligonucleotide library was incorporated into a phage library and used for selection against PDZ domains. D The successful selections were subjected to NGS analysis, and a subset of peptides and domains were selected for biophysical validations.

The coverage and quality of the library were analyzed, using the NGS data. Out of raw sequence reads 62% were of right length and of those 14.4% had non-synonymous mutations. This means that 53% of raw reads of the library matched sequences in the designed library. The coverage confirmed was 11 787 out of 12 454 sequences, or 94.7% of the library design. The distribution of number of reads per sequence spanned from 153 to 1 with a median of 15 and a mean of 18 reads per sequence, which means there are no major biases in the constructed library (Figure 14b).
The library composition of sequenced commercial peptide phage display libraries have some different biases in their composition\textsuperscript{153}. I therefore analyzed if there were any residue bias in the 666 peptides not sequenced in compared to what is confirmed by sequencing and what is typically observed in peptide-phage libraries (Figure 14a). The amino acid composition and the mean pI of the peptides that were in the library design sequences confirmed or not confirmed by NGS can be seen in (Figure 14A). The peptides that are missing from the sequencing are on average more basic, with an isoelectric point at 8.06 compared to 6.3 for the sequenced set. This is explained by the fact that the missing sequences contain an over representation of arginine, 13.0\% compared to 5.8\% in the peptides sequenced. The missing set seems to also be overrepresented in prolines representing 16.2\% of the not sequenced set and 6.5\% in the sequenced set. I randomly sampled 666 peptides in 10 sets, and no one of the randomly sampled peptide sets had as high frequency of proline or arginine as the not sequenced set, the average and standard deviation of proline was 7.02 ± 0.4 and arginine 6.3 ± 0.2. The fact that prolines is over represented in the set not sequenced is contrary to combinatorial phage libraries where proline usually are over represented compared to the number of codons it has\textsuperscript{153}. 
Figure 14. The naïve phage library was sequenced deep enough to confirm a coverage of 94.7%. A The table shows the frequencies of amino acid residues in the library design in the part that is confirmed by sequencing and the part that is not confirmed by sequencing. B The graph shows the read distribution between the sequences that got sequenced.
Phage display selections and sequence analysis of binding enriched phage pools

I tested the performance of the phosphomimetic ProP-PD library against six PDZ domains, the 1-3 PDZ domain of Scribble and DLG1 (Figure 15A). These proteins are in complex with Lgl involved in establishing and maintaining apico-basal cell polarity in epithelial cells, they are also involved in neuronal development synaptic organization cell migration, and cell proliferation. They have also previously been successfully used as bait proteins in selections against combinatorial and proteomic peptide phage display. The domains bind class I PDZbm, and their interactions have been suggested to be up and down regulated by ligand phosphorylation. This makes them suitable test cases for exploring the potential of Phosphomimetic ProP-PD, for discovering interactions that are enabled or disabled by phosphorylation. The selection was carried out in four rounds of phage display and each selection was performed in triplicate. The progression through the selection was monitored by phage pool ELISA. Phage pools from the third and fourth round of selections were barcoded through PCR and subjected to NGS.

The NGS analysis generated a large number of DNA sequences that were sorted based on their barcodes. The sequences were translated into their corresponding peptide sequences, and then matched against the library design. The fractions of NGS count in the sequencing as compared to the enriched selection are illustrated in figure 15B. The frequencies in the NGS counts reflects the enrichment through the phage display selection which roughly correspond to the relative strength of the binding within the given set.

Each peptide pair was analyzed if the wild type or the phosphomimetic variant dominated the selection. We hypothesized that for if particular target is selected only or primary as the phosphomimetic variant the interaction would be enabled by phosphorylation at that position and disabled if it was only or primary selected as wild type. A number of peptides have a significant difference in the ratio of mimetic/total peptide compared to wt/total peptide (as determined by t test) those are indicated with an asterisk in Figure 15B.

A PWM of the peptides found in the sequencing was made using web logo 3, if the peptide was sequenced as both mimetic and WT both sequences was used in the creation of the PWM if only one sequence was present only that went in to making the PWM (Figure 16A). All domains bind to peptides with the class I PDZbm where Scribble PDZ1 and 3 have a preference for leucine in the terminal position and the other domains have a preference for valine.
To determine what affect the phosphomimetic substitution has at any given position in the motif, I calculated the ratio between mimetic count and total count for each peptide. The average of this for each position along the motif was then compiled in a selectivity matrix (Figure 16B) ratios between 0.5 and 1 (blue) indicate a preference for mimetic residue at the given position, and 0 to 0.5 (red) indicate a selectivity for the wild type residue. From this analysis it is clear that the domains are selective for phosphomimetic mutations at p-3 and to some extent p-1 is favorable for interactions, which suggested that phosphorylations at these sites light enable interactions.

Notably the C-terminal peptide of MCC is picked up as a Scribble PDZ1 ligand with a significant selectivity for phosphomimetic substitution at p-1 (Figure 16B). This is in accordance with Pangon et al. that found the phosphorylation at p-1 of MCC to be important for Scribble interaction \textsuperscript{83}, and supports the relevance of these findings made through phosphomimetic ProP-PD The selectivity matrix of Scribble PDZ1 suggests that this might be a general for p-1 in Scribble PDZ1.
Figure 15. A The first three PDZ domains of Scribble and DLG1 were individually subjected to phosphomimetic ProP-PD and the outcome was analyzed by NGS. B The NGS results were mapped to the library design and the fraction of each peptide compared to the total selection is depicted in the bar graphs as mean ± SD n=3 (Scribble PDZ2 n=2). The peptides that have a significant difference in frequencies between the wild type peptide and the mimetic peptide as determined via T-test are indicated with an asterisk.
Comparing phosphomimetic ProP-PD to Pre-phosphorylation of ProP-PD library

In two papers from 1997 and 2003 the group of Gianni Cesareni showed that it is possible to use tyrosine kinases phosphorylate a phage library \(^{159}\) and used it to probe interactions dependent on phospho-tyrosine \(^{160}\). We were interested to see if this approach would work for finding proteomic interactions based on serine/threonine phosphorylation and how it compares to phosphomimetic ProP-PD. We therefore generated a ProP-PD library that consisted of the 4872 wild-type sequences in the phosphomimetic ProP-PD library. The coverage was determined by NGS to 98%. To phosphorylate the library prior to selection we acquired activated RPS6KA1, which is one of the kinases enriched in the binders for the mimetic library to Scribble.

First it was established that the library got phosphorylated by the kinase by ELISA using an anti phospho-ser/thr/tyr antibody (Figure 17B). The kinase was then used to phosphorylate the naïve library and the in phages in prior to each round of selection. In parallel, selections were performed with and without pre-phosphorylation with Scribble PDZ1 as the bait protein. Binding enriched phage pools were analyzed as before, which revealed signif-
icant differences between the selections for two peptides, particular peptides: the TMFLRETSL\textsuperscript{COO-} peptide from the GUCYA2 protein and the ARVSKE\textsuperscript{PLCOO-} peptide from the MAPK12 protein. In the case of the TMFLRETSL\textsuperscript{COO-} peptide, it was significantly enriched in the selection against the pre-phosphorylated library. The peptide has a potential RPS6KA1 p-1 binding site, and the results therefore suggest that its phosphorylation leads to stronger interaction with Scribble PDZ1 as reflected by the selection results. In contrast, the top binder in the unphosphorylated selection MAPK12, which lacks RPS6KA1 motif, is less strongly selected upon library phosphorylation. This may reflect that it is outcompeted by the GUCYA2 peptide in the pre-phosphorylated selection (Figure 17).

Figure 17. A library containing the 4827 wild-type sequences used in the phospho-mimetic ProP-PD was created and used in selections against Scribble PDZ1 after pre-phosphorylation RPS6KA1. B. The viability of phage library phosphorylation was assessed using ELISA. C. Simultaneous selections were made with and without pre-phosphorylation n=2 and subjected to NGS sequencing after successful selection. The * indicate statistically significant differences in selection as determined by t test. The GUCYA2 peptide TMFLRETSL with a potential RPS6KA1 phosphorylation motif for phosphorylation at p-1 is more enriched in the pre-phosphorylated selections.

Are the differences in the sequencing counts driven by a sequence bias in the library?

To ensure that the differences in peptide counts are not biased by the frequencies of the peptides in the input library we analyzed the frequencies of the selected peptides in the sequenced library. Comparing selected peptides in the library and the peptides that was not selected we see that the mean counts in the sequenced library is slightly higher than for the peptides not selected but there is no significant difference (Figure 18A). It is also im-
important to see if the sequences with high counts in the selection have higher starting count compared to peptides with lower count in the selection. As determined by plotting the counts in library compared to the counts in the selection we can see that there is no correlation in high copy number in the sequencing and high copy number in number in the library (Figure 18B).

![Figure 18. Library biases. A the count in the library sequencing of selected peptides compared to not selected is has a slightly higher average but not significant. B There is no correlation within the selected peptides and the number of reads it had in the library sequencing.](image)

**Biophysical validation**

To confirm that the results generated through phosphomimetic ProP-PD result from affinity differences between the wild-type peptides and their phosphomimetic counterpart, and that this in fact translates in to affinity differences between unphosphorylated and phosphorylated ligands. I determined the affinities for a select set of pairs through two distinct biophysical methods namely MST and ITC.

For the experiment recombinantly expressed Scribble PDZ1 was used and unphosphorylated, phosphorylated and mimetic peptides, with modifications at distinct sites, p-1 MCC HTNETSL\(^{COO^-}\), p-3 PRS6KA2 RL/STRL\(^{COO^-}\), and p-6 TANC1 KRSFIESNV\(^{COO^-}\). FITC-labeled peptide was held at a fixed concentration and varying concentrations of Scribble PDZ1 was used. Scribble PDZ1 is selective towards phosphorylation of the p-1 site of the MCC peptide as determined by the \(K_D^{WT}/K_D^{P}\) ratio of 3. The effect of mimic mutation at the same site was 1.6 (Figure 19A and C). The same selectivity is observed for p-3 phosphorylation of RPS6KA2, where the \(K_D^{WT}/K_D^{P}\) ratio is 5.3. In this case the \(K_D^{WT}/K_D^{min}\) ratio is even higher at 12.7 (figure 19A and C) this is in agreement with the phosphomimetic ProP-PD results where the difference between mimetic and wild type is greater for RPS6KA2 than for MCC. The phosphorylation of RPS6KA2 is has to our knowledge
not been confirmed experimentally but its homolog RPS6KA1 has. I therefore measured the affinity of this peptide to VRKLPSTTL\textsuperscript{COO-} where an increase in affinity from 1 \(\mu\text{M}\) to 0.39 \(\mu\text{M}\) was found upon phosphorylation by MST which gives a \(K_D^{\text{WT}}/K_D^p\) ratio of 2.5 (figure 19A and C). The TANC1 peptide (phosphorylation at p-6) is in the phosphomimetic ProP-PD selected mainly as wild type. The affinities for the peptide reflect this as it is more selective towards the unphosphorylated peptide \(K_D^p/K_D^{\text{WT}} = 2.4\).

To confirm the results an analogous method ITC was used for the phosphorylated and unphosphorylated peptides of MCC RPHTNETSL\textsuperscript{COO-} and PRS6KA2 MKRLTSTRL\textsuperscript{COO-}. Here the peptide was varied and the domain was constant. The selectivity for phosphorylated peptides was confirmed with a \(K_D^{\text{WT}}/K_D^p\) ratio of 1.6 for MCC and 4.5 for RPS6KA2. Lastly in the phosphomimetic ProP-PD selection Scribble PDZ1 was specific to wild type residues at p-2 and this was also tested using ITC with the MAPK12 ARVSKETPL\textsuperscript{COO-}. The affinity for the wild type peptide was determined to 1.2 \(\mu\text{M}\) \(K_D^{\text{WT}}\) but the phosphorylated peptide completely fail to bind the domain in the micro molar range (25\(\mu\text{M}\)-2200 \(\mu\text{M}\)) (Figure 19 B and C).

The same set of peptides used in MST with Scribble PDZ1, MCC HTNETSL\textsuperscript{COO-}, RPS6KA2 RLTSTRL\textsuperscript{COO-} and TANC1 KRSLVESNV\textsuperscript{COO-} was used in affinity determination by MST for DLG1 PDZ2. According to the phosphomimetic ProP-PD selectivity matrix (Figure 16 B) DLG1 PDZ2 should preferentially bind to p-1 and p-3 phosphorylated ligands (although the effects at p-3 were not statistically significant) and preferentially to the unphosphorylated peptide at p-6. Of the peptides used in the experiment only the TANC1 peptide had been selected by the DLG1 PDZ2 domain. For p-1, the \(K_D^{\text{WT}}/K_D^p\) ratio is 0.1 indicating that the domain is not selective for the phosphorylated MCC peptide. This is contrary to the selection results, although MCC was not selected for DLG1 PDZ2 and this domain prefers valine not leucine at the terminal position so the larger leucine might affect the phosphopeptide-DLG1 PDZ2 interaction negatively. The \(K_D^{\text{WT}}/K_D^p\) ratio of DLG1 PDZ2 for RPS6KA2 (p-3) is 1.3, and there is not a significant selectivity towards the phosphorylated peptide, the \(K_D^{\text{WT}}/K_D^{\text{min}}\) ratio however is higher at 6.5. In the case of TANC1 (p-6) peptide, it was found that DLG1 PDZ2 is selective towards the wild type peptide with a \(K_D^p/K_D^{\text{WT}}\) ratio of 2.3.
Figure 19. The affinity of Scribble PDZ1 for unphosphorylated phosphorylated and phosphomimetic peptides. A Microscale thermophoresis affinity measurements with FITC labeled peptides of unphosphorylated, phosphorylated and mimetic variants of MCC HTNETSL^COO-, RPS6KA2 RLTSTRL^COO, TANC1 KRSFIESNV^COO and unphosphorylated and phosphorylated RPS6KA1 VRKLPSTTL^COO- with a fixed concentration of 25-50 nM titrated with varying concentrations of Scribble PDZ1. B ITC affinity measurements with a constant concentration of Scribble PDZ1 was titrated with varying concentrations of unphosphorylated or phosphorylated peptides of MCC RPHTNETSL^COO-, MAPK12 ARVSKETPL^COO, and PRS6KA2 MKRLTSTRL^COO-. K_D values were determined and an average of 3 measurements ± SD are depicted in the graphs and table C. Statistical difference in binding affinity was determined for the peptides with unphosphorylated phosphorylated and mimetic with one way anova comparing the phosphorylated and mimetic to the unphosphorylated and where there are phosphorylated and unphosphorylated peptides with unpaired students t test. *** P≤ 0.001 ** p≤ 0.01. ns = not significant. For representative titrations if the ITC results see paper III figure 3.

The conclusion of the phosphomimetic ProP-PD experiments coupled with the biophysical validation is that a difference in peptide count between a wild type peptide and a mimetic peptide is reflects a difference in affinity between the phosphorylated and unphosphorylated ligands, as seen by the K_D^WT/K_D^P selectivity ratio.

In paper III we analyze the binders found for Scribble and DLG1 in relation to known binders. The overlap between known binders and binders found by phosphomimetic ProP-PD is 13 for Scribble and 16 for DLG1. The majority of the new interactions were those who are enabled by phosphomimetic substitution (paper III figure 4). We also analyzed Scribble and DLG1
interactions in a larger context by looking at enriched Gene Ontology terms and enrichment in KEGG pathways, where we found that our interactions were enriched in Hippo signaling, WNT signaling, Neuroactive ligand receptor interaction and oxytocin signaling pathway (paper III figure 4 table EV7). We also analyze what kinases are enriched in the selections in terms of phosphorylation of the ligands and found that Scribble ligands were enriched in binding motifs for RPS6KA kinases (paper III Table EV 8).

Comparison between Phosphomimetic ProP-PD and pre-phosphorylation ProP-PD

To compare the phosphorylated library with the phosphomimetic library, one get more results in the phosphomimetic selections. These selections are also much more controlled in terms of what peptides on the phage harbor the glutamate substitution in phosphomimetic ProP-PD compared to what peptides get phosphorylated in the pre-phosphorylation ProP-PD. There is also in our experiments no way of quantifying how many of the peptides on a phage that are phosphorylated. If the goal is to distinguish between those who preferentially bind to serine or phospho-serine it is important that most peptides on a phage get phosphorylated. In our phosphorylation experiment the peptide that got selected more frequently is a peptide that already binds to Scribble unphosphorylated. Also no kinase will phosphorylate all potential phosphorylation sites, which might be both a disadvantage and an advantage, depending on the scientific question. There are of course ways to optimize the phosphorylation experiment in terms of kinase selected, amount of kinase, time the phosphorylation is allowed to go on etcetera. There might also be phage specific things to consider like the linker composition and length between the peptide and the fusion coat protein. In this study we get more and clearer results with the phosphomimetic ProP-PD that I think is the better method of the two of them.

Instead of using a kinase one could postulate to chemically modify amino acids after peptide production to better mimic phosphorylation. This is primarily possible on primary amines and cysteine. Unpaired cysteine are usually not used in phage display since they do not display properly\textsuperscript{161} and primary amines are present in lysine which from the start are longer than a phosphoserine. Another interesting potential path to study the effect of ligand phosphorylation in phage display is to expand the genetic code of bacteria to produce phages with peptides incorporated with phosphoserine/threonine residues. This is typically done by using specific engineered tRNA carrying the unnatural amino acid that match the amber stop codon UAG\textsuperscript{162}. Recently Barber et al. made a split mCherry library in bacteria encoding 31 amino acid residues long peptides where the central amino acid was amber encoded serine. This library was and used in experiment finding novel and known phosphopeptides binding to 14-3-3 proteins, interaction determined by reconstituted mCherry was sorted through FACS. Incorpora-
ition of non-natural amino acids has actually been done in pIII phage display before, where they incorporated several unnatural amino acids to replace tyrosine and then used in directed evolution\textsuperscript{163}. The main disadvantage they faced was the effectiveness of incorporation of the unnatural amino acid compared to the normal tyrosine. Since then extensive work has been done to make the incorporation of non-natural amino acids more efficient. Rogerson et al. optimized an aminoacyl-tRNA/tRNA pair to efficiently incorporate phosphoserine and a non-hydrolysable analogue at a level more comparable to normal amino acid incorporation\textsuperscript{164}. The setup for a phospho-ProP-PD library should be the same as the phosphomimetic to be able to compare WT and phosphorylated directly. Therefore the system needs to incorporate phospho-serine at approximately the same rate as the normal serine to not bias the selection in the amplification step. However supplementing the media with the non-natural amino acid will still be more expensive than using phosphomimetic proteomic peptide phage display.

Part B structural details of PDZ phosphopeptide binding paper III and IV

Having uncovered that Scribble PDZ1 bind to p-3 phosphorylated ligands we decided to explore the structural basis for the phosphopeptide recognition. To investigate the structural reasons for phosphopeptide binding explored the interaction through protein NMR. \textsuperscript{1}H \textsuperscript{15}N HSQC titrations with the free PDZ domain, and the domain with phosphorylated and unphosphorylated RPS6KA2 and MCC were performed. The NMR structure of Scribble PDZ1 bound phosphorylated RPS6KA2 was determined from the \textsuperscript{1}H \textsuperscript{15}N HSQC titrations, but since the peptide was not \textsuperscript{15}N-labeled the residues of the peptide could not be assigned in the structure. The RPS6KA2 peptide binds to the canonical binding site between the $\beta_2$, $\beta_3$ sheets on one side and the $\alpha_2$ helix on the other. The PDZ domain showed a large conformational change in the $\beta_2$ and $\beta_3$ sheets compared to an already published NMR structure of Scribble with no peptide bound (1X5Q Figure 20A). Three surface exposed positively charged residues lining the binding pocket was identified as potentially interacting with the phospho group in the process of binding (Figure 20B). These residues were mutated in to alanine and subjected to ITC titrations with the phosphorylated and unphosphorylated RPS6KA2 (Figure 20C). The K746A mutation made the domain bind tighter to the unphosphorylated and slightly weaker to the phosphorylated ligand compared to the wild type domain, and thereby loosing its selectivity for the phosphopeptide, $K_D^{WT}/K_D^P$ selectivity ratio of 1. R801A mutation decreased the affinity for both the phosphorylated and unphosphorylated peptide but still bound tighter to the phosphorylated peptide compared to the unphosphorylated, $K_D^{WT}/K_D^P$ selectivity ratio of 1.6. The R762A however made the
phosphorylated peptide bind slightly tighter and the phosphorylated significantly weaker shifting the selectivity to prefer the unphosphorylated peptide $K_D^{P}/K_D^{WT}$ reversed selectivity ratio of 2.7. This gives the conclusion the R762 residue is the gatekeeper for interaction with p-3 phosphorylation for Scribble PDZ 1.

Figure 20. A The superposition of the structure of Scribble PDZ1 bound to phosphorylated RPS6KA2 (blue cartoon PDB 6ESP) and the ligand free domain (white cartoon PDB 1X5Q) shows that the binding site opens up on binding the phosphopeptide as compared to the ligand-free domain. B Surface exposed basic residues lining the binding pocket used for mutational analysis. C Affinity of phosphorylated RPS6KA2 (blue) and unphosphorylated RPS6KA2 (red) binding the different mutants of the Scribble PDZ1 domain. Using two-way anova with Holm-Sidak’s multiple comparisons test shows that there is a significant change in the affinity, compared to the wild type domain, towards the phosphorylated and unphosphorylated RPS6KA2. * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$. D the HSQC spectra do not move upon binding of the phosphopeptide in the R762A mutant. E Table of affinities for the different mutants.
In the $^1\text{H}^\text{15N}$ HSQC spectra of the domain bound to the phosphorylated and unphosphorylated MCC and RPS6KA2 peptides you can see certain residues moving (chemical shift) compared to the spectrum of the unbound domain because they undergo structural changes when binding to the peptide. In paper IV we go in to detail in the movement of the residues upon binding to the peptides. We found that the $^\text{15N}$ chemical shift of a number of residues had shifts in different directions if they bound to the unphosphorylated or the phosphorylated RPS6KA2 peptide. One of the residues where we observe this phenomenon is the residue R762 along with 13 other residues in the structure. The large observed shift changes for the RPS6KA2 peptide where the selectivity ratio is $\approx 5$ between phosphorylated and unphosphorylated does not carry over to the MCC peptide there the shift change for the two peptides are very similar and the selectivity difference for the peptides are not that large either $\approx 2$. This can be represented in two dimensions where the $\Delta^1\text{H}$ shift is on the y-axis and the $\Delta^\text{15N}$ shift on the X (paper IV figure 3). A HSQC spectra was made with the R726A mutant and the RPS6KA2 peptide showing that this residue do not move in the spectra anymore (Figure 20D).
Part C Expanding the phosphomimetic ProP-PD to more PDZ domains.

Since we showed the viability of phosphomimetic ProP-PD against Scribble and DLG1 PDZ domains we wanted to explore the generality of PDZ and phosphopeptide interactions. Nine domains from nine different proteins were therefore subjected to phosphomimetic phage display (Figure 21).

The NGS results were analyzed as described previously (Figure 22). The enriched peptides and the selectivity matrix for the nine selections are shown in Figure 22. Like Scribble PDZ1, we found that InaDL6 and SNTB1 prefers phosphomimetic substitution over wild type peptide at p-1 and p-3. In contrast, the PDZ domain of Shank 1 showed a preference for wild type peptides, suggesting that ligand phosphorylation may enable interactions with a subset of domains but disabling interactions with others.

Figure 21. The effect of phosphomimetic mutations in PDZ binding was further explored with a set of nine additional PDZ domains using phosphomimetic ProP-PD. The domains where SNTA1 PDZ, SNTB1 PDZ, DLG3 PDZ2, DLG4 PDZ3, NHERF3 PDZ4, MAGI1 PDZ4, Shank PDZ, InaDL PDZ 6, and MPDZ PDZ10
Figure 22. A and B Phosphomimetic ProP-PD was performed on mime additional domains. InaDL 6 and SNTB1 have similar selectivity matrices compared to the Scribble domains (Figure 16). C MST affinity measurements using FITC labeled peptides at a constant concentration 25-50 nM and varying concentrations of Shank1 PDZ. n=3 error bars is ± SD.
We find that Shank 1 selectively binds wild type RPS6KA2 preferentially to its phosphomimetic counterpart, which is somewhat contrary to previously reported results where Shank have been shown to interact with the RPS6KA kinases. In Y2H screen the phosphomimetic glutamate substitution of RPS6KA3 ETAL and the wild type STAL showed interaction with Shank1 \(^{165}\). The affinity between Shank 1 and the unphosphorylated, phosphorylated and mimetic variants of RPS6KA2 and MCC were determined and indeed Shank 1 PDZ binds tightest to the unphosphorylated variant in both cases (RPS6KA2 \(K_D^P/K_D^{WT} = 4\) and MCC \(K_D^P/K_D^{WT} = 3\)) (Figure 22 and table in Figure 19).

To gain additional information on phosphopeptide binding to PDZ domains we searched the literature and found two recently published structures of PDZ domains is bound to p-3 phosphorylated peptides. One of the structures is of MAGI1 PDZ2 where the phosphorylated RPS6KA1 peptide (VRKLPSTTL \(^{COO^-}\)) \(^{88}\) and the other is of SNX27 PDZ domain bound to the C-terminus of p-3 phosphorylated LRRC3B (PDDITVV \(^{COO^-}\)). In the case of MAGI1 PDZ2, the affinity for the phosphopeptide is essentially the same as for the unphosphorylated peptide, and phosphorylation can therefore be considered to have neutral effect. MAGI1 PDZ2 has a lysine in the \(\beta_3\) sheet one residue earlier in the sheet compared to the Scribble arginine 762, which may provide the domain a tolerance for phosho-binding (Figure 23A and B) in the SNX27 case the affinity conferred by peptide phosphorylation had a selectivity \(K_D^{WT}/K_D^P = 12\). Interestingly the SNX27 PDZ domain does not have an arginine in the site corresponding as Scribble PDZ1 R762. Instead an arginine in the \(\beta_2\) sheet (corresponding to Scribble PDZ A743) seems to be critical for its phosphopeptide binding (Figure 23B). The domains thus accomplished phosphopeptide binding through an alternative way.

To explore the plasticity of the phosphopeptide binding, we investigated if we could restore the preference for phosphorylated RPS6KA2 in the Scribble PDZ1 R762A mutant. We therefore generated the Scribble PDZ1 R762A/A743R double mutant. We find that this drastically reduced the affinity for wild type RPS6KA2 as compared to both the R762A and the wild type domain. This might be the result of charge repulsion between A743R and the basic residues at p-7 and p-6 in the peptide. In contrast the affinity for the phosphorylated peptide was slightly increased and the selectivity of the domain for p-3 phosphorylation was thus reinstated (\(K_D^{WT}/K_D^P = 4\)) (Figure 23B and C). The arginine however is far away to rescue the selectivity for p-1 phosphorylated peptide of MCC (figure 23C).
Figure 23. A A structure based sequence alignment of the PDZ domains used in this study and the domains that have structural evidence for phospho-serine/threonine peptide binding. B Comparisons of the structures of MAGI1 PDZ2 bound to phosphorylated RPS6KA1 (left PDB 5N7F), SNX27 bound to LRRC3B (middle PDB 5E1A) and Scribble PDZ1 (right PDB 6ESP. For illustrative purposes the RPS6KA1 peptide of 5N7F is depicted in the binding site) C binding affinities for phosphorylated and unphosphorylated RPS6KA2 (hollow) and MCC (meshed) of the R762A and R762A/A743R Scribble domain. The Selectivity for p-3 phosphorylation is restored by the double mutation.
Is it possible to determine phosphopeptide-binding preference for PDZ domains based on structure-based sequence alignment?

A structure-based sequence alignment of the PDZ domains used in this study plus SNX27 PDZ domain and MAGI1 PDZ2 PDZ domains is depicted in Figure 23A. Marked with red boxes are the residues that have been used to make Scribble PDZ1 mutants. There are a number of proteins that have a basic residue in corresponding position to Scribble PDZ1R762, among the domains in the alignment. Scribble PDZ2 and 3 have a corresponding basic residue at R762. They also display similar profile in the selectivity matrix for position p-1 and p-3. The DLG1 PDZ 1 and 2 domain have preference for phosphomimetic substitution at p-1 and p-3. The DLG1 PDZ 1 and 2 domain have preference for phosphomimetic substitution at p-1 and seem to be indifferent to mimetic mutation at p-3 also carry a basic residue corresponding to R762.

There are two interesting cases, DLG1 PDZ3 and DLG4 PDZ 3, which have preferences for p-3 and p-1 mimetic substitution respectively, and none of them have any basic residue in the Scribble PDZ1 key position, the MAG1 PDZ2 lysine position nor the SNX27 PDZ arginine position. In structure 2I0I in PDB DLG1 PDZ3 is bound to HPV18 E6 and the distance between the lysine in the loop between α1 and β4 that is 7.8 Å away from a p-3 glutamate on the E6 peptide, which seems too far for electrostatic interaction. SNTA1 and SNTB1 have basic residues at positions corresponding to R762 and A743, which explain their preferences seen in the selectivity matrix. Both NHERF3 PDZ4 and MPDZ PDZ10 have a basic H or K in the Magi1 PDZ2 position that might affect the preference for mimetic binding in the p-1 position. Interestingly, Shank1 have an arginine in the same position as the SNX27 but does not bind preferentially to p-3 phosphorylation ligands, as shown by MST, this might be due to repulsion from a glutamate, which is situated in the MAG1 PDZ2 lysine position.

The conclusion I draw from this is that the R762 of Scribble PDZ1 and R743 of SNX27 seem to be beneficial for binding of p-3 phosphorylated ligands, and R762 also for p-1 phosphorylated ligands. The other residues in the β2, β3 sheets will affect the binding based on charges, hydrogen bonds, and steric hindrance.
Concluding remarks

As a method phage display and proteomic peptide phage display has proved to be a viable method for finding motif-based interactions. The aim of this thesis has been to explore the use of peptide-phage display to capture selectivity differences in motif-based interactions of proteins depending on allosteric binding of calcium, protein engineering and on ligand modification by phosphorylation.

In the cell, interactions need to be dynamic because of the need to relay information through signaling pathways. Many of the dynamic interactions are motif-based. Binding motifs have often moderate affinities towards their targets and subtitle changes in concentration, or modifications, of either the motif or the target of the motif can decide if they will be binding to each other or not. Evolution seems to prefer motif-based interactions of moderate affinity 32. In paper II we explore the specificity changes conferred by artificial evolution towards higher affinity for a given ligand. We show that affinity towards a wide variety of targets gets increased and the specificity is decreased. The conclusion to draw from this is that the investigation of selectivity or specificity changes between engineered domains and their counterparts in the proteome needs to be explored if the goal is to use the domain in cells to act as an inhibitor.

In the first paper we used combinatorial phage display to investigate differences in binding preferences of Sorcin in presence and absence of Ca\(^{2+}\). Through the experimental protocol used we found the same motif for both selections, and thereby failed to distinguish specificity differences between calcium bound and calcium free Sorcin. Through affinity measurements it was determined that a ligand with the selected motif has a 3.5 times higher affinity towards the calcium bound protein, and therefore the ligand is selective for the calcium bound protein. If the selections had been sequenced through NGS we might have obtained sufficient information to discern differences between the between the presence and absence of Ca\(^{2+}\) in the number and the motifs of the peptides that were found in the analysis.

When this project started the literature of how ligand phosphorylations affected PDZ domain-PDZbm interactions was mainly focused on how it can interfere with binding. Well-studied examples of how phosphorylation enables motif-based interactions with proteins such as 14-3-3 and SH2 domains have established crucial roles of the phosphorylation for binding. But the ability of phosphorylation to tune affinity, and to change the selectivity of
the ligand between distinct domains has been less explored. In this thesis I identified a potential phosphorylation switch in the peptide of RPS6KA2 where the unphosphorylated ligand preferentially bind to Shank 1 PDZ domain, and the phosphorylated ligand has a higher affinity for Scribble PDZ1. As more and more information gets published on PDZ PDZbm interactions that are either enabled or unaffected by phosphorylation and domains where the same phosphorylation disable the interaction it will become evident that this is a common way for a cell to regulate cell signaling, not just for PDZ domains.

The effort of discovering binary PPIs through large-scale methods like Y2H or ProP-PD yields a lot of information of what proteins bind to each other independent of PTMs. As there are well over 200 000 confirmed phosphorylation sites in human it is becoming increasingly important to find out what PTMs are important for motif-based protein interactions, in order to understand cell signaling in health and disease. In this thesis I have expanded the scope of phage display to the world of phospho-dependent motif-based protein interactions, an area of research where chemically synthesized peptide libraries, like spot arrays, have been the dominating method for large-scale discovery. The phosphomimetic ProP-PD is readily scalable to include other parts of the proteome, and can therefore work with other types of domains than PDZ domains. The possibility of further developing the approach using genetic code expansion, might be a road worth pursuing, not just for exploring the effects of serine/threonine phosphorylation but for other types of simple PTMs like acetylation or methylation of lysines.

The importance of exploring conditional interactions cannot be stressed enough; I have opened up novel avenues for research. I predict that dedicated proteomic libraries where you look at differences in binding of the same motif in two different states, like phosphorylated and unphosphorylated, will become increasingly useful over the next years to come.
Populärvetenskaplig sammanfattning

Bakgrund


Resultat

Den metod som förenar mina artiklar i min avhandling är fagdisplay. Fager eller bakteriofager är virus som infekterar bakterier. I bakterierna förökar sig sedan viruset genom att göra många kopior av sig själv och ta sig ut genom bakteriens cellvägg.

Fagerna består i princip enbart av ett proteinhölje och DNA som kodar för de proteiner som fagen har. Genom att ändra i DNA sekvensen hos fagen kan man få den att producera modifierade protein, de modifierade proteiner- na kan då visas på fagens yta. Det är enkelt att skapa fager med protein som är modifierade med olika sekvenser och på det sätter skapa sig ett ”bibliotek” av fager som kan bestå av miljontals olika proteinsekvenser i den modifierade delen. Fagiblotken används sedan i fagdisplayförsök där man tillsät- ter fagerna till proteiner som är fastsatta på en yta. Fager som specifikt binder till proteinet kommer stanna kvar när man tvättar bort de fager som inte bundit. Sedan kan några fager som inte bindit specifikt föröka sig i bakterien. Gör man detta ett par rundor så kommer man att anrika fager som binder hårt och specifikt till proteinet. Vilka de är kan man ta reda på genom att sekvensera DNA.

Vi använder oss i de flesta fall av fagbibliotek som innehåller sekvenser som motsvarar sekvenser som finns i proteiner hos människan. Vi i använden dem i fagdisplayselecttioner för att hitta vilka proteiner de binder till, som exempel PDZ domäner. I artikel två använde vi oss av ett fagbibliotek som innehöll peptidsekvenser från människans proteiner. Detta för att hitta vilka proteiner en PDZ domän som har blivit modifierad för att binda hårdare till ett särskilt protein från ett virus som orsakar livmoderhalscancer. Teorin var att eftersom proteinet hade modifierats för att binda starkare till en viss pep- tid skulle den också bli mer exklusiv. Vi fann dock att den modifierade PDZ
domänen band till fler typer av proteiner än den ursprungliga domänen, och att den band till dessa starkare.


Slutsats

Att hitta interaktioner som beror på fosforylering är viktigt för att förstå hur en cell fungerar och reagerar på signaler. Vi har visat att fagdisplay med fosforyleringshärmande aminosyror är ett gångbart sätt att hitta interaktioner som beror på fosforylering. Vi har också visat på hur vissa PDZ domäner kan uppnå starkare bindning till fosfopeptider. Mina resultat kommer förhoppningsvis ligga till grund för att fortsätta utveckla fagdisplay för att hitta interaktioner som beror på modifieringar av proteiner, vilket är viktigt att ta reda på om man ska förstå vad som händer i sjuka celler.
Acknowledgments

There are so many people I want to thank for making this thesis happen.

And of course the first one I want to thank is my supervisor Ylva. Thanks for taking a chance on me as your first PhD student even though I wasn’t really a student in biochemistry. And thank you for the patience with all my stupid questions and the long (pause for emphasis) struggles I had to find reliable, consistent methods for biophysical validations. Thanks also for giving me an outlook to the international community, and the science performed elsewhere in our field, to keep us aiming to do better research. Thanks also to Per my co-supervisor for accepting me to do my thesis as a master student in your lab. With out that opportunity I would not be here today. And with that I would also want to thank Greta and Emma letting me play a small part of the NCBD-CID project.

Thanks to Roland and Chi without your scientific contributions to the research presented in this thesis there would not be much to write about, at least for papers III and IV. The same thanks goes to Philip M. Kim and Clair Jouhyun Jeon for letting me stay in your lab in Toronto and helping us out with the scripts for sorting the NGS results.

Thanks to Andreas for the collaboration in paper II and your enormous enthusiasm for science, data and methods.

Thanks to the scilife drug discovery and development, biophysical screening and characterization platform and Anette, you guys were the answer to the struggles mentioned above.

I also want to thank the members of the Ivarsson group both present and former. Caroline for sharing my office with me for the last year and for all the gummy bears I have stolen. You were the organizational sunshine this group needed. Susanne thanks for sharing my nerdiness for PDZ domains, for the nice discussions about science, politics, Sweden, Denmark and skåne. Also thanks for the remembrance that this is just a job. Thanks to Leandro for always being so positive and so understanding and helpful with all of our questions, and “suggestions” about the sequence data. Thanks to Ali for being my first PhD colleague expanding our small group with new ideas and perspectives. Also thanks with all the help surrounding Cell lab work and western blotting. Thanks to Vikash and Cissi for showing that there is a science career after a PhD. Thanks to my project students Louise, Erik and Phrita, even though nothing that we worked with made it in to this thesis it
was always fun to work with you and I have learnt a lot from supervising you. Thanks to all other project students and visiting PhD students our lab has had over the years, Jana, Illaria, Nastasia, Deria, Lena, Lena, Lena, Dimitri, Elias and the people I will undoubtedly have forgotten in this list.

Thanks to the people that have worked in the biochemistry program, during my time here, that has made my days more fun. Present and previous PhD students: Angelica, Christian, Eldar, Åsa, Huan, Emil, Erika, Dirk, Vladimir, Thilak, Guilia, Edward, Joana, Derar and Daniella; Postdocs and researchers: Nisse, Gun, Sara, Helena N, PeO, Sandra, Khyati, Mustafa; P.Is: Gunnar, Mikael, Doreen, Erik. Special thanks to Francoise and Johanna A for the help with teaching, and to Helena D for letting my supervisor know that this hobby of mine with being a student representative here and there might be worth the time I spent on it. And maybe not so much a thank you as a good luck to the Kamerlin group, I think you will fit grate in to this department.

Thanks to the admin staff Johanna J, Eva, Lina, Mariam and Gunnar, for help and for all the lunch discussions when I want to speak Swedish.

Thanks to the rest of the department, the costume parties were fun, and sorry I had to bore you with endless PhD meetings.


Jag vill också rikta ett tack till alla fantastiska lärare jag haft som elev och student, hade det inte varit för eran entusiasm för min frågvishet hade jag nog aldrig tagit mig hit.

程念彭，考考，Ethan，我爱你。我会永远记得我们第一次见面

Till sist vill jag tacka dig mamma, jag är så ledsen att du aldrig fick uppleva den här dagen. Jag är så stolt och så tacksam för allt du givit mig. Även att det står på flera ställen i det här tacket att den här avhandlingen aldrig hade blivit av utan dem stämmer det There are so many people I want to thank for making this thesis happen.

Best,
Gustav Sundell
References


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1716

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-359434