Peptide-mediated interactions in high-resolution 3-dimensional structures

Amelie Stein

TESI DOCTORAL UPF / 2010

Thesis supervisor: Prof. Dr. Patrick Aloy
Structural and Computational Biology
Institute for Research in Biomedicine, Barcelona

Abstract

Peptide-mediated interactions in high-resolution 3-dimensional structures

Proteins and protein interactions are involved in virtually all processes of life. Here we study interactions between globular domains and short linear motifs, which form a small interface ideal for transient interactions. Despite the small number of contacts involved, these domain-motif interactions (DMIs) are known to be highly specific *in vivo*. We have identified hundreds of instances of DMIs in high-resolution 3-dimensional (3D) structures to analyze the molecular basis of their high specificity. Furthermore, we have derived structural parameters to identify DMIs in 3D structures in a more general, motif-independent way. An important class of DMIs are kinase-substrate interactions. By combining the phosphorylation motif with different kinds of contextual information, we could predict substrates of the human kinase Aurora A. Lastly, we have incorporated DMIs into our database of 3D interacting domains (3did) to disseminate our results to the scientific community for future research.

Interacciones mediatos por péptidos en estructuras tridimensionales de alta resolución

Los procesos moleculares subyacentes a la mayoría de funciones biológicas implican la participación directa de una infinidad de proteínas y múltiples interacciones entre ellas. En esta tesis estudiamos un tipo particular de estas interacciones, de carácter transitorio y altamente específicas, dónde un dominio globular en una proteína reconoce un corto péptido lineal en otra (DMIs). En concreto, identificamos múltiples casos de DMIs en estructuras tridimensionales (3D) de alta resolución y analizamos las bases moleculares de su especificidad. Además, derivamos parámetros estructurales globales que nos permiten identificar nuevos casos de DMIs. Así mismo, y como caso práctico, combinamos el motivo de fosforilación propio de la quinasa humana Aurora A con diversas clases de información contextual para predecir y validar 90 nuevos substratos. Por último, incorporamos las caracterizadas DMIs en nuestra base de datos de interacciones en 3D (3did) con el fin de diseminar nuestros resultados entre la comunidad científica.

Acknowledgements

Much like, as discussed in this thesis, the context of protein interactions mustn't be underestimated, the context of a PhD student is vital for her success. I would like to thank everyone that has contributed to this work, and in particular ...

- my supervisor, Patrick Aloy, for supporting and guiding my work and letting me pursue my own ideas,
- the members of my Thesis Advisory Committee, Carme Caelles, María Macías and Baldo Oliva for reviewing my work, and particularly to the former two for acquainting themselves with this partly rather theoretical matter,
- our collaborators Teresa Sardon and Isabelle Vernos, for introducing us to the world of Aurora kinases, as well as Denis Shields and Rich Edwards for joint work on linear motifs in 3D structures, Miquel Pons for bringing us closer to intrinsically unstructured proteins, Anastassis Perrakis for a project on transcription factors, and Modesto Orozco and Manuel Rueda for collaborative research on dynamics in domain interactions, although these projects are not discussed in detail here,
- my colleagues Manuel Alonso, Arnaud Ceol, Roberto Mosca, Roland Pache, Sasha Panjkovich and Andreas Zanzoni for interesting discussions, proof-reading, food for thought, and coffee breaks,
- the authors of several indispensable and freely available tools, among them Don Knuth for TEX, Leslie Lamport for LATEX, Oren Patashnik for BIBTEX, and Warren DeLano for PyMol,
- and last but not least my family, especially my parents and my husband.

Preface

Proteins are key elements in almost all events of cellular life. They seldom act alone, but rather in interaction with one or more partners. Some proteins form parts of large macromolecular complexes that execute core functionalities of the cell, like protein production or degradation, while others transmit information in signalling networks to co-ordinate these processes. Regulation of cellular functions requires the integration of information from different sources, e.g., the presence of signalling molecules on the cell's surface needs to be considered as well as the state inside the cell, for example, whether the DNA has successfully been duplicated in order to proceed with cell division. This implies that regulation must be sensitive to a number of parameters, and that errors in regulation can lead to diseases such as cancer.

In this thesis, we study the molecular basis of regulatory protein interactions from a structural point of view. Currently, the only source of molecular details of protein interaction interfaces are high-resolution 3-dimensional (3D) structures. Many interactions in signalling networks are transient and occur through the recognition of a short linear motif in one protein by a globular domain in another, forming a small interface. The peptide stretches containing linear motifs are frequently found in unstructured regions of proteins and adopt a well-defined structure only upon binding to the domain. While binding motifs for almost 50 domains had been described, when we began this work, it was assumed that 3D structures of these peptide-mediated interactions were very scarce. However, based on the known motifs we could identify hundreds of these interactions in 3D structures, involving 30 different domains. This allowed us to study the contribution of the motif vs. that of the flanking regions, and to estimate the likelihood of cross-talk for different peptide-binding domains. Furthermore, we have exploited the fact that, while many motifs occur in unstructured regions, they do adopt a well-defined conformation upon binding. This conformation is not only typical for partic-

ular groups of peptides, but shared across families. We have implemented a strategy that identifies domain-peptide interactions in 3D structures based on this characteristic structure, thereby greatly extending the set of domains for which interactions with motifs are structurally classified. Indeed, we have also found instances of peptide-mediated interactions with this approach that had not been described in motif resources so far. We are collecting these interactions in our publicly accessible database of 3D interacting domains (3did). The ready availability of so many structures of peptide-mediated interactions will enable more studies to help understand them, and possibly allow the development of synthetic peptides or drug-like components to interfere with these interactions e.g. in case of disease.

In an additional project, we have applied our knowledge on peptide-mediated interactions to the identification of substrates for the human Aurora A kinase. This kinase plays an essential role in the cell cycle, but only few substrates are known, which cannot explain all the functions that Aurora seems to be associated with. In this case, it was not enough to consider the binding motif alone, as it is found in almost all human proteins. Instead, we focused on proteins involved in the cell cycle processes that Aurora A acts in. Doing so we could reduce the set of putative substrates to less than 100, including several that could suggest novel functions for Aurora. Experimental verification shows that our method has an accuracy of 80%. It is applicable to many other human kinases, thus promising to help reveal several of their substrates and functions.

Abbreviations and Resources

3D	3-dimensional	
3did	Database of 3D interactions	`http://3did.irbbarcelona.org` Stein et al. (2009b)
AD	activation domain (yeast two-hybrid)	→ page 2
AP	affinity purification	→ page 4
BLAST	Basic Local Alignment Search Tool	Altschul et al. (1997)
CATH	Protein Structure Classification: Class, Architecture, Topology, Homologous Superfamily	`http://www.cathdb.info` Cuff et al. (2009)
DBD	DNA-binding domain (yeast two-hybrid)	→ page 2
DDI	domain-domain interaction	→ page 30
DMI	domain-motif interaction	→ page 32
DNA	deoxyribonucleic acid	
ELM	Eukaryotic Linear Motif database	`http://elm.eu.org` Puntervoll et al. (2003)
HMM	hidden Markov model	
IntAct		Aranda et al. (2010)
IUP	intrinsically unstructured protein	→ page 11
MINT	Molecular Interaction Database	Ceol et al. (2010)
mRNA	messenger RNA	
MS	mass spectrometry	
ORF	open reading frame	
PCNA	proliferating cell nuclear antigen	
PDB	Protein Data Bank	`http://www.pdb.org` Berman et al. (2000)

PDZ domain	postsynaptic density 95; discs large; zonula occludens 1 domain	
Pfam	Protein Families	`http://pfam.sanger.ac.uk` Bateman et al. (2004)
PPI	protein-protein interactions	→ page 2
PPII-helix	polyproline-II-helix	
PSSM	position-specific scoring matrix	
pS, pT, pY	phosphoserine, phosphothreonine, phosphotyrosine	
PTB domain	phospho-tyrosine binding	
PTM	post-translational modification	→ page 23
	PyMol	`http://www.pymol.org` DeLano (2008)
RNA	ribonucleic acid	
SH2 domain	Src-homology 2 domain	→ Fig. 1.13
SH3 domain	Src-homology 3 domain	→ Fig. 1.13
SCOP	Structural Classification of Proteins	`http://scop.mrc-lmb.cam.ac.uk/scop` Andreeva et al. (2008)
SLiM	Short Linear Motif	
SVM	support vector machine	
tRNA	transfer RNA	
UniProt	Universal Protein Resource	UniProt-Consortium (2009)

Single- and 3-letter-abbreviations for protein-coding amino acids

A	Ala	Alanine	I	Ile	Isoleucine	R	Arg	Arginine
C	Cys	Cysteine	K	Lys	Lysine	S	Ser	Serine
D	Asp	Aspartic acid	L	Leu	Leucine	T	Thr	Threonine
E	Glu	Glutamic acid	M	Met	Methionine	V	Val	Valine
F	Phe	Phenylalanine	N	Asn	Asparagine	W	Trp	Tryptophan
G	Gly	Glycine	P	Pro	Proline	X		any amino acid
H	His	Histidine	Q	Gln	Glutamine	Y	Tyr	Tyrosine

Table of Contents

List of Figures

1 Introduction

1.1 Proteins and protein interactions

Proteins are key players in virtually all biological events that take place within and between cells. They fulfill a variety of functions, from structural scaffolds in the cytoskeleton or in virus capsids to transport proteins like hemoglobin to regulatory proteins that control the flow of cellular information, such as kinases and phosphatases, or the expression of genes, such as transcription factors. Over the last decade, the availability of large-scale sequencing of DNA, together with the computational detection of open reading frames (ORFs), has given us access to the *genomes* of a number of model organisms, among them the bacterium *E. coli*, baker's yeast (*Saccharomyces cerevisiae*), worm (*C. elegans*), fly (*Drosophila melanogaster*), zebrafish (*Danio rerio*), mouse (*Mus musculus*), and human (*Homo sapiens*). The Ensembl database, which focuses on vertebrate genomes, currently supports over 50 species (Flicek et al., 2010), while Ensembl Genomes lists far more than 100 species across the taxonomic space (Kersey et al., 2010). Having the list of proteins and their sequence has in turn allowed the study of *proteomes*. But to understand the function of a protein, knowing the sequence is not enough. For example, structural information often provides crucial information about how a protein works, ideally in the form of high-resolution three-dimensional (3D) structures which reveal atomic details. Furthermore, proteins seldom act in isolation – they often accomplish their function as part of large molecular machines or in intricate regulatory networks of transient protein-protein interactions that regulate the activity of those molecular machines. Therefore, to understand how a cell works, it is not sufficient to know the list of its parts, but we need to understand how the individual components interact – in other words, we need to put them into context.

Proteome: The set of proteins in a given cell type or tissue, may also denote the full set of proteins encoded by an organism's genome.

1.1.a) Protein interaction detection

Literature curation of protein interaction data

For many years, studies of the interactions of individual proteins have been performed using classical genetic, biochemical and biophysical methods of interaction detection. In literature curation efforts, many of these interactions have been introduced into databases such as IntAct (Aranda et al., 2010) or MINT (Ceol et al., 2010), to make them available to the scientific community (Orchard et al., 2007). In the last decade, several methods for the detection of protein-protein and protein-DNA interactions have been developed and applied to large datasets, up to whole proteomes. To a lesser extend, protein-lipid and protein-small molecule interactions have also been screened. The work in this thesis focuses on protein-protein interactions (PPI), and the two most commonly used methods for large-scale protein interaction detection are described below. Studies using these methods have provided us with data on the protein interaction networks, or *interactomes*, of several model organisms, among them yeast, worm, fly and human.

Yeast two-hybrid

The yeast two-hybrid method was developed by Fields and Song (1989). It is based on the idea that two independently inactive halves of a transcription factor, when brought together by interacting proteins, will activate transcription of the specific reporter gene. In a yeast two-hybrid setup, a transcription factor is split into two parts, the activation domain (AD) and the DNA-binding domain (DBD), thus exploiting the modularity of globular domains. To test whether proteins X and Y interact, X is fused to the DBD, and Y is fused to the AD (Fig. 1.1). If X and Y do interact, they will bring the two parts of the transcription factor together and thus activate transcription of the genes targeted by the specific DBD, which can easily be detected. There are several possibilities for the reporter gene activated upon transcription; some color the cells hosting positive interactions, while others allow them to live by encoding an essential amino acid or nutrient. Based on the construction of the assay, yeast two-hybrid should detect direct interactions between proteins X and Y. Nevertheless one cannot exclude that a third protein, Z, mediates the interaction, so that the actual binary interactions are X:Z and Y:Z. Conversely, it is also possible that fusion to the AD or DBD domains blocks the interac-

Domains and modularity → 1.3.a), page 18

tion interface of X and Y, such that an interaction is not observed although possible between unmodified proteins (Aloy and Russell, 2002b). Moreover, the interactions tested with yeast two-hybrid happen in the nucleus, where transcription can be activated. However, this is not the native environment for many tested proteins, meaning that they may have different interaction properties, potentially resulting in false positives or false negatives. Despite these limitations, the method has been applied successfully to a variety protein sets from different organisms (Fields, 2009) which was possible because yeast two-hybrid assays are applicable to large numbers of proteins in parallel. Since the first large-scale yeast two-hybrid experiments in yeast in 2000 and 2001, thousands of interactions have been elucidated (Uetz et al., 2000; Ito et al., 2001), leading to the first drafts of the yeast interactome. A number of other model organisms have also been screened, including human (Stelzl et al., 2005; Rual et al., 2005). Similarly, proteins of interest for a specific tissue or disease can be tested for interactions among themselves as well as for PPI with other proteins. Network analysis is promising to help understand complex diseases better and, eventually, identify therapeutic approaches (e.g., Pujana et al. (2007); Zanzoni et al. (2009)).

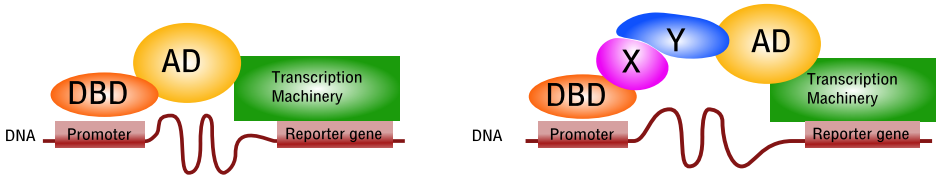

Figure 1.1: The yeast two-hybrid system for experimental interaction detection. (left) A transcription factor binds with the DNA-binding domain (DBD) to it specific promoter. The activation domain (AD) then recruits the transcription machinery, leading to transcription of the reporter gene. (right) In the yeast two-hybrid system (Fields and Song, 1989), to test whether proteins X and Y interact, X is fused to the AD and Y is fused to the DBD. If X and Y interact, transcription can be observed. Illustration inspired by Fields (2009) and Aloy and Russell (2002b).

Domain-based yeast two-hybrid

Interaction detection methods like yeast two-hybrid show whether two proteins interact, but not which parts of the protein mediate the interaction, or, more exactly, which is the interaction interface. A recent twist is the use of protein domains instead of full proteins in yeast two-hybrid assays (Boxem et al., 2008). This is based on the observation that many PPI are mediated by interactions of globular domains in those proteins. It is possible to

Domain-domain interactions
→ 1.4.b), page 30

split proteins into their domains because the latter usually fold independently – random protein fragments may not be amenable to such an interaction discovery procedure because they often cannot adopt a functional structure. Domain-based yeast two-hybrid allows to detect which part of a protein mediates the interaction, or at least delimits the region in which the interface lies. To achieve this higher resolution, one construct of each fragment of protein X fused to the AD is required, combined with one construct of each fragment of protein Y fused to the DBD, so that the number of experiments is much higher than for a simple test of whether X and Y interact.

Affinity purification

While yeast two-hybrid serves the detection of binary protein interactions, affinity purification (AP) methods are most suited for the discovery of protein complexes. They work by fusing a tag to the protein of interest, which can be used for purification, and expressing this hybrid protein in the cell. Proteins bound to the tagged one (with sufficient strength) are co-purified along with it, and subsequently the "complex" is analyzed by mass spectrometry (MS) (Puig et al., 2001). Affinity purification has been applied on a genome/proteome scale to protein complexes in yeast (Gavin et al., 2006; Krogan et al., 2006), where hundreds of known and novel protein complexes could be identified. Human proteins have been searched for interactions among co-purified complexes using a similar method (Ewing et al., 2007). The difficulty in interpreting AP results in terms of binary protein interactions is that having X, Y, Z and A in the same complex does not mean that they all interact directly, i.e., there isn't necessarily a molecular interface between each possible pair. Two models are currently used for the expansion of AP data to binary interactions, the spoke model and the matrix model (see Fig. 1.2, center and right) (Bader and Hogue, 2002). In the spoke model, each co-purified protein is assumed to interact with the tagged protein, while in the matrix model, all possible interactions among the purified proteins are assumed to exist. As Fig. 1.2 shows, neither model necessarily produces an accurate representation of the interactions within the complex. In principle, it should be possible to test for binary interactions among the members of a complex with other techniques; however, some pairs of proteins do not necessarily interact independently, but need others to stabilize them. Thus, while AP is very useful in

Expansion to binary interactions: spoke and matrix model

discovering strongly connected groups of proteins, mappings to binary interaction networks should be handled with caution. Like with yeast two-hybrid, if the purification tag disturbs or blocks an interaction interface, the respective binding partners will not be found. In addition, ribosomal or heat shock proteins may be co-purified if some protein in the complex was not fully translated at the time of cell-lysis, although they have no functional connection to the complex (Aloy and Russell, 2002b). To address the latter issue, measures such as the socio-affinity index (Gavin et al., 2006) and interaction confidence scores (Ewing et al., 2007) have been developed that estimate whether an association between two proteins is likely to be functional.

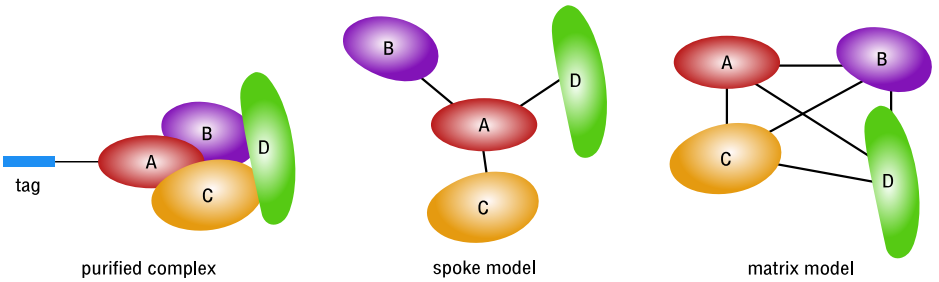

Figure 1.2: Experimental interaction detection: affinity purification. Protein A is expressed in the cell fused to a tag (light blue) that is used for purification after cell lysis. Proteins interacting with A with sufficient strength are co-purified along with it. The two most common models for generating binary interactions from co-purification data are the spoke model (center) and the matrix model (right) (Bader and Hogue, 2002). As the illustration shows, both models lead to the assumption that there is a direct interaction between proteins A and D, which is not the case in complex shown on the left. Therefore, to get more reliable data on whether two members of a complex actually interact, additional information should be used, such as scoring based on frequent co-occurrence of two proteins (Gavin et al., 2006) or data from additional experiments (e.g., Hernández and Robinson (2007)).

1.1.b) Interactome networks

The data produced by high-throughput interaction discovery studies using methods like yeast two-hybrid or AP has been combined with the literature-curated data from individual experiments to create interactomes for individual species. For some model organisms, most notably yeast (*Saccharomyces cerevisiae*), a large number of interactions have already been observed, especially

when considering the size of the proteome (Fig. 1.3). For other species, though, many interactions remain to be detected (Table 1.1). None of the current methods for interaction detection can be expected to be successfully applied to all proteins in a proteome. Membrane proteins, for instance, have a different native environment than proteins occurring in aqueous solutions, and thus do not behave well in either of the methods described above. Special protocols have been developed to handle them (Kittanakom et al., 2009; Gavin et al., 2006). Furthermore, depending on the method repeated screens may be required for high-throughput interaction discovery techniques before a result can be considered reliable. Based on the coverage of individual studies and on comparisons of results obtained using the same as well as complementary methods of interaction discovery, researchers have estimated how many interactions should exist in some model organisms (von Mering et al., 2002; Goll and Uetz, 2006; Venkatesan et al., 2009). However, although current interactome networks are far from complete, they illustrate the connectivity and complexity of the events inside a cell. Despite some issues with the widely used interaction detection methods described above, these techniques have given us access to vast amounts of data that would otherwise remain elusive.

Organism	Number of genes	Number of proteins	Number of interactions
yeast (*Saccharomyces cerevisiae*)	6,000	6,000	60,000
worm (*C. elegans*)	19,000	22,000	6,000
fly (*Drosophila melanogaster*)	13,000	18,000	19,000
human (*Homo sapiens*)	∼25,000	∼70,000	53,000

Table 1.1: Genome, proteome and interactome sizes for selected model organisms. The estimated number of genes is taken from `http://www.ornl.gov/sci/techresources/Human_Genome/faq/compgen/shtml` Proteome estimates are based on UniProt (UniProt-Consortium, 2009), Flybase (Wilson et al., 2008), Wormbase (Harris et al., 2010) and SGD (*Saccharomyces* Genome Database, Weng et al. (2003)). The interactome sizes given here refers to the number of interactions currently discovered (Aranda et al., 2010; Ceol et al., 2010; Prasad et al., 2009), which in many cases is only based on a subset of the proteome, and incomplete due to other reasons as e.g. discussed in Venkatesan et al. (2009).

1.2 Protein structure and disorder

High-throughput interaction discovery experiments as those described above indicate only that two proteins interact, but not how this happens. Some more

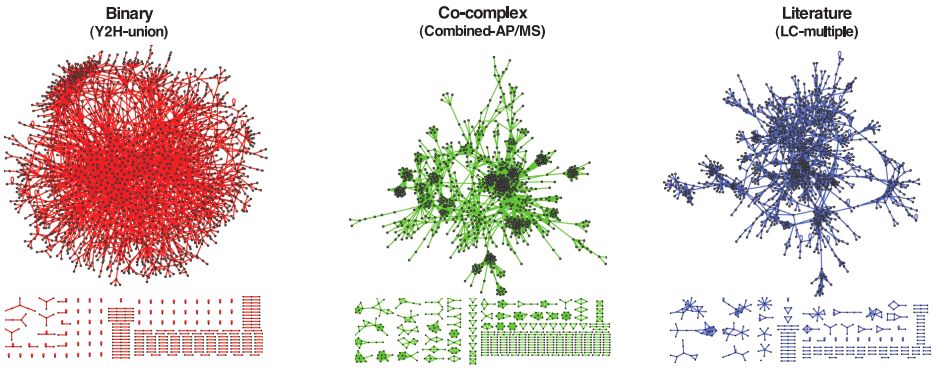

Figure 1.3: The current state of the yeast interactome. Yu et al. (2008) compared the yeast protein interaction networks based on high-quality datasets from yeast two-hybrid, affinity-purification-based methods and literature-curated data. They estimate that current data covers about 20% of the actual yeast interactome.

recent variants, such as the domain-based yeast two-hybrid method (Boxem et al., 2008), can delimit the region of the proteins that are involved in forming the interface. However, neither is able to provide information about the atomic details or the molecular mechanism of the interaction. Currently, this level of detail can come only from high resolution three-dimensional (3D) structures, in which the residue-contacts are resolved and the protein interaction interfaces characterised. This information is collected in the Protein Data Bank, a worldwide repository for 3D coordinates of proteins, nucleic acids and other biomolecules.

The *primary structure* of a protein is its sequence of amino acids. It encodes the higher-order structures, the *secondary structure* – α-helices and β-sheets, which are locally defined substructures – and the *tertiary* or 3D structure of the protein, which is the characteristic fold assumed by the backbone. In the case of protein complexes, the *quaternary structure* describes how the components relate to each other, which includes the atomic details of the interaction interfaces. The first 3D structures of proteins were solved 50 years ago and greatly influenced the work of scientists in many areas of the life sciences, as they revealed molecular details that allowed an interpretation of their function that was not attainable before. Because the structural characterization of proteins is often very time-consuming, it is not feasible to solve

Protein-protein interfaces in 3D structures → 1.4, page 28

structures of all proteins, let alone all possible interactions or complexes. It may be possible, though, to model structures based on those of sequentially similar proteins (Chothia and Lesk, 1986) or to derive general principles of protein structures and use those to predict structures of new proteins. It has also become clear that, since proteins are not rigid structures, knowing their dynamics often is crucial for understanding their functioning (e.g. Rueda et al. (2007)). Furthermore, some proteins may not assume a well-defined structure, but rather fluctuate among an ensemble of possible conformations. The recent focus on dynamic and unstructured proteins has influenced the field of structural biology in that, for many proteins and complexes, multiple structures will be required to understand the function and dynamics of the components, and that the importance of methods which can capture proteins in movement has increased. Below, the most commonly used techniques for structure determination and modelling are briefly outlined.

1.2.a) Protein structure determination

X-Ray crystallography

X-ray crystallography is based on the fact that X-rays interact with the electrons in molecules, including macromolecules such as proteins. This leads to a specific pattern of scattered X-rays. However, the scattering of an individual protein is very weak. To strengthen the signal and at the same time retain the regularity of the scattering pattern, the molecules must be arranged in a very regular fashion – in a crystal. This in turn means that very high concentrations of the protein or complex to be crystallized are required, and that appropriate conditions need to be found in which a regular crystal forms. In addition, not all crystals yield scattering patterns that are suitable for an interpretation of their molecular details (Fig. 1.4). In some cases, for example, the electron densities may only allow tracing the backbone or identifying the fold, but not the positioning of the individual side chains (Branden and Tooze, 1999). The electron density maps are often discontinuous, which makes knowing the protein sequence(s) essential for their interpretation. If the conformations of the molecule differ among the elements in the crystal, the scattering patterns do not coincide, and thus the electron densities will not be well-defined at these sites. This impedes the identification of the structure of these regions, which

is frequently observed in loops connecting stable elements of the protein core. For the same reason, IUPs cannot be studied using X-ray crystallography. The fact that structures are derived from proteins in a rigid crystal and at high concentrations may also mean that some proteins assume a conformation that is not their native one. In addition, the tight packing of molecules in a crystal may lead to the observation of contacts that are not related to the biological function of the proteins. Some proteins have several alternative conformations, e.g., depending on interaction partners or catalytic activity. In principle, each of those conformations can be observed if a crystal of this conformation can be produced. An advantage of X-ray crystallography is that, given suitable crystals, there is no size limitation. This has allowed the structural analysis of entire virus capsids or large macromolecular machines such as the proteasome (Groll et al., 1997) or the ribosome (Ban et al., 2000). Currently no other method is able to provide atomic-level details of such large structures.

Nuclear Magnetic Resonance spectroscopy

The requirement for crystallization create difficulties in the structural study of proteins using X-rays, partly as it may be time-consuming to identify the conditions under which a given protein or complex crystallizes well, but more so because some regions or even entire proteins do not adopt stable structures that are amenable to crystallographic studies, such as intrinsically unstructured proteins (IUPs). Nuclear Magnetic Resonance (NMR) spectroscopy is applied to molecules in solution – a more natural environment than a crystal – and thus offers a way to study these proteins. In addition, it allows the study of protein dynamics, and thus structures solved by NMR often provide different conformations (see Fig. 1.5, page 12). The method is based on measuring the spin (magnetic moment) of some atomic nuclei, including ^{1}H, which is abundant in proteins, and ^{13}C and ^{15}N, which can be introduced by growing the protein-expressing cells on media enriched in these isotopes. Radio frequency pulses, applied to molecules in a strong magnetic field, cause them to emit radiation while transitioning back to the equilibrium state, the so-called chemical shift. This data allows identification of the atom's neighborhood, with different radio frequencies providing different kinds of information. Two frequently used types are COSY (correlation spectroscopy), which

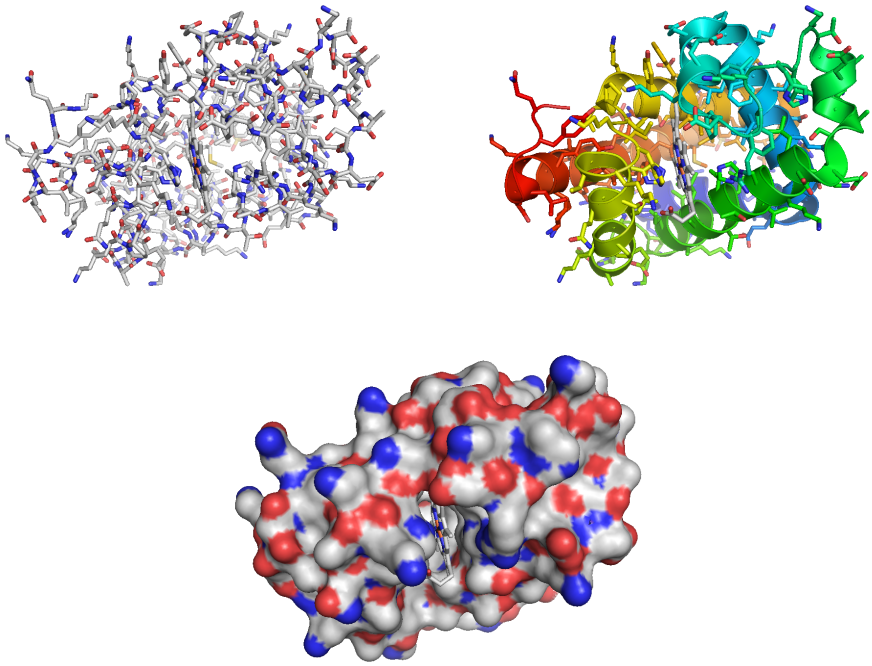

Figure 1.4: Different visualizations of a protein structure. The 3D structure of myoglobin (PDB:2ekt, Harada et al. (2007)) is shown in three different visualizations here, which are all in the same orientation. (top, left) The positioning of the individual side chains is only possible at high resolution. (top, right) The backbone of the protein, here colored in a rainbow scheme from blue at the N-terminus to red at the C-terminus, can also be observed in structures of lower resolution. The position of residues and contacts, which may be critical for a functional interpretation, are often missing at low resolutions though. (bottom) The surface visualization reveals cavities such as that for the heme in this structure, which are difficult to deduce from other visualization.

gives information about covalently linked, i.e., sequentially adjacent residues, and NOESY (nuclear Overhauser enhancement spectroscopy), which reveal residues that are close in space, independent of whether they are close in sequence or not (Branden and Tooze, 1999). The analysis of NMR data is not trivial, as many peaks in the spectra overlap, so that unequivocal assignment of residues to peaks is not always possible. The work of Kurt Wüthrich and colleagues on this problem has enabled the application of NMR to proteins

(Braun et al., 1983) and was awarded with the Nobel prize in 2002. Like for X-ray crystallography, it is essential to know the amino acid sequence for the interpretation of NMR data. Because of the complexity of the spectra, the size of proteins amenable to NMR is usually restricted to relatively small proteins. Nevertheless, some large protein assemblies have been studied with NMR using special techniques, such as the GroEL-GroES chaperone (Fiaux et al., 2002) or the proteasome (Sprangers and Kay, 2007). In the latter case, the molecule's high symmetry was exploited to make the analysis of the spectra possible (Bax and Torchia, 2007).

Other techniques for structure determination

Other methods for determining the structure of include cryo-electron tomography (e.g., Frank (2002); Nickell et al. (2006)) and small angle X-ray scattering (SAXS, e.g. Neylon (2008)). Electron tomography is used to study very large structures, such as the capsoids of viruses or molecular machines composed of dozens of proteins (e.g., Beck et al. (2007)). SAXS is applied to proteins in solution, making it suitable for the study of IUPs. For example, it can be used to identify the architecture of domains in a protein and their interactions. However, neither of these methods is currently able to provide the atomic detail of interaction interfaces required in this work.

1.2.b) Intrinsically unstructured proteins

While structures solved by X-ray crystallography appear to be rigid, those solved by NMR show that proteins do not necessarily assume one well-defined conformation in their native state, but that they are flexible molecules, and that some regions show more variation than others (cf. Fig. 1.5 on page 12). Indeed, this lack of a defined structure had already been observed in small regions of crystallographic structures, such as loops that connect the stable α-helices and β-sheets in the cores of protein folds, where the electron densities in the corresponding regions were not well-defined. Initially scientists assumed that these regions were merely linkers between the well-structured parts, and that only the latter had functional relevance. This assumption arose, at least in part, from the observation of several structures that clearly implied the

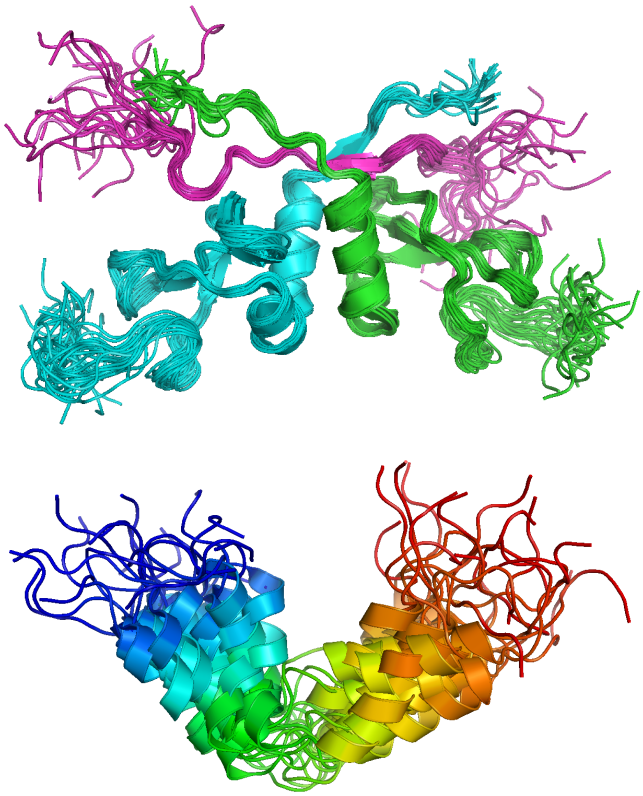

Figure 1.5: Flexibility in protein structures solved by NMR. (top) A chromo_shadow domain dimer (green and cyan) binding a P.V.L peptide (magenta) at the dimer interface (PDB:1s4z). These domains are found in the heterochromatin-associated proteins and are involved in the regulation of gene expression (Thiru et al., 2004). The figure shows that, while the core helices and the binding peptide are structurally well-defined, the tails exhibit considerable flexibility. (bottom) The Arf protein is a tumor suppressor that regulates p53 function. It is mostly unstructured under some conditions, but forms two helices connected by a flexible linker as shown here under others. The structure is colored in a "rainbow" scheme, from blue at the N-terminus to red at the C-terminus (PDB:1hn3, DiGiammarino et al. (2001)).

The structure-function paradigm

protein's function, especially among initially solved structures (Fersht, 2008). However, more recent research has shown that disordered or unstructured regions have great functional importance, e.g. as regulatory proteins, and

that a reassessment of the correspondence between a well-defined 3D structure and a protein's function, the so-called structure-function paradigm, is required (Wright and Dyson, 1999; Tompa, 2002). Also, it is not only short regions that can be unstructured: more than half of the eukaryotic proteins are estimated to contain long unstructured regions (>30 residues), and about 20% of all eukaryotic proteins are expected to be fully unstructured (Dunker et al., 2001; Oldfield et al., 2005). While their lack of structure does not permit Intrinsically Unstructured Proteins (IUPs) to fulfill functions such as those of metabolic enzymes, or transport proteins like hemoglobin, which are tightly bound to stereochemical properties (Bhattacharyya et al., 2006), it is precisely their natively unstructured nature that allows them to bind multiple different partners without the need to undergo energetically costly transformations. Therefore IUPs are ideal candidates for regulatory proteins (Tompa, 2002), which are particularly important in higher organisms that need to integrate a multitude of signals. IUP and unstructured regions often contain many *linear motifs* that bind to specific *recognition domains* which are typical in signal transduction (Dunker et al., 2005). Upon binding to a partner protein, IUPs often do assume a well-defined structure. Two mechanistic models have been suggested for how the protein adopts the conformation for the interaction: the conformational selection model, in which the conformation suitable for this interaction exists among an ensemble of structures, and the induced folding model, in which the bound conformation becomes defined by binding to the partner (Stein et al., 2009a). Because of their flexibility, the same region of an IUP can bind different partners in different conformations (Freund et al., 2002; Demarest et al., 2002). It has been observed that the distribution of amino acids found in the PDB differs from that in the proteome (Gerstein, 1998). This may be due to the fact that most structures in the PDB have been solved by X-ray crystallography (Fig. 1.6), to which IUPs on their own are not amenable as they will not crystallize. NMR studies allow a glimpse on the ensemble of structures encoded in an IUP. If the IUP adopts a well-defined structure upon binding to a partner, and this complex can be crystallized, the interaction can also be analyzed by X-ray crystallography. From an evolutionary point of view, proteins with a well-defined structure that enables their function are under strong constraints, as any mutations need to keep this structure functional and retain its ability to fold properly, while destabilizing/destructive mutations often lead to non-viable organisms. As examples such as sickle-cell anemia show, even single point mutations can

Linear motifs → 1.3.b), page 21

Conformational selection and induced folding

IUPs and evolution

have a strong effect on the 3D structure of a protein (Branden and Tooze, 1999). Unstructured regions, on the other hand, are less limited by such constraints, and are thus much more tolerant towards mutations. Hence, changes are found more often in unstructured regions than in other regions of proteins. Thus, IUPs are more flexible not only from a structural, but also from an evolutionary point of view. It is also interesting to note that IUPs are much more common in eukaryotes than in eubacteria and archaea, indicating that they are a relatively recent development in evolutionary terms (Dunker et al., 2000; Dunker and Obradovic, 2001). This correlates with the increased complexity of higher eukaryotes, particularly in terms of regulation, which became necessary in particular with the advent of multicellular organisms. In the context of protein interaction networks, it has been observed that many hubs are IUPs and/or have IUPs as interaction partners, which again reinforces the observation that IUPs are important in regulatory networks (Dunker et al., 2005; Uversky et al., 2008). It has been suggested that the flexibility of IUPs makes them key for the evolution of protein interaction networks (Stein et al., 2009a). Furthermore, a number of IUPs have been found to be implicated in diseases, especially relating to problems with cellular regulation, as it happens e.g. in cancer, but also in diseases caused by protein aggregation, such as Alzheimer's (Uversky et al., 2008; Stein et al., 2009a).

Hubs: Proteins in interaction networks that have many different interaction partners.

1.2.c) Modelling

As pointed out above, structural characterization of all proteins is not feasible. However, based on knowledge extracted from the already solved structures it may be possible to predict the structures of new sequences, although details like side chains cannot always be placed well. Methods attempting to tackle this problem can be split into two classes: comparative methods are based on libraries of known structures and try to identify a good template, while *ab initio* or *de novo* methods try to predict the protein's structure from scratch. Both types rely on an energy function to evaluate the model they have created, with the goal to minimize the global energy. Comparative methods are based on the concept that, while a multitude of sequences exist, the number of 3D folds they assume is actually relatively limited and has been estimated to lie around 1,000 (Chothia, 1992). Around 700-800 folds are already structurally characterized (Kiel et al., 2008), and *structural genomics* initiatives such as

the Protein Structure Initiative (PSI, Dessailly et al. (2009)) and Structural Proteomics in Europe (SPINE, Alzari et al. (2006)) attempt to increase the coverage of different folds or domains by concentrating on proteins for which no template with similar sequence can be identified. The two commonly used comparative approaches are *threading* and *homology modelling*. In threading, the sequence is fitted ("threaded") into many different folds, and each of these relatively rough models is evaluated in order to estimate whether it is acceptable from an energetic point of view. In homology modelling, the structure of a homolog (related protein) is identified via sequence alignment. A detailed model is then built based on this template (Martí-Renom et al., 2000; Eswar et al., 2008). Loops and other unstructured regions pose difficulties for both methods, as they are much more flexible than the core of a fold and may vary considerably in length. In *ab initio* protein structure prediction, the main problems are the large number of possible conformations that individual amino acids can take with respect to each other (Levinthal, 1969), and the fact that the interplay of the physicochemical forces involved in folding still are not fully understood (Service, 2008). Some *ab initio* methods mimic folding, e.g. by first assembling local structures such as α-helices and β-strand that can then be combined into more complex structures (e.g., Bradley et al. (2005)). Others implement stochastic methods to search the space of possible conformations efficiently. Either approach requires vast computational resources such as supercomputers or distributed computing, and is therefore currently only feasible for very small proteins (Service, 2008). Current methods for structure prediction aim to generate models of well-defined structure, on which they may regularly prove themselves in the Critical Assessment of Technique for Protein Structure Prediction competition (Moult et al., 2007). IUPs or proteins that require interaction partners or chaperones to fold are more difficult to predict.

Protein domains → 1.3.a), page 18

1.2.d) The Protein Data Bank

Virtually all structures solved with either of the methods described above are submitted to the Protein Data Bank (PDB), a repository for structures of biomolecules that was created in the 1970s at the Brookhaven National Laboratory in Upton, NY, USA, with the intention to make atomic 3D coordinates of proteins and nucleic acids available to the scientific community. Initially,

the structural data was distributed on magnetic media, nowadays the worldwide PDB (wwPDB) has mirrors in Europe, Japan and the USA (Berman et al., 2000, 2007). Virtually all scientific journals require coordinates be submitted to the PDB before the corresponding publications are accepted. As Fig. 1.6 shows, the number of new structures per year has increased dramatically since the early years, mainly due to technical advances such as cloning and expression of protein in bacteria, providing the large amounts required for purification and structural studies, or the automatized testing of multiple conditions in parallel to identify the appropriate ones more quickly. Another reason for growth are the structural genomics projects mentioned above. Other initiatives such as 3Drepertoire (`http://3drepertoire.org`) have gone yet another step further and attempt to solve structures of interactions, so that not only folds but also interaction interfaces are becoming available at a higher pace. Nevertheless, the coverage of proteins and complexes that are difficult to handle with current techniques for structure determination, such as membrane proteins and IUPs, remains low.

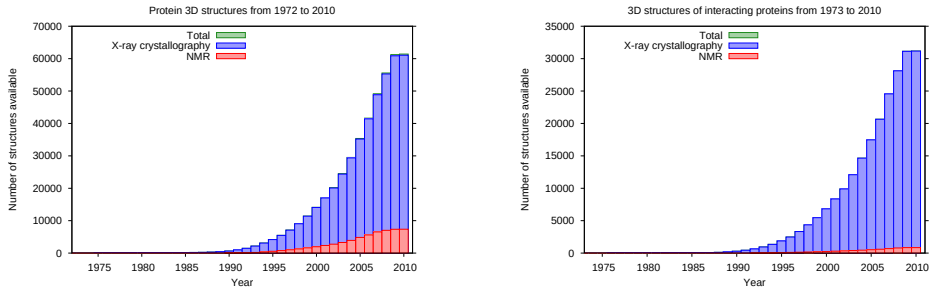

Figure 1.6: Growth of the Protein Data Bank (PDB). (left) Total number of 3D structures of proteins and protein interactions available in the PDB. The vast majority of structures is solved using X-ray crystallography (blue), and about 10% using nuclear magnetic resonance (NMR) spectroscopy (red). The number of 3D structures solved by other methods (visible part of the green bar indicating the total number of structures) is marginal. (right) About half of the 3D structures contain protein interactions. The fraction of interactions solved by NMR is even smaller, which may be due to the fact that NMR studies are usually limited to a relatively small number of amino acids, as the spectra tend to become very complicated. Note that the data for 2010 only covers the first quarter of the year.

1.3 Modularity in proteins and protein interactions

Recent work indicates that novel phenotypes rarely arise from radically new proteins, but instead from new combinations of available functionalities, or by rewiring existing pathways (Bhattacharyya et al., 2006; Carroll, 2005). This is in line with the observation that the number of genes does not increase as dramatically as one might expect from the different complexity in their phenotypes/appearance. While the proteome sizes increase more than the genome sizes (see Table 1.1), the number of protein families is relatively constant; only a few domain families are specific to higher organisms (Bhattacharyya et al., 2006). A notable example is the SH2 domain, which is almost exclusively found in multicellular organisms and is discussed further below.

Modularity has been observed in several biological systems. For instance, promoters are physically distant of the genes they control (see also Fig. 1.1). Furthermore, they are also separate entities functionally speaking: placing a new gene under the control of an existing promoter will activate its transcription along with that of the others. Likewise, adding a new promoter upstream of a given set of genes will create a new regulatory input for their activation (Kirschner and Gerhart, 1998). Classic metabolic enzymes integrate the ability to bind their substrate and their catalytic functionality all in one protein, and especially the latter is tightly coupled to precise stereochemical requirements. While this allows for efficient reaction rates, the evolution of new functionalities based on such a template is difficult. Similarly, regulation through allosteric effectors is intimately coupled to an enzyme's structure, and mutations to allow different effectors would likely impair the structure's stability and/or functional activity (Bhattacharyya et al., 2006). In other words, due to the integration of substrate binding and catalytic activity, such enzymes have little modularity, and the individual aspects cannot be modified independently. There are much more modular proteins, though: they contain one or more globular domains with a variety of functions, and the flexible linker regions between these domains often contain *linear motifs* which bind to specific domains, thereby mediating interactions between different proteins as well as also intramolecular binding events.

1.3.a) Protein domains

Protein domains can be thought of as "building blocks" of proteins. They are usually between 50 and 300 residues long and tend to have a compact, often globular structure and usually fold and evolve independently (Ponting and Russell, 2002; Jin et al., 2009). Domains have a variety of functions, from catalysis such as kinases and phosphatases to recognition of small molecules, lipids or peptides, which may be post-translationally modified. This modularity makes them ideal building blocks for proteins. Many bacterial proteins contain only one domain, while multidomain proteins are frequently found in higher eukaryotes (Fig. 1.7 on page 20, see also Fig. 1.13 on page 34). In many cases, a domain's N- and C-terminus are close in space, which allows their insertion into other proteins (at appropriate sites, e.g., in surface loops or in linker regions) without disrupting the host protein's structure (Stein et al., 2009a). The independently folding nature of protein domains allows studies to focus on them rather than the full proteins, which may be advantageous if the full protein is too large to handle with a given technique, or to narrow down regions of the protein that are involved in a particular function, such as in the domain-based interactome of *C. elegans* (Boxem et al., 2008). Using combinations of domains with different functions as the building blocks, new proteins with novel functions can be created relatively easily by common evolutionary events such as duplication, recombination, insertion and deletion. The different independent properties combine to generate a new, specific function, with e.g. recognition domains to bind a substrate and a catalytic domain to modify it. This allows the same functionality, e.g., a catalytic process, to be used in multiple different scenarios in the cell, each coupled with a different (set of) domain(s) to recruit the appropriate substrate (Bhattacharyya et al., 2006). The function of the domains in multidomain proteins is usually independent, although some fine-tuning can occur over the course of evolution. An example of this is discussed with the intramolecular inhibition of the Src kinase in section 1.4.c), page 35.

Post-translational modifications → 1.3.c), page 23

Several resources for protein domain annotations are available. The Structural Classification of Proteins (SCOP) contains manually annotated 3D structures, grouped by fold, superfamily and family (Andreeva et al., 2008). Similarly, CATH classifies proteins in their hierarchy by Class, Architecture, Topology and Homologous Superfamily (Cuff et al., 2009). These domains are based on

the observation of repeatedly occurring folds among the known 3D structures of proteins. The advantage of manual, structure-based annotation is that it can capture domains that are discontinuous in sequence or otherwise show little similarity on the sequence level. For the same reasons, it is very difficult to reliably transfer such classifications to the large number of instances where no 3D structure is available. SMART (Simple Molecular Architecture Research Tool) (Letunic et al., 2009) and Pfam (Finn et al., 2010) offer sequence-based definitions of protein domains. The domain definitions are created from multiple sequence alignments, which means that they represent conserved elements found in several proteins or families. With large-scale sequencing projects, much more protein sequences are available than structures, so that these domain definitions are built on a broader, and probably less biased, basis than domains curated from structures. The domain definitions are captured in Hidden Markov Models (HMMs), which can be used to search any sequence for occurrence of instances of the respective domains. Both Pfam and SMART have manually refined some of their domain definitions and incorporated these modifications into the HMMs, so that highly reliable definitions are available for some domains. If 3D structures for a domain are available, it is linked to the PDB. However, with 3D structures available for just over 4,000 of the almost 12,000 domains currently in Pfam, structural data is missing for about two thirds of the cases. It has also been shown that over a third of Pfam domains contain at least short unstructured regions (Chen et al., 2006), indicating that these may be more flexible in adapting to different binding partners than domains with well-defined 3D structures in the native state, and that it may not be easy to get a 3D structure of them.

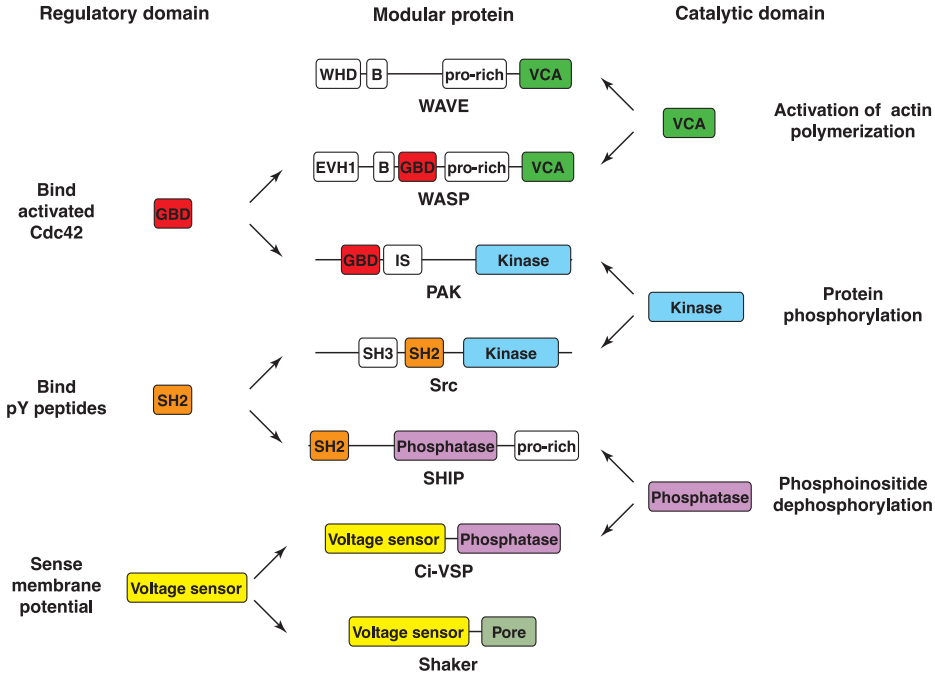

Figure 1.7: Examples of domain architecture in multidomain proteins. The VCA (verprolin homology, cofilin homology, acidic) domain is found in two actin-regulating proteins, WAVE and WASP. However, as it is combined with different other modular elements, which results in different activities of the two proteins. The GBD (GTPase binding domain) in WASP and PAK (p21-activated kinase) mediates binding to activated Cdc42. Kinases form one of the largest protein families in eukaryotes, and the kinase domain is found in many different proteins and domain architectures. The intramolecular interactions of the Src kinase are show in Figure 1.13 on page 34. The SH2 (Src homology 2) domain, which binds phosphorylated tyrosine (pY), is also found in a large number of proteins, including many kinases and phosphatases, where it is used for substrate recruitment and intramolecular inhibition. The phosphatase Ci-VSP from *Ciona intestinalis* contains a voltage-sensing domain, which regulates its activity dependent on the membrane potential. Shaker is involved in a voltage-gated channel, combining voltage-sensitivity with a transmembrane domain. The proline-rich regions (pro-rich) found in several of the proteins shown here are recognized by a number of domains, including GYF, SH3 and WW (see section 1.3.b)). Like the proteins illustrated here, many more fusions of these and other domains are found in Nature, enabling the re-use of known functionalities in a variety of combinations to generate new connections. Figure taken from Bhattacharyya et al. (2006).

1.3.b) Linear motifs

Linear motifs, also called Short Linear Motifs (SLiMs), are short stretches that occur in otherwise unrelated proteins and mediate a common function, e.g., as localization signals or as recognition motifs for specific domains. In addition to their shortness, many motifs are *degenerate*, i.e., only a few residues are critical for their recognition, with the remaining positions being variable or even arbitrary. For example, the core motif recognized by SH3 (Src homology 3) domains is PxxP, i.e., a proline (P) followed by two arbitrary residues (x) followed by another proline. Often, *regular expressions* are used to describe motifs, which allows more flexibility. For example, the Aurora kinase in yeast recognizes and phosphorylates `[KR].[ST][ILV]` motifs, where `[KR]` can be either lysin (K) or arginine (R), `[ST]` is either serine (S) or threonine (T), and `[ILV]` can be one of leucine (L), isoleucine (I) or valine (V). The dot (.), like the x in the SH3 motif above, stands for arbitrary residues. Thus, peptides RRTT, KNSL and RVSI all match the pattern. Additional symbols frequently used in regular expressions for linear motifs are listed in Table 1.2. Support for regular expressions is built into several programming languages, thus searching for motif occurrences in protein sequences is simple. An alternative representation for motifs are position-specific scoring matrices (PSSMs, e.g., Obenauer et al. (2003)). PSSMs contain more detailed descriptions, they hold a value for each residue in each position, related to how often it is observed. In motif detection, this is used to compute a score for the motif occurrence, while a regular expression either matches or does not match. However, more data is required to build a PSSM. In addition, the key residues as represented in a PSSMs are more difficult to grasp for humans working with them. The largest resource for such motifs is the Eukaryotic Linear Motif database (ELM, Puntervoll et al. (2003)), which collects motifs described in the literature along with their recognition domain and makes this data available to the scientific community. If possible, information about proteins or 3D structures containing the respective motifs is provided. Starting with around 80 motifs in 2003, ELM now covers almost 150 motifs (Gould et al., 2010), including cleavage sites, motifs recognized by globular domains as discussed below, and sites for post-translational modifications.

Domain-motif interactions → 1.4.c), page 32

Due to their short and degenerate nature, SLiM detection based on sequence alone is difficult (Neduva and Russell, 2005). One reason for this is that the

Symbol	Name	Meaning
.	dot	Any amino acid allowed
[...]	character class	Amino acids listed are allowed
[^ ...]	negated character class	Amino acids listed are not allowed
$\{\min, \max\}$	specified range	$\min$ residues required, $\max$ allowed
^	caret	Matches the N-terminus of the sequence
$	dollar	Matches the C-terminus of the sequence
?	question	One amino acid is allowed, but is optional
$*$	star	Any number of amino acids are allowed (including zero)
$+$	plus	One or more amino acids are allowed
|	alternation	Matches either expression it separates
(...)	parentheses	Used to either (1) mark positions of specific interest; e.g. the amino acid being covalently modified. or (2) group parts of the expression

Table 1.2: Brief description of regular expressions as used in ELM motifs, available at `http://elm.eu.org/help.html`

signal-to-noise ratio is poor in many datasets (Chica et al., 2009; Davey et al., 2009). Furthermore, motifs are often found in unstructured regions, which are difficult to align (Perrodou et al., 2008), making the identification of short over-represented elements even more difficult. Thus, it is necessary to first identify sets of proteins which are likely to share a motif, e.g., by virtue of common interaction partners, specific post-translational modifications (PTMs, see below) or sub-cellular localizations.

In order to bind their recognition domain or become modified, motifs need to be accessible. Thus, they are often found in IUP and unstructured regions (Fuxreiter et al., 2007), but less common in well-ordered structures such as coiled coils or globular domains (Neduva et al., 2005), although it has recently been shown that a number of domains do occur in loops on the surface of domains (Via et al., 2009). Nevertheless, many current tools for motif detection focus on unstructured regions and exclude well-structured regions, thereby focusing on the with higher motif density. The relatively high conservation of domains is another reason to exclude them, as their conservation would mask the – usually much weaker – signal for a linear motif. As an example, Neduva et al. (2005) search unstructured regions of proteins with a common interaction partner for linear motifs, and successfully identify a number of cases that

are described in ELM as well as new ones. However, they cannot always unequivocally determine the binding domain for these motifs. Like several other methods for motif discovery, Neduva et al. (2005) is based on TEIRESIAS, a generic program for pattern identification (Rigoutsos and Floratos, 1998). However, TEIRESIAS does not consider evolutionary relationships among the motif-containing sequences, thus requiring a lot of post-processing to identify possibly biologically relevant patterns. Furthermore, there is only limited support for ambiguity (such as [KR] or [ST] in the Aurora pattern above) or wildcards of variable length. Therefore, Edwards et al. (2007) developed a novel algorithm, SLiMFinder, that is specifically geared towards the detection of short, over-represented motifs in biological sequences. Given a set of sequences, SLiMFinder first generates possible motifs occurring in them, including flexible gaps and ambiguous positions. This step is based on unaligned sequences because reliable alignments of unstructured regions, in which peptides are often found, is difficult (Perrodou et al., 2008). Next, SLiMFinder ranks the motifs according to their probability of arising by chance. The tool was able to successfully rediscover a number of known ELM motifs and has a low false discovery rate. Additional features for motif discovery are available, for example, it is possible to specify residues that are required to be in the pattern, such as those found to be post-translationally modified in the peptides of interest.

1.3.c) Post-translational modifications

In several linear motifs, recognition is not only dependent on the amino acid sequence itself, but also on post-translational modifications (PTMs) of residues therein (Fig. 1.8 on page 25). In general, PTMs are chemical moieties covalently attached to an amino acid. They are often found in unstructured regions or IUPs, but also in loops on the surface of domains. Phosphorylation is probably the most commonly found and best-studied PTM, but a variety of others have been observed as well, including acetylation and methylation, which are frequently found in histones, where they affect chromatin organization and epigenetic regulation of gene expression (Seet et al., 2006; Ubersax and Ferrell, 2007; Sims and Reinberg, 2008). In general, PTMs may affect a protein's activity, interaction properties or localization. The process of adding such a PTM is relatively cheap in terms of energy requirements, especially

in comparison to production of proteins from mRNA or the transcription of RNA from DNA. Furthermore, it happens on a much shorter timescale and is thus suitable for rapid changes and fast processes such as signalling cascades. Lastly, PTMs are almost always reversible, so that it is not only possible to activate a given protein or pathway, but also to deactivate it once this particular function is not needed any more. In recent years, the availability of mass spectrometry (MS) for high-throughput studies of post-translationally modified proteins has allowed the detection of large numbers of phosphorylation sites. In particular, some methods have been developed specifically for the identification of phosphorylation sites in proteins that are normally expressed in low copy numbers in the cell, which had made their study difficult before (Stein et al., 2009a). Databases such as phospho.ELM (Diella et al., 2004) and PHOSIDA (Gnad et al., 2007) make sites identified in such studies as well as those curated from the literature available to the scientific community.

Domains that recognize post-translationally modified residues include 14-3-3 and FHA (forkhead-associated domain), which bind phosphorylated serine (S) and threonine (T), the bromodomain, which recognizes acetylated lysine, the phospho-tyrosine binding (PTB) and Src homology 2 (SH2) domains, which bind phosphorylated tyrosine (Y), and the ubiquitin-binding domain (UBD) which recognizes the addition of ubiquitin to lysine residues (Yaffe and Smerdon, 2004; Seet et al., 2006). These domains have been termed "readers" of PTMs (Pincus et al., 2008). Analogously, the enzymes adding the respective modifications can be considered "writers". In the case of phosphorylations, these enzymes are kinases, which form one of the largest families of genes in eukaryotes, covering roughly 2% of the proteome (Manning et al., 2002; Ubersax and Ferrell, 2007; Turk, 2008). Like the domains binding the modified motifs, kinases recognize specific patterns in proteins which they then phosphorylate (cf. above). In addition, kinases may recruit substrates by binding docking motifs to surfaces integral to the kinase domain (often found in S/T kinases) or to interaction domains such as SH3 or SH2 (often found in Y kinases). In both cases, the substrate recognition site may be distal from the site that is phosphorylated (see Fig. 1.9 on page 26). This increases the local concentration of substrate close to the active site, and also creates a dual requirement for substrate specificity, as both the docking motif and the sequence recognized at the active site need to be present in a putative substrate, with appropriate spacing to assume both interactions. It is possi-

ble that the docking motif must be modified or *primed* itself in order to be recognized, further increasing substrate specificity. Another possible effect is that binding to the docking site may allosterically influence the activity of the kinase (cf. 1.4.c) and Fig. 1.13). While the integration of docking receptor and kinase domain in S/T kinases is less flexible, a possible advantage of such a structure is that the resulting proteins are smaller and more compact (Bhattacharyya et al., 2006; Ubersax and Ferrell, 2007; Pawson and Kofler, 2009).

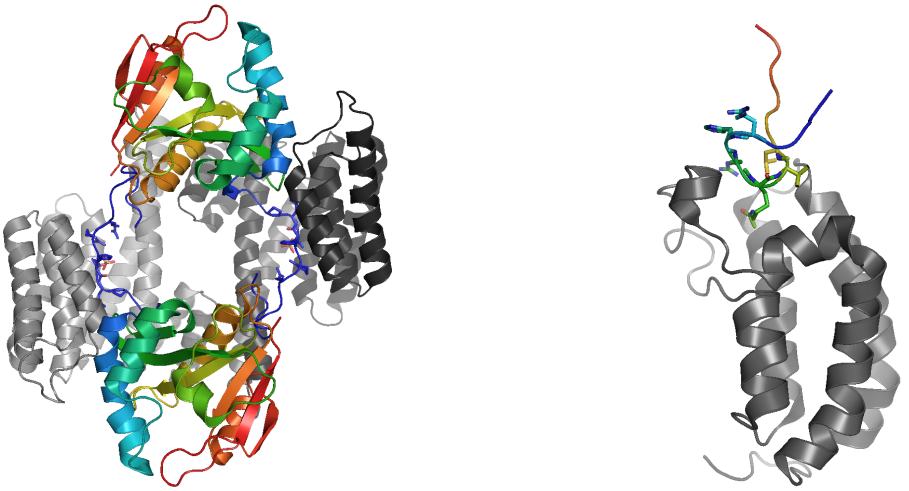

Figure 1.8: Recognition domains that bind post-translationally modified residues. Domains are shown in gray, motif-containing ligands in a rainbow color-scheme. (left: PDB:1ib1) The 14-3-3 domain binds phosphorylated serine or threonine residues (Yaffe and Smerdon, 2004). They usually form dimers, with each molecule binding a phosphorylated ligand. (right: PDB:2rny) The bromodomain binds acetylated lysine, which is frequently found in histones. Therefore the bromodomain is often part of proteins involved in chromatin regulation (Seet et al., 2006)

The third component of PTM-dependent signalling systems are "erasers", enzymes that remove the modification. In the case of phosphorylation networks, these are phosphatases. It appears that phosphatases by themselves are less specific than kinases. However, they often interact with regulatory subunits that affect their specificity. It has been observed that most S/T kinases have a preference for phosphorylating S, while many S/T phosphatases have a preference for dephosphorylating T. Despite the bulkier nature of Y in comparison

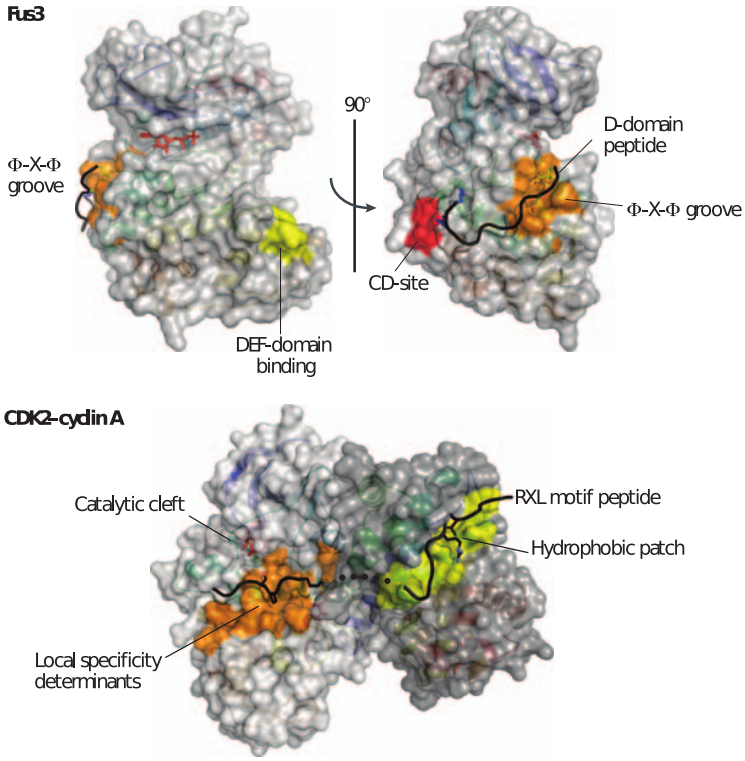

Figure 1.9: Docking sites on kinases have important roles in substrate recognition, both by increasing the local concentration in proximity to the active site and by selecting only substrates that have both the docking and the kinase motif. (top) The MAPK Fus3 in complex with a peptide from Far1, a cyclin-dependent kinase inhibitor, which is bound via a Φ-X-Φ motif, where Φ is a hydrophobic residue. A regular expression to describe this pattern might be `[AILVMF].[AILVMF]`, however details which hydrophobic residues are accepted and which are not depend on the individual binding site. ATP (red) is shown in the active site. (bottom) The complex formed by cyclin A (dark gray) and the cyclin-dependent kinase 2 (CDK2, dark gray) binds a substrate with an RXL motif. The motif-binding site (yellow) is located on cyclin A, while the active site and its surrounding specificity-determining residues (orange) are on CDK2. Exchange of the interaction partner can thus modify the substrate specificity of CDK2. Figure taken from Ubersax and Ferrell (2007).

to S/T, there are dual-specificity phosphatases that remove phosphates from either of the three residues, in addition to dedicated S/T and Y phosphatases (Ubersax and Ferrell, 2007).

The triangle of tyrosine kinases, phosphatases and SH2 domains raises an interesting question regarding the evolution of the pY-based signalling system, which is used extensively in cell-cell signalling in multicellular animals. Pincus et al. (2008) have applied comparative genomics studies to study the chicken-and-egg question of how such a system, based on three independent but functionally interrelated components, could have evolved. They compared metazoan genomes to that of the unicellular *Monosiga brevicollis*, which contains SH2 domains, as well as to genomes of other organisms that contain only parts of the Tyr signalling machinery, and found indications that SH2 domains and Tyr phosphatases appeared before Tyr kinases. Considering the fact that some promiscuous S/T kinases can also phosphorylate Y, it may have been a selection advantage to specifically respond to and remove those modifications. The additional complexity, due to new pTyr-based signalling connections that do not interfere with existing S/T-phosphorylation-based pathways, may have contributed to the emergence of multicellular organisms (Manning et al., 2008; Pincus et al., 2008; Mayer, 2008).

1.3.d) The role of modularity in protein interaction networks

An advantage of having modular elements not only in proteins, but also in protein interaction networks is the possibility of re-using the same module in different contexts. Different levels of modularity have been observed in protein interactions. For instance, in the case of docking sites, the binding motif can be transferred to other substrates, but the docking site itself cannot easily be inserted into other proteins. Independently folding domains in multidomain proteins, possibly combined with motifs in linker regions or on substrate proteins, are more modular as each element is functional in separation from the others and can thus be re-used in different contexts. The ultimate separation of catalytic activity and substrate recruitment is found in adaptors and scaffolds, proteins with the function of co-localizing enzyme and substrate(s). This may allow the same catalytic protein be used in different pathways leading to different outcomes, into which it is recruited by the respective scaffolds (Bhattacharyya et al., 2006). For example, the yeast MAP kinase (MAPK) kinase kinase (MAPKKK) Ste11 participates in three different signalling cascades: the mating pheromone response, the filamentous growth and the high osmolarity response pathway. To yield signalling

Adaptors are proteins that tether and co-localize two other proteins for interaction, *Scaffolds* do the same with three or more proteins.

MAPK: mitogen-activated protein kinase; S/T kinases that respond to various extracellular stimuli and form the terminal components of three-kinase cascades.

specificity and eliminate cross-talk, the scaffold proteins Ste5 and Pbs2 recruit Ste11, together with the respective pathway-specific MAPK (Fus3 or Hog1) into a complex. Pbs2, in addition to its scaffold function, also acts as a MAPK kinase (MAPKK) and phosphorylates Hog1 in the high osmolarity response pathway, while in the other two pathways this function is carried out by MAPKK Ste7. Phosphorylation of the MAPK eventually leads to transcription of the appropriate genes required in the three respective situation (Ubersax and Ferrell, 2007; Schwartz and Madhani, 2004).

Due to their independent nature, domains may be inserted on the surface of an existing protein without disrupting its structure. This allows for quick evolution of new functions as it combines two established, stable and functional modules – namely, the "host" protein and the domain – and thus should be less subject to viability constraints than the trial-and-error-process of single point mutation. Motifs, on the other hand, are usually not transferred by insertion. Instead, as they only require a few key residues for binding, they may arise from a few mutations in accessible and often unstructured regions. Along with the insertion of the corresponding motif-binding domain in either a protein acting catalytically on the motif-bearing one, or in a scaffold to tether both for interaction or modification, this allows the creation of a new interaction, and thus a new functional connection between two proteins, on a relatively short timescale in evolutionary terms.

1.4 A more structured view of protein interactions

As described above, 3D structures are paramount to our understanding of protein function. Structures of interacting proteins reveal the molecular details of the binding interfaces, which often provide important clues as to how binding is achieved, and how the interaction might behave in the presence of competitors, which may be host or pathogenic proteins, but also drug-like components that aim to interfere with it. Furthermore, the molecular details provided in 3D structures can help explain the question of specificity, i.e., why do proteins A and B interact but not A' and B, although A and A' are similar? Understanding specificity in turn is crucial for protein interface design (e.g., Mandell and Kortemme (2009)) or the search for drug-like components (Wells and McClendon, 2007).

By exploring all interactions of known 3D structure as stored in the PDB we could divide protein interactions involving the modular domains described above into two main categories, on the basis of their contact interfaces (Aloy and Russell, 2006): domain-domain and domain-peptide interactions. Domain-domain interactions involve the binding of two globular domains, creating a relatively large contact interface. In contrast, domain-peptide interactions have much smaller interfaces, formed between a globular domain and a linear motif. A third type of interaction found in 3D structures are complexes, which usually have extended interfaces and employ cooperative effects. The elements and interactions in complexes are not necessarily based on modular domains. All three types are illustrated and described in more detail below.

1.4.a) Complexes

Protein complexes form the large molecular machines in our cells. They may consist of dozens of proteins, sometimes in combination with nucleic acids, and fulfill catalytic and structural functions that are conserved throughout the tree of life, such as gene transcription (polymerases, e.g. Cramer et al. (2001)), RNA translation (the ribosome, e.g. Ban et al. (2000)) and degradation of proteins that are misfolded or not needed any more (proteasome, e.g. Groll et al. (1997)). Several Nobel prizes have been awarded for the solution of 3D structures of these complexes over the last years. Currently, even larger assemblies are being tackled, for example, structures of the nuclear pore complex have recently been solved by cryo-electron tomography (Beck et al., 2007), with further analysis of the individual components and their connections by integration of different kinds of biophysical data by (Alber et al., 2007).

The interfaces between the elements of a complex frequently extend over large areas, with deep clefts and interlocking elements (Fig. 1.10 on page 31). Cooperative effects are often involved in these interactions. The elements of complexes have evolved to fit and function together perfectly, shown by the high degree of conservation observed for many complexes. The elements of a complex usually remain bound together over their lifetime and may even be unstable on their own, which may be related to the cooperative effects in their interaction. The absence of individual components may also render a complex

non-functional. It has been suggested that cells exploit this to activate and inactivate complexes over the course of time, e.g., in different phases of the cell cycle, without needing to produce and degrade all the elements every time. Instead, most members of the complex are pre-produced, and just individual crucial elements are expressed in time for complex activation, and degraded for deactivation. This has been termed *just-in-time assembly* (de Lichtenberg et al., 2005). While the individual components of complexes are usually not modular in the sense of the modular domains discussed above, a modularity has been observed in the architecture of complexes. Here it refers to the re-use of different proteins in various complexes, either independently or in the form of modules, groups of two or more proteins that always appear together (Gavin et al., 2006).

1.4.b) Domain-domain interactions

One of the earliest solved 3D structures, hemoglobin, shows the interaction of four globular domains. Many later structures confirmed that domains are often responsible for mediating the association between two proteins. DDIs can occur inter- and intramolecular, with an average interface size of around $2,000\mathring{A}^2$ (Chakrabarti and Janin, 2002). They have been found in a variety of cellular processes, from interactions between kinases and regulatory subunits in the cell cycle to gene transcription (Jeffrey et al., 1995; Bowman et al., 2004). Due to their modular nature, domains are usually stable independently, so that two domains can associate to form a stable interface, but also dissociate after their function in interaction is fulfilled without impairing their individual functionality or stability. In comparison to the interaction surfaces in complexes, those in DDIs are more flat, and employ cooperative effects less frequently (Fig. 1.11 on page 32). Yet, changes in the conformation of a domain upon formation of some DDIs have been observed (Stein et al., 2010). Some domains have been observed to interact with a large number of different domains, often using different, though possibly overlapping, interfaces (Aloy and Russell, 2004; Kiel et al., 2008). This is illustrated for the Ras-like G domain fold in Fig. 1.12 (page 33). Our database of 3D interaction domains (3did) which is discussed in more detail below, contains a collection of known 3D structures of inter- and intrachain DDIs, including data on the interfaces formed between these domains (Stein et al. (2005, 2009b) and section 3.4,

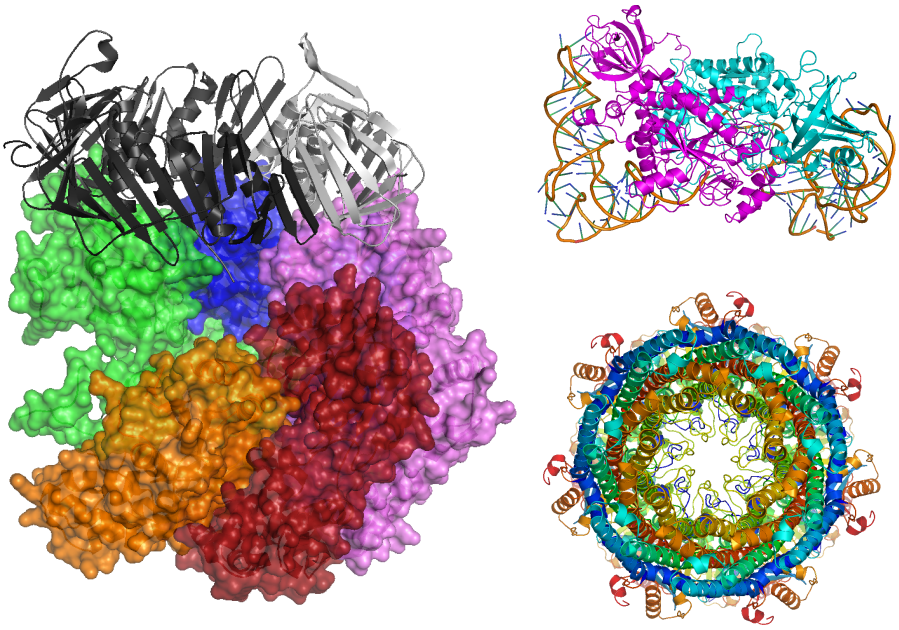

Figure 1.10: 3D structures of protein complexes. (left: PDB:1sxj) The clamp loader or replication factor C (RFC) complex (surface representation, colored), bound to the proliferating cell nuclear antigen (PCNA, cartoon representation, gray). The interfaces between the components of the RFC complex are much more extended than those between the three subunits of the PCNA (cf. Fig. 1.11) (Bowman et al., 2004). (right, top: PDB:1asy) The aminoacyl tRNA synthethase complex illustrates an interaction involving both proteins and nucleic acids (Ruff et al., 1991). (right, bottom: PDB:3ipm) The proteasome consists of 28 subunits in a highly symmetrical assembly (Groll et al., 1997). Its function is the degradation of ubiquitin-tagged proteins, either because of folding problems or because the protein is not needed in the cell any more.

page 147).

Inspired by the detection of domain-domain interactions in structures, Margalit and colleagues have analyzed PPI data in order to suggest DDIs that mediate the functional associations between the two proteins in question (Sprinzak and Margalit, 2001). They exploit DDIs of known structure, but also pairs of domains frequently found in interacting proteins, as structural data only available for about a third of all domains (cf. 1.3.a)). More recently, proteins have been dissected into domains in high-throughput interaction screens to

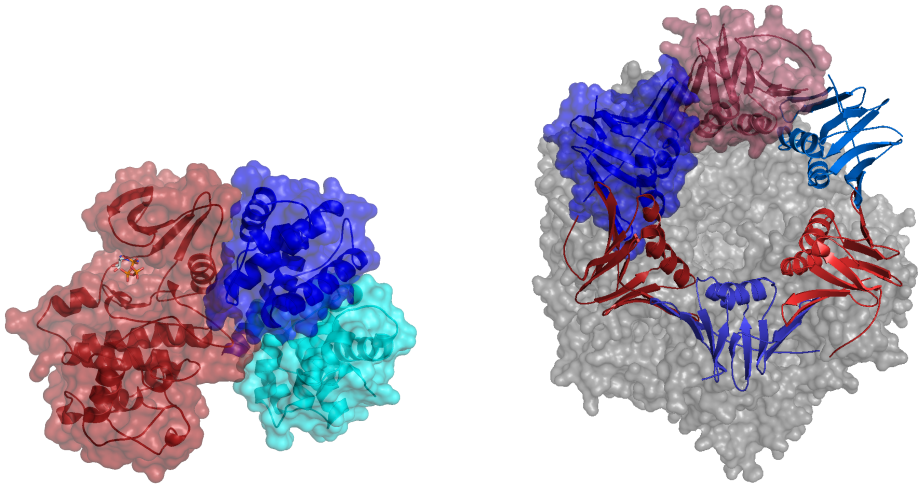

Figure 1.11: Interfaces in domain-domain interactions (DDIs). In comparison to the interfaces in complexes shown in Fig. 1.10, the domain-domain interfaces are smaller and flatter. Nevertheless they are large enough for stable assemblies. (left: PDB:1fin) The interaction between CDK2 (red) and Cyclin A (blue/cyan) is mediated by a DDI between the kinase domain in CDK2 and the N-terminal cyclin domain (blue) (Jeffrey et al., 1995). (right: PDB:1sxj) PCNA consists of 3 subunits, each consisting of an N-terminal and a C-terminal domain (blue and red, respectively). They associate via DDIs to form a ring that slides along the DNA during transcription (Bowman et al., 2004).

find out which domain pairs are responsible for the association (Boxem et al., 2008), which is possible, again, because of the independently folding nature of domains.

1.4.c) Peptide-mediated or domain-motif interactions

Domains do not only bind other domains, but also peptides containing linear motifs as those discussed above (Fig. 1.13). These domain-motif interactions (DMIs) have a much smaller interface, 350$\mathring{A}^2$ on average, which makes them well-suited for transient interactions such as those in signalling cascades. Due to their transient nature, DMIs are estimated to be under-represented in current protein interaction networks (Pawson and Linding, 2005). Despite the low number of contacts, the motif is essential to establish binding; in fact,

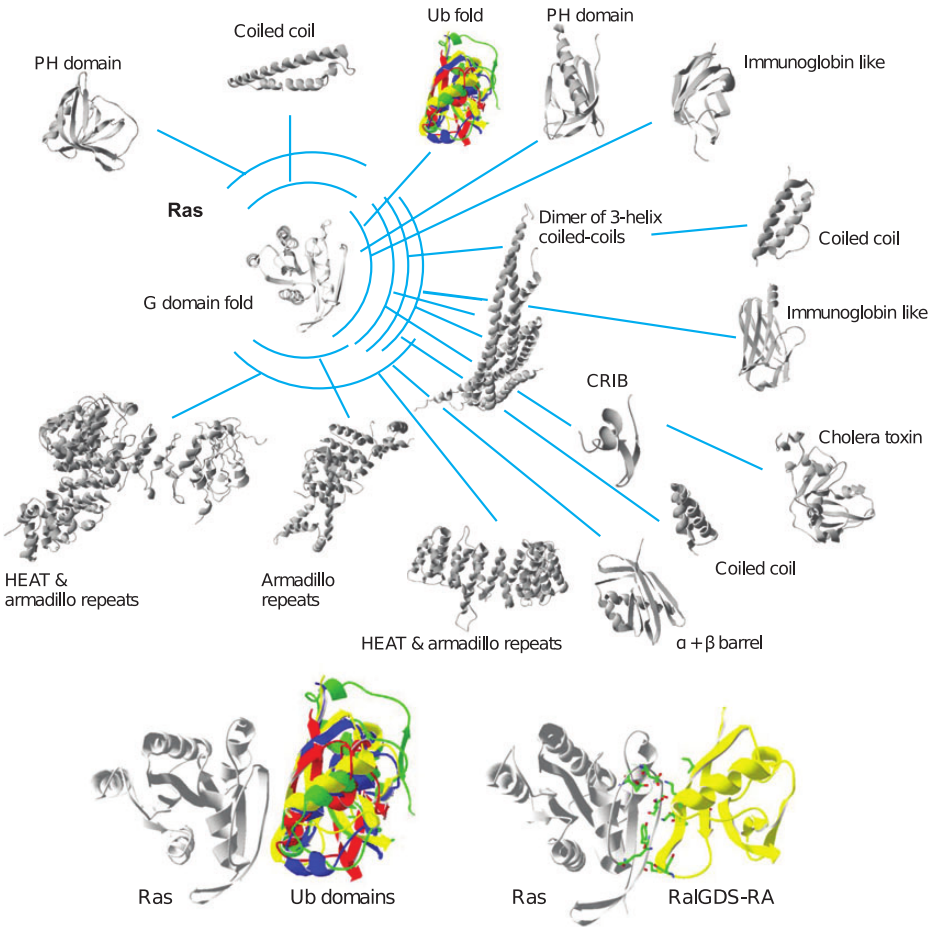

Figure 1.12: Different DDI interfaces for the G domain fold. Figure taken from Kiel et al. (2008). (top) The Ras domain forms a G domain fold. It binds a large number of effectors using different interfaces. (bottom, left) Four structurally similar Ras effectors (red, blue, green, yellow) from the family of ubiquitin-like (Ub) domains in interaction with Ras are superimposed, showing that they use the same binding interface. (bottom, right) The molecular contacts between the two domains determine both specificity and affinity of a DDI. Shown here are the contacts between Ras and RalGDS.

it has been shown that a short peptide with the motif alone is sufficient for binding (e.g., Kim et al. (2001)). Nevertheless, the residues in the flanking regions or *context* contribute to the interaction as well (Stein and Aloy, 2008;

Chica et al., 2009). We have analyzed the contribution of these regions based on 3D structures of DMIs and shown that the context contributes about 20% of the binding energy on average (see section 3.1, page 47).

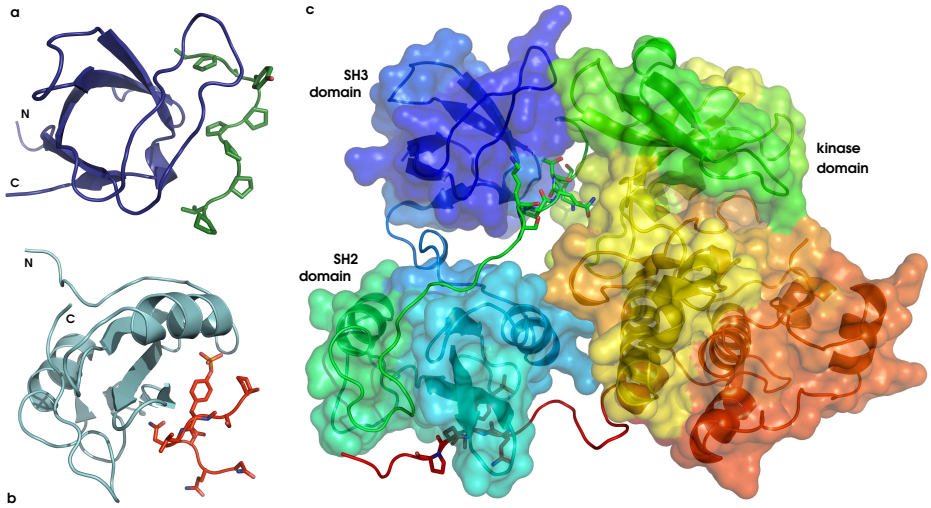

Figure 1.13: Domain-motif interactions in 3D structures. (a) The SH3 domain bound to a proline-rich peptide, its canonical recognition motif (Li, 2005). (b) The SH2 domain in interaction with a pY-containing peptide (Pawson et al., 2002). (c) Intramolecular interactions between the SH2 and SH3 domains and peptides in linker regions keep the Src kinase in a repressed conformation, unless external ligands for both domains are present. Note that the SH3-binding linker region (green) adopts a PPII-helix shape even though it is not proline-rich. PDB:1fmk (Xu et al., 1997). Figure shown in Stein et al. (2009a)

Different methods have been developed to study the binding preferences of individual peptide-binding domains. Alanine scanning, the iterative replacement of each residue in a peptide by alanine, can identify residues which are crucial for the interaction. In addition, replacement by residues other than alanine can reveal whether alternative peptides would bind with higher affinity, or whether some residues abolish binding, such as a proline C-terminal to the phosphorylation site of Aurora A (Ferrari et al., 2005). In phage display, peptides are fused to phage proteins such that they are displayed on the phages' surface, allowing easy testing of whether a given domain will bind the peptide *in vitro* (Smith and Petrenko, 1997). It is applicable on large datasets and has been used to identify the binding preferences of a large number of domains, including proteome-wide scans of domain families such as PDZ, SH3 and WW

(Tong et al., 2002; Landgraf et al., 2004). The PDZ (postsynaptic density 95; discs large; zonula occludens 1) domain is a typical peptide-binding domain, which specifically recognizes C-termini through β-strand addition (Remaut and Waksman, 2006), although exceptions have been observed in which the bound structure mimics a C-terminus (Hillier et al., 1999). PDZ domains are frequently found in fly and human, with 81 and 214 instances, respectively. In yeast, though, there are only two occurrences, which have low sequence identity to canonical PDZ domains such that it is not clear if these putative domains are functional (Harris and Lim, 2001; Stein et al., 2009a). Canonical PDZ motifs are based on their preferential recognition of peptides based on position -2, with 0 being the C-terminal residue (Nourry et al., 2003), although recent studies indicate that this classification may need revision (see below).

Motifs bound in DMIs are frequently observed in unstructured regions (Russell and Gibson, 2008), which is exploited by several methods for the identification of DMIs (Neduva et al., 2005; Edwards et al., 2007). Their location in unstructured regions makes them amenable for post-translational modifications (PTMs, cf. 1.3.c)), on which a number of DMIs depend. In those cases, binding domains only recognize the modified peptide, thus making the interaction dependent on dynamic events in the cell. The often large and/or charged modifications are critical for recognition by the domain, and often those motifs are based on even fewer residues than modification-independent DMIs. This dependence on PTMs is another factor that impedes the detection of DMIs in high-throughput methods, as these modifications may not be present under the method's conditions. A few exceptions to the requirement for PTMs have been found, though, e.g. SH2 or PTB domains that bind unmodified peptides. In these cases, the interaction interface is extended in comparison to that of the PTM-dependent DMI, and the affinity is lower (Pawson et al., 2002). Similarly, unique binding profiles have been observed for other peptide-binding domains, such as an SH3 domain that recognizes proline-free motifs (Tonikian et al., 2009).

Structurally, like DDIs, DMIs are found in interactions between two different proteins (interchain) as well as in intramolecular interactions. For example, the Src kinase contains an SH2 and an SH3 domain, which hold it in a repressed conformation by intramolecular contacts (Fig. 1.13 on page 34). SH2 binds phosphorylated Y (pY), while SH3 generally binds proline-rich regions

that adopt a specific structure, the polyproline-II-(PPII)-helix (Li, 2005). In this particular case, the linker region between the SH2 domain and the kinase domain, despite not being proline-rich, adopts a PPII-helix conformation and binds to the SH3 domain, which helps to keep the kinase repressed (Xu et al., 1997; Pawson, 1997). When the C-terminus of Src is phosphorylated, it is bound by the SH2 domain, complementing the repressive action of the SH3:peptide interaction (Fig. 1.13). SH2 domains are almost exclusively found in metazoan organisms (cf. 1.3.c)). Studies in the non-metazoan *Monosiga brevicollis* have shown that phosphorylation of the C-terminal Y residue and its interaction with the SH2 domain do not inhibit kinase activity in the organism's Src homolog. Thus, it is assumed that the peptide-binding domains initially served substrate localization and only subsequently evolved the ability to repress the kinase in absence of suitable substrates (Li et al., 2008; Pawson and Kofler, 2009).

Despite the short and degenerate nature of motifs, DMIs are highly specific *in vivo* (Zarrinpar et al., 2003; Yun et al., 2009). Data from 3D structures provides molecular details of the interfaces between domain and motif and thus critical information on how this specificity is achieved. However, 3D structures as well as *in vitro* experiments like alanine scanning of phage display capture the molecular binding preferences of a given domain, but are independent of whether the domain and a putative binding peptide actually occur at the same time, in the same sub-cellular localization *in vivo*. As explained below, there are a number of factors influencing the specificity in PPIs *in vivo*, which help achieve the high specificity observed in DMIs (section 1.5, page 38).

1.4.d) Availability of structural data and modelling

Given the demanding nature of structural studies, it is likely not feasible to cover entire interactomes with high-resolution 3D data of the interfaces (Aloy and Russell, 2005). However, computational methods should be able to bridge this gap when given structural data of sufficient variety from which to learn how PPI interfaces "work". Similar to homology modelling, it is possible to model the interaction of two proteins on a homologous interacting pair (Aloy and Russell, 2002a). An increased coverage of distinct domain-domain (Fig. 1.14 on page 37) and domain-peptide interfaces will improve

the reliability of such methods. In the case of DDIs, automatic detection is relatively easy, as domains can be identified from sequence with high reliability. For peptide-mediated interactions, the difficulty lies in the detection of true motifs – the regular expressions often used for motif encoding will yield many false positives, due to their shortness and degenerate nature. More complex models like PSSMs, when available, would require a cut-off value to filter likely true hits from likely false ones. In either method, only measures such as proximity to the recognition domain and location of the peptide in the appropriate binding groove would be able to confirm a true instance of the motif. In this work, we have identified a large number of DMIs in 3D structures in a semi-automated fashion. This could provide a basis for a more automated yet reliable detection of DMIs in 3D structures.

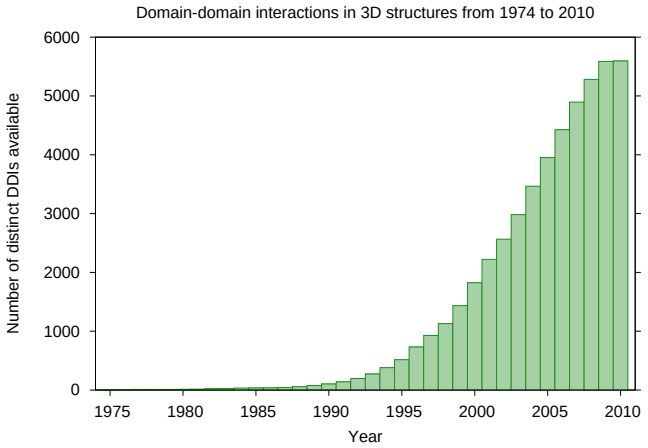

Figure 1.14: Domain-domain interactions (DDIs) for which 3D structures are available. The number of distinct DDIs for which 3D structural templates are available has increased steadily over the last 15-20 years. Analogous to an increase in fold coverage improving homology modelling, the growing availability of 3D structures of DDIs allows the application of methods such as InterPreTS, which predict protein interactions based on 3D structures of homologous domain pair, to ever larger sets of proteins (Aloy and Russell, 2002a, 2003).

1.5 Specificity and contextual information

The interaction classes described above exploit the modular nature of domains and motifs to generate different combinations of functionalities and connections. At the same time, though, they are often involved in highly specific events such as signalling pathways. Thus, modularity does not imply that every protein containing domain A binds every protein containing domain B, or that every SH3 domain binds every proline-rich peptide. A range of specificities has been observed; some domains are rather promiscuous, while others are highly specific. In particular, domain-motif interactions (DMIs) are known to be very specific *in vivo*. This may be surprising, given the small number of residues involved in these binding events. However, the direct molecular contacts between motif and domain are only one aspect of the specificity in such an interaction. Several other factors also play a role in determining specificity. For instance, an important issue is whether two proteins – even though they may bind when brought together *in vitro* – actually meet in the cell, or if they are expressed at different times or in different sub-cellular localizations. We refer to such factors as *context*. Several important aspects of the context are discussed below and illustrated in Fig. 1.15 (page 39).

Zarrinpar et al. (2003) have studied the case of the PxxP-motif in Pbs2 in yeast, which only binds to the SH3 domain in Sho1, but none of the other yeast SH3 domains. By introducing single point mutations in the motif but apart from the key residues, they observed that even small changes would alter the binding profile. Thus they concluded that specificity in these interactions evolves not only to bind the functional interaction partner well, but also to avoid binding to other proteins or cross-talk – in other words, to maximize specificity (see also option c in Fig. 1.15). In addition, the authors found that the PxxP motif in Pbs2 binds to SH3 domains from organisms other than yeast, indicating that this fine-tuned evolution of the specificity profile can only consider competitors that it actually encounters *in vivo*. More recently, Tonikian et al. (2009) have performed a large-scale study on SH3 domains and their binding peptides in yeast (Tonikian et al., 2009). They scanned billions of peptides in phage display experiments for binding to any of the SH3 domains, and found that many SH3 domains have a binding profile that corresponds to one of the canonical patterns, `[RKY]..P..P` and `P..P.[KR]`. However, they also reported several SH3s with clearly divergent peptide recognition profiles,

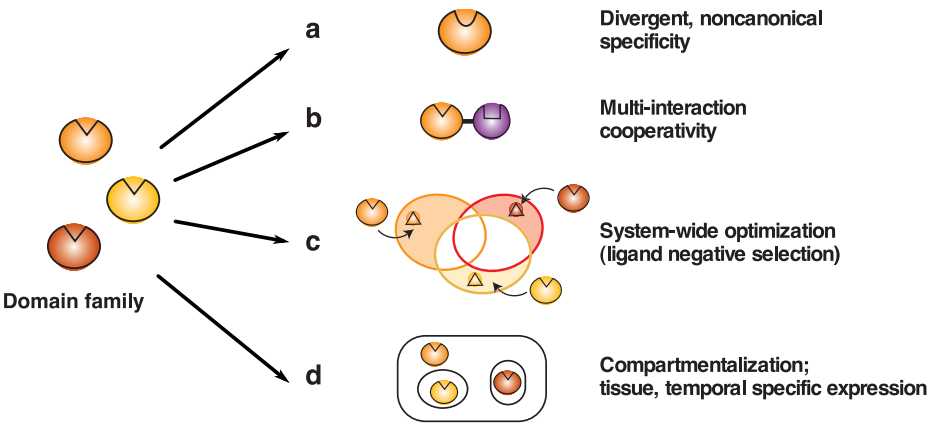

Figure 1.15: Contextual information helps determine specificity. Several factors determining the *in vivo* specificity of recognition domains are illustrated here. Usually, these factors are orthogonal, i.e., they are independent of each other and may be combined to further increase specificity. (a) The evolution of a non-canonical binding interface. In this case, the molecular binding surface recognizes only a small subset of the peptides generally bound by this domain. (b) Adaptors and scaffolds (cf. 1.3.d)) require multiple binding constraints be fulfilled. They may include distinct domains, but also several copies of the same, so that repeated, appropriately spaced motifs are needed. (c) Fine-tuning of the peptides and binding domains into niches of specificity, so that cross-reactivity with other members of the binding domain family is avoided. (d) Domains and peptides that do not occur in the same cell and sub-cellular localization at the same time are not at risk of cross-talk. This is reflected in the observation that binding to non-natural competitors, such as domains from a different organism, occurs even with fine-tuned interactions (Zarrinpar et al., 2003). Figure taken from Bhattacharyya et al. (2006).

including one domain that recognizes peptides without proline residues. Using the peptide binding profiles generated in this study together with protein interaction data, Tonikian et al. (2009) trained a Bayesian algorithm with which they identified novel interactions in the yeast interactome that are mediated by SH3-peptide binding. Their study confirms that peptide-binding domains evolve into niches of specificity. Similar studies have been carried out for the PDZ domain, where an even broader variety of specificities has been observed (Chen et al., 2008; Tonikian et al., 2008). ELM currently lists several discrete motifs for this domain, `.[ST].[VIL]$`, `.[VYF].[VIL]$` and `.[DE].[IVL]` (as described above, PDZ domains usually bind C-terminal peptides, which is indicated by the $ in the pattern). These consensus motifs

are based on the observation of preferentially recognized peptides by different members of the PDZ family, differing mainly in the -2 position, with 0 being the C-terminal residue (Nourry et al., 2003). However, in a recent study of all putative C-terminal PDZ-ligands across the mouse proteome, Stiffler et al. (2007) have shown that the 157 mouse PDZ domains analyzed in their work do not fall into distinct binding classes. Instead, their specificity is evenly distributed throughout selectivity space. Based on this rich dataset, the group created multidimensional profiles to represent the binding preferences of each of those PDZ domains. To achieve this, they built a model of 38 contacting residue pairs in PDZ domains from 3D structures and considered sequences from both peptide and domain Chen et al. (2008). A major advantage in training their system was that, in addition to large sets of quantitative binding data, they also knew which pairs of domains and peptides did not interact.[1] These models are capable of successfully predicting PDZ-motif interactions in mouse given sequence information only. They have also been tested on fly and worm PDZ domains and found – perhaps unsurprisingly – to work with lower accuracy in those organisms. Tonikian et al. (2008) studied PDZ-binding peptides with a phage-display-based approach in the human and worm proteomes. They identified at least 16 specificity classes, which significantly extend the classifications based on position -2, yet the variety is not as broad as in the mouse study by MacBeath and colleagues described above. Tonikian et al. (2008) also performed mutational studies on PDZ domains and showed that, while the recognition of C-termini is robust, some mutations can alter the specificity of the domain even when the changed position is not in direct contact with the peptide.

These three recent studies on recognition domain specificity suggest that many other discrete patterns may need revision, too, in particular for application of motifs in predictive methods. In addition, they indicate that regular expressions may not be the most suitable form to capture the binding preferences of individual domains, but that more complex models will be required. Nevertheless, the more generic motifs do provide useful information on the preferences of recognition domain (sub)families, not least in cases where large-

[1] The lack of negative interaction data is a major issue in the development of reliable prediction methods, see e.g. Aloy and Russell (2006). Recently published sets of proteins unlikely to interact (e.g., Smialowski et al. (2010)) may prove useful in addressing this problem.

scale screening data of the binding profile of the specific domain and organism of interest are not available.

High-resolution 3D structures provide crucial information on the interaction interface and thus specificity-determining regions, which makes them great choices for studying protein interactions in detail (e.g., Yun et al. (2009)) as well as for the development of predictive methods. For domain-domain interactions (DDIs), Aloy and Russell (2002a) developed a method to predict whether two proteins will interact, based on structural templates of homologous proteins. However, while templates are available for thousands of DDIs (Fig. 1.14, page 37), until recently it was assumed that there were only a handful of 3D structures of interactions between domain and linear motifs (DMIs) (Neduva and Russell, 2006). Therefore, approaches seeking to exploit the high-resolution data from 3D structures for peptide-mediated interactions were often based on a small set of structures which was combined with other data, e.g. from phage display experiments. An example is the method to predict PDZ-peptide interactions by Chen et al. (2008) described above. For a few other domains, including SH3 and WW, the SPOT (Specificity Prediction Of Target) algorithm has been applied, which uses phage display data in combination with information from 3D structures of the DMI to predict binding (Brannetti et al., 2000; Brannetti and Helmer-Citterich, 2003). However, sufficiently diverse structural as well as peptide-binding data is required to make these methods applicable to all families of peptide-binding domains. In general, such methods align the sequences of their templates to those of the sequences queried, so that the molecular contacts affected by the difference between the respective domains and peptides can be mapped. Evaluating the effect of these changes on binding – e.g., based on physicochemical or statistical data on which contacts are favorable and which are not – leads to a score on whether the tested protein pair is likely to interact or not. Given sufficient data, it may be possible to fine-tuned such a system so that it ranks protein pairs by their "interaction strength", a measure similar to affinity. A, different structure-based approach is protein-protein *docking* which is based on individual structures and thus does not need a template of the interaction. Docking programs attempt to identify the interface, but are currently not capable of predicting whether the two proteins in question will actually bind or not (Aloy and Russell, 2006). A more recent development in the field of DMIs is a method that identifies binding sites for a given motif on

Docking: a method to predict the binding interface between two molecules, often using shape complementarity or simulations.

the surface of a selected globular domain (Petsalaki et al., 2009). An advantage of this method is that structures of globular domains are available in relatively large numbers, compared to interaction interfaces and particularly interfaces of domain-peptide interactions. However, interfaces predicted by either method require verification by e.g. mutational studies or structure determination.

There are other factors besides individual recognition surfaces that determine the *in vivo* specificity of an interaction, though (Fig. 1.15 b-d). This includes cellular contextual information like sub-cellular localization or expression patterns, which determine whether two potentially interacting proteins actually meet *in vivo*, but also combinatorial effects from having multiple recognition surfaces that bind a single interaction partner. For example, multidomain proteins often have different specificity-determining regions, both on their protein-binding domains and on their catalytic domains. Their concerted effect allows to bind and modify substrates with a higher specificity than each domain would permit individually. Similarly, adaptors and scaffolds combine multiple modular protein-recognition elements to recruit two or more proteins into a complex, thus ensuring that the appropriate substrates are brought close to their modifying enzymes. In general, the use of several different, often orthogonal measures to ensure specificity, from molecular recognition surfaces to cellular contextual factors appears to be frequently used in Nature (Stein et al., 2009a). With high-throughput methods, various kinds of contextual data are becoming available at a high speed. For example, the interaction detection methods described above are providing drafts of the interactomes of several model species, while high-throughput mass-spectrometry (MS) analyses reveal which proteins and positions are post-translationally modified. The question of compartmentalization has been addressed by O'Shea and colleagues, who identified the sub-cellular localization of 75% of the yeast proteome using GFP (green fluorescent protein)-tagged proteins (Huh et al., 2003). For other species, localization data is more sparse, though some is available in the Gene Ontology resource (Ashburner et al., 2000). In predictive methods, one can considerably improve the reliability by combining different types of contextual information. For example, Linding et al. (2007) have integrated data on protein interaction, *in vivo* phosphorylation, co-expression and other factors to identify the kinases responsible for experimentally determined phosphorylation sites. They report that 60-80% of the specificity in their phosphorylation

Adapters and Scaffolds → 1.3.d), page 27

network is contributed by contextual factors.

In this thesis, we will show that there are many more structural instances of peptide-mediate interactions in 3D structures than previously estimated. In a first step, we identify hundreds of DMIs in structures based on ELM patterns (Stein and Aloy (2008), see section 3.1, page 47). Using this atomic-resolution data, we study specificity and cross-talk, and the role of the flanking regions around the motif – the molecular context – in establishing specificity. Subsequently we derive parameters to describe linear motifs structurally, and use them to identify even more instances of DMIs among the set of 3D structures, involving both domains known to bind motifs and others for which this had not been described before (Stein and Aloy (2010) and section 3.2, page 65). Access to a large set of 3D structures of DMIs, as provided with the work presented in this thesis, will allow for the extraction of valuable biological data, such as the parameters governing the relationship between a peptide-binding domain and the corresponding consensus motif. Such data may in turn be used for scanning protein sequences of occurrence of the motif, to predict interaction partners based on the binding domain, or to offer molecular explanations for interactions discovered in studies with lower resolution, such as high-throughput interaction detection experiments. In another project presented as part of this thesis, we predict substrates of the human Aurora A kinase based on the occurrence of the consensus motif, and combine this with different kinds of contextual data (Sardon et al. (2010), see section 3.3, page 107). This is necessary because the phosphorylation pattern for Aurora is rather degenerate, and molecular details of the substrate recognition interface are not available, so that we could not use such data for substrate selection. However, an important consideration is the fact that Aurora acts at spindles, restricting the set of putative substrates from the entire human proteome to a few hundred spindle proteins. By integrating this with MS phosphorylation data and site conservation, we generate a ranked list of putative substrates, several of which are confirmed experimentally.

2 List of publications

Only articles 1, 3, 5, 8 and 9 will be discussed in this thesis.
Authors marked with a * contributed equally to the respective work.

1. Stein, A., Russell, R. B., and Aloy, P. (2005).
   3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res*, 33(Database issue):D413–D417.

2. Stein, A. and Aloy, P. (2008).
   A molecular interpretation of genetic interactions in yeast. *FEBS Lett*, 582(8):1245–1250.

3. Stein, A. and Aloy, P. (2008).
   Contextual specificity in peptide-mediated protein interactions. *PLoS ONE*, 3(7):e2524.

4. Parthasarathi, L., Casey, F., Stein, A., Aloy, P., and Shields, D. C. (2008).
   Approved drug mimics of short peptide ligands from protein interaction motifs. *J Chem Inf Model*, 48(10):1943–1948.

5. Stein, A., Panjkovich, A., and Aloy, P. (2008).
   3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res*, 37(Database issue):D300–D304.

6. Stein, A.*, Pache, R. A.*, Bernadó, P., Pons, M., and Aloy, P. (2009).
   Dynamic interactions of proteins in complex networks: a more structured view. *FEBS J*, 276(19):5390–5405.

7. Littler, D. R., Alvarez, M., Stein, A., Hibbert, R. G., Heidebrecht, T., Aloy, P., Medema, R. H., and Perrakis A. (2010).

Structure of the FoxM1 DNA-recognition domain bound to a promoter sequence. Accepted for publication in Nucleic Acids Research.

8. Stein, A. and Aloy, P. (2010).
Novel peptide-mediated interactions derived from high-resolution 3D structures. Submitted.

9. Sardon, T.*, Pache, R. A.*, Stein, A.*, Molina, H., Vernos, I., and Aloy, P. (2010).
Uncovering novel substrates for the Aurora A kinase. Submitted.

10. Stein, A.*, Rueda, M.*, Panjkovich, A., Orozco, M., and Aloy, P. (2010).
A systematic study of the energetics involved in structural changes upon association and connectivity in protein-protein interaction networks. In preparation.

3 Results

3.1 Contextual specificity in peptide-mediated protein interactions

Proteins are key players in virtually all biological events that take place within and between cells. They seldom act in isolation and often accomplish their function as part of large molecular machines, whose action is co-ordinated through intricate regulatory networks of transient protein-protein interactions. Consequently, much effort has been devoted to unveiling protein interrelationships in a high-throughput manner (Rual et al., 2005; Stelzl et al., 2005). However, such experiments only tell whether two proteins interact, but do not reveal atomic details of the molecular mechanism. This information is currently only available in the form of high-resolution 3-dimensional (3D) structures. By exploring all interactions of known 3D structure as stored in the Protein Data Bank (PDB) (Berman et al., 2000) we could divide protein interactions into two main categories: domain-domain and domain-peptide interactions (Aloy and Russell, 2006). Domain-domain interactions involve the binding of two globular domains from different proteins, thereby creating a large contact interface of $2.000\mathring{A}^2$ on average. In domain-peptide interactions a globular domain recognises a short *linear motif*, creating a relatively small interface. Such interactions are found predominantly in signalling and regulatory networks (Pawson and Nash, 2003). Linear motifs are short patterns of around 10 residues with a common function (i.e. binding to a globular domain) that occur in otherwise unrelated proteins. In isolation these motifs bind their target proteins with sufficient strength to establish a functional interaction. They are frequently found in disordered or unstructured regions, and adopt a well-defined structure only upon binding. Usually just a few residues in the motif are fixed to a specific amino acid, or restricted to a small set of residues, while several positions may be arbitrary. For exam-

ple, SH3 domains bind PxxP patterns, where 'x' denotes an arbitrary amino acid. Due to their transient nature, peptide-mediated interactions are difficult to handle biochemically, and thus underrepresented in high-throughput experiments (Pawson and Linding, 2005). The largest collection of manually curated information about linear motifs is provided by the Eukaryotic Linear Motif Database (ELM) (Puntervoll et al., 2003). Despite their shortness, peptide-mediated interactions are extremely specific *in vivo*. Several studies have pointed out that contextual information, such as localization and expression patterns, but also the residues surrounding the motif, is important in determining specificity (Zarrinpar et al., 2003; Stiffler et al., 2007). To study this class of interactions and especially the role of the context in detail, we first identified all instances of motifs as defined in ELM in interaction with the corresponding domain in the PDB by a combination of motif-based search and visual inspection (Stein and Aloy, 2008). We found 810 interactions, involving 47 different motifs and 30 recognition domains, which form a set of 383 non-redundant peptide-mediated interactions. In order to analyse the contribution of motif and context, which we here define as all residues outside the defined ELM motif, we used FoldX (Guerois et al., 2002) to compute the binding energy of the native interaction, and again after the mutation of all motif and context residues, respectively, to alanine. We found that the context contributes, on average, 21% of the binding energy; thus, as expected, the motif provides the majority. We also studied and quantified the topological and energetic variability of interaction interfaces within each interaction type (combination of recognition domain and peptide class), finding a much higher heterogeneity in the context residues than in the consensus binding motifs. Finally, we performed a peptide exchange among peptides binding the same domain to study specificity and assess the potential for cross-talk. Using *in silico* alanine scanning (Schymkowitz et al., 2005) we discovered that the contribution of motif residues is favorable in both native and constructed (non-native) peptide-domain interactions, but that many non-favorable contacts occur in the context, in native but especially in constructed interactions. These findings confirm that the context is crucial in determining interaction specificity.

Stein A, Aloy P. Contextual specificity in peptide-mediated protein interactions. PLoS One. 2008; 3(7): e2524.

3.2 Novel peptide-mediated interactions derived from high-resolution 3D structures

Many biological responses to intra- and extracellular stimuli are regulated through complex networks of transient protein interactions in which a globular domain in one protein recognizes a linear peptide in another. These peptide stretches are often found in unstructured regions of proteins, and contain a consensus motif complementary to the interaction surface displayed by their binding partners. Many currently available methods for the *de novo* discovery of linear motifs exploit the fact that these usually occur in unstructured regions, in combination with other characteristics such as a common interaction partner (Neduva et al., 2005) or other shared biological features, such as co-localization (Edwards et al., 2007). In this work, we focus on another aspect of domain-motif interactions (DMIs), namely that the peptides, when bound to their recognition domain, adopt a specific, well-defined structure. While analyzing the peptide-mediated interactions in high-resolution 3D structures we had identified previously (Stein and Aloy, 2008), we noticed that many bound peptides have a particular stretched and elongated shape, which can be observed across different families of peptides. Comparison with other instances of peptide-mediated interactions shows that this particular structure can be as characteristic for a motif as the peptide itself. Here we analyzed these structural properties, and used the results to train a Support Vector Machine (SVM) to recognize linear-motif-like peptides. In combination with other structure-encoded features, such as molecular contacts and interface size, we identified around 10,000 putative DMIs. To establish the biologically significant motifs among them, we clustered the data by sequence and topology and derived consensus motifs for all topological clusters, or *interaction types*, with sufficient non-redundant sequential information. We found significant patterns for around 150 interaction types, involving almost 100 different domains. In a benchmark procedure, we showed that our method successfully rediscovers $2/3$ of the DMIs with high-resolution 3D structures derived from ELM patterns, as described previously (Stein and Aloy, 2008). Finally, to cross-validate our identification strategy, we scanned interactome networks from four model organisms (human, fly, worm, and yeast) with our newly derived patterns to see if any of them occurred more often than expected. Indeed, we found significant over-representations for 64 domain-motif

interactions, 46 of which had not been described before, involving over 6.000 interactions in the model species interactomes for which we could suggest the molecular details determining the binding. Knowledge of the atomic details of an interaction interface, which is provided along with the consensus motifs identified in this work, is critical for the planning of further functional studies as well as for the development of interfering elements, be it drug-like compounds (Neduva and Russell, 2006; Russell and Gibson, 2008), novel engineered binding proteins (Taussig et al., 2007) or peptides in synthetic circuits or networks (Reményi et al., 2006; Mandell and Kortemme, 2009). The transient interactions involving regulatory proteins and small interfaces typical for DMIs make them interesting candidates for all these applications. In addition, this extended set of structures for peptide-mediated interactions will allow for the extraction of valuable biological and biophysical data, such as the relationship between peptide-binding domains and the motifs they recognize. Moreover, the high-resolution 3D data can provide interface details to protein-protein interactions discovered in high-throughput experiments. Indeed, the significantly enriched motifs derived in this work may offer molecular explanations for over 6,000 protein interactions in current model species interactomes.

Novel peptide-mediated interactions derived from high-resolution 3-dimensional structures

Amelie Stein[1] and Patrick Aloy[1,2,*]

1. Institute for Research in Biomedicine and Barcelona Supercomputing Center. c/ Baldiri i Reixac 10-12, 08028 Barcelona, Spain.

2. Institució Catalana de Recerca i Estudis Avançats. Pg. Lluís Companys 23, 08010 Barcelona, Spain.

* Correspondence should be addressed to:
Tel:+34 934039690; Fax:+34 934039954; Email: patrick.aloy@irbbarcelona.org

Keywords: protein interactions, peptide-mediated interactions, interaction motifs, three-dimensional structure, interactome networks, linear motifs

Abstract

Many biological responses to intra- and extracellular stimuli are regulated through complex networks of transient protein interactions where a globular domain in one protein recognizes a linear peptide from another, creating a relatively small contact interface. These peptide stretches are often found in unstructured regions of proteins, and contain a consensus motif complementary to the interaction surface displayed by their binding partners. While most current methods for the *de novo* discovery of such motifs exploit their tendency to occur in disordered regions, our work here focuses on another observation: upon binding to their partner domain, motifs adopt a well-defined structure. Indeed, through the analysis of all peptide-mediated interactions of known high-resolution three-dimensional (3D) structure, we found that the structure of the peptide may be as characteristic as the consensus motif, and help identify target peptides even though they do not match the established patterns. Our analyses of the structural features of known motifs reveal that they tend to have a particular stretched and elongated structure, unlike most other peptides of the same length. Accordingly, we have implemented a strategy based on a Support Vector Machine that uses this features, along with other structure-encoded information about binding interfaces, to search the set of protein interactions of known 3D structure and to identify unnoticed peptide-mediated interactions among them. We have also derived consensus patterns for these interactions, whenever enough information was available, and compared our results with established linear motif patterns and their binding domains. Finally, to cross-validate our identification strategy, we scanned interactome networks from four model organisms with our newly derived patterns to see if any of them occurred more often than expected. Indeed, we found significant over-representations for 64 domain-motif interactions, 46 of which had not been described before, involving over 6.000 interactions in total for which we could suggest the molecular details determining the binding.

Author Summary

Protein-protein interactions are paramount in any aspect of the cellular life. Some proteins form parts of large macromolecular complexes that execute core functionalities of the cell,

while others transmit information in signalling networks to co-ordinate these processes. The latter type, of more transient nature, often occurs through the recognition of a small linear sequence motif in one protein by a specialized globular domain in the other. These peptide stretches are frequently found in unstructured regions of proteins, contain a consensus pattern complementary to the interaction surface displayed by their binding partners, and adopt a well-defined structure only upon binding. The information that a peptide adopts a particular structure on interaction, even though it may be unstructured on its own, is currently only available from high-resolution three-dimensional (3D) structures, and this can be as characteristic as the consensus motif itself**.** In this manuscript, we present a strategy to indentify novel domain-motif interactions (DMIs) among the set of protein complexes of known 3D structures, which can not only provide information on the consensus motif and binding domain, but also allows ready identification of the key residues on both the motif and the domain side. A detailed knowledge of an interaction interface is critical to plan further functional studies as well as for the development of interfering elements, be it drug-like compounds [1,2], novel engineered binding proteins [3] or peptides in synthetic circuits or networks [4,5]. The transient interactions involving regulatory proteins and small interfaces typical for DMIs make them interesting candidates for all these applications.

Introduction

Proteins are key players in all aspects of cellular life. They seldom act alone, but rather in combination with other molecules. Some proteins form parts of large macromolecular complexes that execute core functionalities of the cell, while others transmit information in signalling networks to co-ordinate these processes. To disentangle the complex network of protein interactions, both complex membership and binary interactions are currently being studied in large-scale experiments in several model organisms [6,7,8,9,10,11]. However, interaction discovery data mostly offers information on whether two proteins do or do not interact, but it cannot provide details on the mode of binding or the interaction interface. Atomic details of protein-protein interactions are only available in high-resolution 3-dimensional (3D) structures, which are collected in the Protein Data Bank (PDB) [12]. A detailed description of the atomic contacts involved in interaction interfaces often reveals the forces that hold two proteins together, and permits to extract conclusions on the potential disruptibility of the interface through, for instance, the action of specific drugs

[13]. Recently, the combination of structural data and assignment of globular protein domains has allowed to distinguish between two main classes of protein-protein interfaces [14]: Domain-domain interfaces tend to be large and stable, while interfaces between a globular domain and a peptide stretch are usually smaller, sometimes with only a handful of key residues involved in the binding event [15,16]. The latter type of interface allows for transient binding, making them ideal for signalling networks. The classification of interfaces into domain-domain or domain-peptide gives information on their size, strength, shape, and other features that may help us understand the interaction between the two proteins and how it reacts to competitors [17], their correct identification being thus critical. In both cases, high-resolution 3D structures provide crucial information on how proteins involved in these interactions recognize each other and achieve a high degree of specificity, which in the case of domain-peptide interactions also includes its context [16]. Furthermore, structures can also help identify key residues in binding pockets [18] or be used as the base of complex models for the prediction of domain-motif interactions (DMIs) [19]. It is thus clear that, the more high-resolution structural details we can compile on DMIs, the better we will understand their function and fast evolving profiles.

The peptides involved in DMIs are characterised by a consensus motif with specific conserved residues that are recognised by the binding domain. Some positions of such motifs are restricted to particular amino acids, while others may allow a set of similar residues, or even arbitrary ones. Consensus motifs are often given as regular expressions. For example, the Src-homology-3 (SH3) domain binds proline-rich peptides, and several variants of the PxxP (‘x’ or ‘.’ denote arbitrary positions) pattern have been observed, including [RKY]xxPxxP (class I; square brackets denote the set of possible residues for this position) and PxxPx[KR] (class II) [20]. Structurally, linear motifs are frequently found in disordered regions [21], thus exposed to potential binding partners and with the ability to adopt a variety of conformations [22,23]. Most motifs assume a well-defined structure upon binding to their recognition domain, like the polyproline type II (PPII) helix adopted by SH3-binding peptides [20], the alpha helix, formed by Nuclear Receptor cofactor peptides [24] or the beta strand in peptides that interact through beta strand addition, as do PCNA- and PDZ-binding peptides [25]. Unstructured regions may adopt different conformations depending on the interaction partner, so that a given peptide could potentially bind more than one domain, each with the appropriate structure [2]. Given the small number of key residues, motifs can arise and vanish spontaneously with only a few mutations. Along with

the modularity of their binding domains [23,26], this allows a rapid evolution to explore novel regulatory interactions relatively easily [27]. In this way, a motif that mediates a particular function or interaction can arise convergently in otherwise unrelated proteins. Due to their transient nature, DMIs are difficult to identify in high-throughput interaction detection experiments [26]. In the last years, several methods have been developed for *de novo* discovery of motifs from sets of sequences assumed to share a feature that e.g. explains interaction with a common partner [28] or other biological factors like co-localisation or particular post-translational modifications [29]. These methods exploit the convergent evolution of linear motifs in looking for patterns that are over-represented among unrelated sequences in the query set. Homologous proteins or regions are removed before the computation of over-represented motifs, and domains or other well-structured regions are often masked because motifs tend to occur in unstructured regions [21]. Sometimes, though not always, the motif-binding domain can be identified in these *de novo* procedures for motif discovery [28]. A problem of these sequence-based motif discovery algorithms is the poor signal-to-noise ratio of many datasets [30,31]. Focusing on the local environment of the motif, i.e., the *flanking regions* or *context*, increases the sensitivity of these methods [30] and may provide additional information in the search for functional interpretations of the novel motifs [31]. However, a potential caveat of methods relying on evolutionary conservation is that they might well miss some of the instances that have arisen very recently [30].

Despite the small size of their binding interfaces, domain-motif interactions are known to be highly specific *in vivo* [32], although they can also show some promiscuity, with similar affinities for native and non-native interaction partners, when tested in isolation [33]. Depending on the given binding domain, cell type and organism, the specificity may be encoded primarily in the motif sequence [18,34], the flanking regions [16], or the network context [35]; probably these factors often work in concert. Traditionally, motif recognition patterns were split by one or two key residues (cf. [36]), but recent work has revealed that a much finer subdivision may be needed [37] or that, at least for some domains, there may not even be clear borders between the recognition profiles of different members of a domain family, but that recognition profiles cover the whole specificity space instead [32]. With their high specificity, regulatory function and small interface, DMIs make excellent candidates for drug targets [13,38,39], and information about high-resolution 3D structures of the interfaces may be crucial in this context [1,2]. A recently published method by

Petsalaki *et al.* [40] searches the surfaces of 3D stuctures for sites that may bind a given motif, based on physicochemical properties. If the binding domain or a set of possible binding domains for a motif are known, this tool could help identify the interface between peptide and domain. A successful prediction of the binding site on the domain would reveal much detail beyond what is given by sequence-based approaches, yet it would not provide the atomic contacts of the interaction. There are other computational tools, such as iSPOT [41,42], that have been designed to predict peptide-binding specificities using also 3D structure information. In this case, however, one needs to know in advance that a given interaction is peptide-mediated and which are the exact residues participating in the interface.

While many current methods for the *de novo* discovery of motifs exploit the fact that they tend to occur in disordered regions, our work here focuses on another observation: upon binding to the domain, motifs adopt a well-defined structure (see also [43]). Indeed, the structure of the peptide may be as characteristic as the consensus motif, and help identify peptides even though they do not match the established consensus. An example is the linker peptide between the SH2 and kinase domains in the Src kinase [44], which adopts a PPII helix and is bound by the SH3 domain although it does not contain a PxxP motif. The interaction topology of the linker binding the SH3 domain is the same as that of intermolecular SH3-peptide pairs, so we consider them to belong to the same *interaction type* [45]. Domains may bind different kinds of peptides in different orientations [46]. During visual inspection of candidates for DMI based on experimentally confirmed motifs stored in the Eukaryotic Linear Motifs database (ELM) [36], we observed that there are several peptide-mediated interactions in structure which did not match the established consensus motifs of their corresponding interaction types and therefore could not be found by a procedure based on known patterns [16]. Furthermore, we noted that linear motifs have a particular stretched and elongated structure unlike most other peptides of the same length. We thus need to, somehow, use this information to identify more instances of DMIs in the databases, and derive the consensus binding patterns governing them. This would provide molecular details for many protein interactions discovered in high-throughput initiatives, and suggest the relevant mutagenesis experiments to tinker with them. In this manuscript we describe our studies of the structural features of known motifs, and use the results, along with other structure-encoded information about interactions, to scan through the set of protein complexes of known 3D structure in order to identify unnoticed peptide-

mediated interactions among them. We compare our results with established linear motif patterns and their binding domains as described in ELM [16], and with other sources of structural descriptions of peptide-mediated interactions. Finally, we cross-validate our newly derived patterns on interactome data, and present a list of novel peptide-binding domains, along with their respective high-resolution 3D structures and consensus motifs.

Results

Structural parameters to capture linear motifs

To exploit structural features of peptides and domain-motif interactions, we first need to establish which parameters are suitable for the separation of known linear motifs from other, presumably non-functional peptides of the same length. We thus selected several structural parameters and applied them to the 631 DMIs of known 3D structure identified in our previous study [16], to test whether they could capture structural properties of functional linear motifs. To ensure that these parameters could partition peptides into true motifs and random cases, we created a control dataset based on the Structural Classification of Proteins (SCOP) [47]. For each SCOP fold, we chose one representative structure and generated all possible peptides of length 4-20 residues, which corresponds to the range of motif lengths in ELM. Although we cannot guarantee that SCOP folds do not contain true linear motifs, it is unlikely as they form well-defined tertiary structures, whereas motifs often occur in unstructured regions and outside domains. Therefore we assume that the SCOP control set constitutes a reasonable collection of negative instances for the identification of linear motifs.

The main structural parameters we developed for motifs are *linearity* and *elongation* (Fig. 1a). The *linearity* of a peptide is a marker of how “flat” it is, how much it deviates from a straight line through the first and last residue. The *elongation* indicates how long a peptide extends in space (Fig. 1a; for details see *Methods*). Together, linearity and elongation should capture our observations described above, namely that linear motifs are more flat and stretched than other peptides. Because flexibility and length of a peptide increase with the number of residues, it is important to only compare peptides with the same length in residues. We computed linearity and elongation for the known DMIs and the SCOP control set and found that, individually, neither of these parameters showed sufficient difference

between the known cases and the SCOP control set, although there was a trend for known DMIs to be longer and more linear than other peptides of the same length in residues, confirming our observations (Supplementary Figures 1a,b). We also assigned secondary structure to the peptides using DSSP [48] and observed that the distributions of values for linearity and elongation differed strongly between the classes of secondary structure. Helical structures were shorter and less linear, while beta sheets and unstructured regions were more linear and longer (Supplementary Figures 1c,d). However, again the differences were not clear enough to use them to separate known DMIs from other peptides. Note that, as described above, some linear motifs act through beta strand addition, and yet others are known to form alpha helices, though most are found in unstructured regions. Since no single parameter was able to divide known DMIs from other peptides, we combined them to see if the trends described above would give a synergistic effect. For pure geometric considerations, a peptide that is flat should also be elongated in comparison to one that is helical or has bends and turns. Indeed, we found that known motifs fell into distinct regions of the space spanned by elongation and linearity, and are further subdivided by classes of secondary structure (Fig. 1b). Thus we concluded that these three factors can be used to separate structures similar to known motifs from other peptides of the same length in residues. Based on these findings, we set out to exploit structural data to find new instances of peptide-mediated interactions among the over 50,000 high-resolution 3D structures stored in the PDB (Fig. 2). However, the SVM will only recognise structural features of the peptide, but not consider any interactions to surrounding domains. Yet we cannot recognise DMIs based on the peptide alone, we need to take the interaction environment into account. We trained a support vector machine (SVM) with the data for linearity, elongation, secondary structure, accessibility and length in residues for all the known DMI peptides, and for a random set of 10,000 SCOP control set peptides (for details see *Methods*). Our first parameter that takes the environment of the peptide into account, accessibility, is required because peptides need to be accessible by other proteins in order to mediate interactions. Additional filters concerning the interaction environment will be described in the following paragraphs.

We generated all possible peptides of lengths 4-20 residues from 52,903 3D structures, excluding regions covered by domains as assigned by Pfam HMMs [49] because motifs are rarely found in these. Note that this creates many overlapping peptides (cf. Fig. 2), which we generate in order to find the largest peptide that is accepted as *linear* and

elongated enough. From the 60,123,359 candidates, only 10,596,512 (18%) peptides in 41,224 structures were accepted by the SVM. Next they were filtered for contacts to neighbouring domains in order to find putative domain-peptide interactions. We intended to identify all peptide-domain interactions, regardless of whether they appear within or between proteins. Therefore, for each candidate peptide that had been accepted by the SVM, we checked for contacts with domains in the vicinity, independent of whether they are part of the same protein or of another. We did not find enough contacts for 2,890,451 peptides (details see *Methods*), meaning that 7,706,061 peptides (73% of the accepted peptides; percentages in the motif discovery pipeline will always refer to the previous number of peptides or DMIs) in 40,199 structures remained. Next we removed some of the overlap that arises due to the way peptides are generated: If a short peptide has been accepted by SVM and domain contacts, and a longer peptide that includes the short one has been accepted as well, we only keep the long peptide (Fig. 3). Note that there may still be partially overlapping peptides in the set, in cases where none of the peptides covers the other completely. This overlap will be addressed later, as we cannot simply join peptides unless the resulting, encompassing peptide is also accepted by the SVM which tests for the typical structural features -- linearity and elongation -- of linear motifs. After the removal of completely overlapping peptides, we end up with 538,689 peptides (7%) in 40,199 structures, which are involved in 782,430 interactions, since one peptide may interact with several domains.

Filtering candidate peptide-mediated interactions

At this step, we noticed that most of the DMI candidates corresponded to intrachain interactions (80%) and, upon visual inspection, many did not seem to be functional. These were often cases in which the domain and putative peptide, while in contact, had no extended binding surface, but rather protruding side chains touching each other, such that it was not clear whether this contact was biological or whether it might be an artifact that arose e.g. due to crystallization or buffer conditions. Other instances arose because of domain definitions that did not perfectly match the structure, in other words, when the Pfam domain assignments did not fully cover the structural units (folds), so that remaining elements (single strands or helices) were identified as binding peptides. Therefore we used domains in the protein structure classification CATH, which are defined on 3D structures [50], to filter out 426,464 (55%) intrachain peptide-mediated interaction candidates that were covered by these domains. Furthermore, many intrachain

interactions were observed between a domain and a peptide close to the domain's boundaries. Accordingly, we removed 87,986 (25% of the remaining) interactions with a sequential domain-peptide distance below 10 residues. In addition, we filtered out contacts between proteins that are not listed in the Protein Quarternary Structure (PQS) database (2%), which contains presumed biological units of protein structures rather than the asymmetric units calculated in the structure determination process. The latter may bare signs of artefacts such as crystal packing. Also, during visual inspection we observed peptide candidates suggested to mediate interactions among multihomomers, which did not appear to be functional (visual inspection). As peptide-mediated interactions are usually heterologous, we removed cases in which the domain-containing protein and the peptide-containing protein form a homomer (1.4% of the candidate DMIs). Note that intrachain DMIs are heterologous as well, such as the SH3-peptide and SH2-phosphopeptide interacctions in the Src kinase described above [44], and structurally of very similar nature as their interchain equivalents [51].

Besides the key residues that form the consensus pattern, linear motifs are characterised by the fact that binding of the motif itself is sufficient to create a functional interaction (e.g., [52]). As we could not perform computationally expensive studies of binding energies for all candidate DMIs, we approximated the binding contribution of the peptide by comparing the domain-peptide interface with the full interface between the two partners. Specifically, we required the interface between domain and peptide to be at least 150Å^2, which holds for over 90% of the known DMIs (Supplementary Figure 2a), but may filter out putative interactions that are due to artefacts. 34,898 (14%) of the candidates had a smaller interface and were thus removed. In addition, to ensure that the peptide is a key player in the interactions detected in our procedure, we required the interface between peptide and domain to cover at least 50% of the total interface between the two proteins, which is true for 65% of the known DMIs (Supplementary Figure 2b). The 50% threshold is intended to reflect our requirement for the domain-peptide-interface to have a major role in this interaction. Some peptide-mediated interactions are formed by multiple domains binding a peptide, e.g. the seven-blade beta-propellers formed by WD40 domains [53]. In those cases we only required all domain-peptide interfaces together to make up 50% of the full interface, and individual domains to contribute roughly equally (see *Methods*). Application of this filter removed 212,125 (95% of the remaining) putative DMIs, so that 10,739 candidates remained.

Clustering of candidate interactions by sequence and topology

In order to identify unnoticed DMIs within the PDB, we needed to classify distinct domain-peptide interfaces and search for regular shared features among the peptides that explain why they bind the domain – a consensus motif. Thus, we needed to group the candidate interactions by topology to separate distinct interaction interfaces. Furthermore, due to the redundancy among entries in the PDB, we needed to create non-redundant sets of peptide-domain interactions. We only attempted to derive a consensus motif if sufficient non-redundant information was available. Note that we did not remove such redundancy in previous steps to capture as many variations of DMIs as possible.

The topological clustering procedure we developed focuses on the residues forming the interface. We computed the fraction of shared peptide-binding residues between each pair of domains from a family, mapping corresponding residues via alignment to Pfam's HMM profile for that family (see Fig. 4 and *Methods*). Next we clustered the interfaces based on the shared peptide-binding residues to separate all interactions for this domain into distinct interfaces. Our method is similar to that by Teyra *et al.*, [54] but relies on multiple instead of pairwise alignment of the domains. In total, we found 822 topological clusters or interaction types, including 547 domains. The largest clusters contain over 700 DMI instances. Domains with many different peptide-binding topologies include protein kinases (8), trypsin (14), Pyridine nucleotide-disulphide oxidoreductase (Pyr_redox_2, 15 topologies), and the immunoglobulin V-set domain (18). However, note that these potential ligands have not been examined for significant motifs yet, so they do not necessarily represent functional DMIs. The sequence-based clustering, which is independent of the topological clustering, serves the creation of a sequentially non-redundant set of peptides bound by a given domain for the derivation of consensus patterns. It is used to create groups of “sufficiently different” peptides per domain, to establish which peptides are different enough to qualify for *motif support,* for which non-redundant data is required. Ideally, a set of unrelated sequences would allow over-represented motifs to be detected easily, as similarities cannot be due to larger conserved regions. To this end, first we need to address the issue of partially overlapping peptides that arises from the way peptides are generated (Fig. 2). A sequence-based clustering procedure working on the pure peptides, which are at most 20 residues long, could not detect small overlaps. If such a small overlap were to contain a motif, it would then be supported (counted) twice. To avoid these

duplicated counts, we first joined peptide-containing stretches such that each protein section that was continuously covered by peptides (i.e. without gaps) was combined into a single region (see Fig. 3). Then these non-overlapping regions were aligned, and their pairwise sequence identity was computed. Clustering by sequence identity was based on combined sequence identity scores for both domain and peptide (see *Methods* for further details). In total, we found 2,490 clusters of 90% sequence identity for the 547 domains, with the largest clusters having over 200 entries. The immunoglobulin V-set domain shows the greatest sequential diversity in its ligands (220 clusters), followed by trypsin (108) and Major Histocompatibility Complex I (105). We are fully aware that 90% sequence identity is a more stringent threshold than what is normally used when creating sets of unrelated proteins. Our reasons to apply such a strict criterion are twofold: we are handling relatively short peptides, on which alignment does not always work reliably, and by selecting a lower threshold, such as 25% or 50%, we would risk getting too different peptides within the same cluster. In addition, we also need to identify motifs even among sets of proteins that are relatively similar (i.e. motifs occurring in the same protein family). Thus the sets of proteins in our clustering should be considered as *non-redundant* rather than *unrelated*. Nevertheless, we explored the possibility of relaxing the sequence similarity threshold to 50% in the clustering procedure (data not shown), and found that, since the clusters are broader and cover more instances, the number of interaction types with sufficient non-redundant information to derive significant patterns (see subsequent paragraph) dropped from 224 to 96.

Deriving consensus motifs

For each topological cluster with at least 3 non-redundant sequences, we attempted to derive a consensus motif using SLiMFinder [29]. SLiMFinder identifies convergently evolved linear motifs in a set of sequences based on their occurrence in unrelated sequences, and computes a probability of their significance. It often suggests more than one possible motif, ranked by their probability of arising by chance. The program requires information about the evolutionary relationship of the motif-containing sequences, in particular which of them are so closely related that they should be considered a single case of support for a candidate motif when it is examined for whether there are enough unrelated proteins matching it. We provided the 90% sequence identity clusters for this purpose, so that only non-redundant cases are counted for motif support. Among our 822 topological clusters or interaction types, only 224 contained at least 3 non-redundant

sequences. These covered 157 domains, with up to 13 clusters per domain. In addition to the sequences and their evolutionary grouping, we included information on modified residues in the peptides, because some recognition domains, like SH2 or 14-3-3, specifically bind peptides that have been post-translationally modified, e.g., by phosphorylation or methylation [23,55,56]. The domain's binding groove recognises the residue with the post-translational modification, so it should be a crucial element in a consensus motif. We searched all accepted peptides for modified residues, and if a particular residue was modified in more than half of the peptides in a cluster, it was required for the motif. Furthermore, we checked for helical peptides, which are another special case for pattern derivation, as helical structures create a regular pattern of residues pointing towards the domain vs. residues pointing away from it. If over 50% of the peptides in a cluster were helical according to DSSP, we enabled the helical pattern derivation feature of SLiMFinder, which takes this particular spacing into account.

For 152 of the interaction types, covering 111 domains, we found at least one significant motif, and for 96 of these, significant motifs were found for all topological clusters (interaction types). The 46 remaining domains did not yield any significant motif, although 3 or more non-redundant sequences were available for pattern derivation. As a curiosity, 36 of the interaction types with significant motifs (31 domains) involved helical peptides, and 20 (18 domains) required at least one modified residue in the pattern. In total, 5,316 interactions in 3D structures are covered by these 152 interaction types for which we could derive a significant motif, including 4,202 inter- and 1,114 intra-chain interactions, respectively. However, 16 clusters among 15 domains contained only intrachain interactions that, upon visual inspection, did not seem to be functional peptide-mediated interactions and were thus excluded. In addition, 8 putative DMI interaction types are always found between proteins that also have a domain-domain interface (DDI), which presumably is more reliable. Our assumption here is that if the DMI is functional, it should also occur independent of a DDI. Hence we modified our method not to accept clusters unless there are interchain instances, and DMI that appear without a DDI in the same protein pair. It is interesting to note that only 18 of the 94 domains for which we find significant patterns in the full dataset are described in ELM. It should be noted though that ELM does not always provide Pfam domain names and thus the overlap could be slightly larger and we are just not able to detect it.

Benchmark of the DMIs identification accuracy

To assess the performance of our method in detecting peptide-mediated interactions and discarding non-functional peptides or interfaces in the PDB, we created a benchmark set of 631 known DMIs [16] and 631 random peptides from the SCOP dataset that do interact with a domain in a different protein (i.e. we only kept interchain training data) and are not fully covered by a domain. To ensure that DMIs are recognised by features beyond similarities among homologous domains and their binding peptides in the training set, we performed the benchmark in a *leave-one-domain-out* fashion, i.e., we removed all peptides binding to a given domain from the training set, and tested the recovery of the corresponding interactions and the detection of its consensus motif using the resulting SVM. For example, in one instance we left out all SH2-binding peptides (the *test set*), then re-ran the full motif discovery pipeline as described above and finally tested how many of them were rediscovered by the SVM trained on the remaining, non-SH2-binding peptides (the *training set*). If a peptide overlapping in at least 3 positions with a test set peptide was accepted and a significant pattern for its domain and topological cluster could be derived using SLiMFinder, the case was classified as “positive”, otherwise (peptide not accepted or non-significant pattern) as “negative”. For known DMIs, we also tested whether the known consensus motif given in ELM scored significantly using SLiMSearch [57], which works similarly to SLiMFinder but allows checking the significance of a predefined motif on a given set of sequences.

After applying the described procedure, we could automatically rediscover 423 of the 631 known DMIs interaction types, which correspond to a sensitivity of 68%. In terms of domains, we correctly recovered cases for 20 out of the 30 domains, i.e., the domain-based sensitivity (67%), very similar to that based on individual cases. Our method did not accept any of the negative cases from the benchmark, indicating that it is highly specific. Analyses on the 208 known DMIs that we could not recover showed that almost half of them (42%) can be explained by the fact that they are covered by domains and thus never considered as peptide candidates by our method, while manually curated sets like ELM and our previous study [16] did not apply such a filter. Other reasons for non-rediscovered positive cases included insignificant patterns (21%) or no pattern determined because of a lack of data (12%), too few contacts between domain and peptide (8%), and insufficient surface contribution of the domain-peptide interface (7%), among others. There are three domains for which the benchmark returns positive results, but no significant pattern is

found when applying our method to the full PDB, which may be due to differences in the data set size in significance computation. Note that cases with too few non-redundant sequences were ignored (i.e., they are not counted as false negatives) for both negative and positive test cases. As an additional independent benchmark, we tested how many of the peptide-mediated interactions from the benchmark set by Petsalaki *et al.* [40] could be identified with our approach. In total, we recovered 298 of the 405 DMIs in their set (74%), which is slightly above the 240 cases (59%) that they correctly predicted from their benchmark. Analysis of the results showed that half of the instances that we missed are due to a low surface contribution, while the other half was covered by domains or not accepted by our SVM.

Cross-validation with interactome networks

To assess the validity of each motif derived by SLiMFinder and confirm that it could indeed occur in different protein interactions, we checked whether it was over-represented in proteins known to interact with a partner that contains the respective binding domain (see *Methods*). For example, although motifs tend to be degenerated, 50 of the 593 (8.4%) human proteins that interact with other proteins containing 14-3-3 domains match the ELM pattern *R[SFYW].S.P*, but the number is reduced to only 206 out of the 7215 (2.8%) of the proteins not interacting with any 14-3-3-containing protein. This corresponds to an enrichment factor of 2.57, which is statistically significant (p-value 2.827e-10, one-sided Fisher's exact test). We used interactome data for selected model organisms with relatively good coverage (yeast, worm, fly, human). We only tested motifs binding to domains of which a structure was solved in this species, to make sure that there is a functional occurrence of it in the species in question, so this validation procedure was limited to 64 domains among the 111 for which we suggest binding motifs. To avoid false positive hits, we required pattern matches be outside of globular domains, as do many sequence-based tools for motif detection [28,29]. In addition, we only counted those motif-containing proteins as support for the DMIs if the interaction cannot be explained by a domain-domain interaction between that protein pair [15]. For each domain, we tested all patterns derived by SLiMFinder with the given parameters (see *Methods*) that is found in one of the selected interactomes. We then computed whether proteins interacting with a partner containing the recognition domain are enriched in hits for the derived motif (one-sided Fisher's exact test, p-value threshold 0.025). We found significant enrichments for 64/90 of the interaction types and 46/64 of the domains, with 1 to 6 patterns per interaction

type enriched. Fig. 5 shows structures for each interaction type found to be enriched in the interactome cross-validation along with its most significantly enriched motif, which is not necessarily the top-ranked by SLiMFinder. Across the interactomes of the four model species considered in this cross-validation, our DMIs described here could offer molecular bases for 5199 interactions in human, 160 in fly, 19 in worm and 941 interactions in yeast. Applying the statistical test to the 44 known ELM motifs for which we found a 3D structure [16] reveals significant enrichment for 72% of the DMIs and 74% of the domains, which is slightly higher than that observed for the patterns derived in this work. Looking at all 66 ELMs we considered for our previous study (i.e., also including ELM patterns for which we did not find occurrences in 3D structure), we find significant enrichments for 55% of the DMI and 59% of the binding domains. This decrease might suggest that motifs with known 3D structures are somehow better defined. Table 1 shows the list of DMIs for which we found a significant enrichment in the interactome networks, together with the best-ranked pattern according to SLiMFinder and the most significantly enriched. The complete list of patterns is provided in Supplementary Table 1.

A look at specific examples

Although it is the general trends that conform the main message of this work, it is always illustrative to look at some specific examples to understand the nature of our results. Among the novel DMIs identified by our approach there are two that, in the meantime, have been included in ELM, but were not listed when our training set was built. These can be considered “blind tests”, since neither the SVM nor other filtering parameters were selected using the information in these domains and peptides. One case is the BRCA1 C Terminus (BRCT) domain, which forms dimers that bind a phosphopeptide [58]. The different orientations of the domain with respect to the peptide are recognised in our procedure, and two topological clusters are generated (see also Figure 4). The best-ranked pattern is *S..FP*; where the S is always found phosphorylated. The ELM annotation also describes a phosphorylated S, along with two similar patterns (*.S..F* and *.S..F.K*).

The second DMI from our results that has been added to ELM is cytoskeleton-associated proteins domain (CAP_GLY), which is involved in the regulation of microtubules [59]. It recognizes short D/E-containing peptides; the consensus motif that we derived is *DE.F* (or *D.{0,1}E.F*) while the ELM pattern is a much longer one: *[ED].{0,2}[ED].{0,2}[EDQ].{0,1}[YF]$.* ELM contains the additional information that the

peptide is always C-terminal, indicated by the $ symbol. In our approach we do not try to establish whether a pattern occurs at one of the termini, because peptides in 3D structures are often truncated, so that what appears to be a terminus won't necessarily be one *in vivo*. In this case, however, neither the ELM CAP_GLY pattern nor the one we derived is significantly enriched in proteins interacting with those containing the domain.

A very interesting example is that of the Bcl-2 protein family, which is crucial in the regulation of apoptosis and has both pro-apoptotic and pro-survival members [60,61]. Many of these are multidomain proteins that contain four conserved Bcl-2 homology (BH) domains. In addition, some members of the extended family only contain one of the BH domains, BH3, which forms a helical peptide that can be bound by multidomain Bcl-2 proteins [62]. As survival despite pro-apoptotic signals is a problem in many cancer cells, this family comprises several interesting drug target candidates. Indeed, a number of small molecule agonists and antagonists have recently been developed and are currently in various stages of clinical trials [62], some of which have been developed based on 3D structures of Bcl-2 and its binding peptide (the BH3 domain). The family is also listed as a peptide-binding domain on the Pawson lab web site [63], named BH1-BH2-BH3-BH4, but a consensus motif for the peptides is not given. The top-ranked significant motif we identified is *L..I[AG]D.[ILV]*, with the large hydrophobic residues pointing into the binding groove. Two other, very similar motifs scored significantly (*LR.I.D.[LV]* and *R.I[AG]D.[LV]*); both also contain the large hydrophobic residues in the appropriate spacing pattern. Structurally, the peptide is always helical, so one might consider replacing the arbitrary positions (.) by anything but proline [^P], because of the helix-breaking properties of this aminoacid. This restriction is also found in other motifs, such as ligands of hormone receptors (Hormone_recep), another all-alpha protein that binds small helical peptide ligands [24].

Finally, in contrast to the three examples above, we could not identify a significant motif for Clp protease (CLP_protease), although sufficient non-redundant sequences were available. Given that Clp proteases degrade peptides with little sequence specificity [64], the fact that our approach could not identify a defined consensus motif should be considered positive for our method.

Discussion

The identification and correct classification of domain-motif interactions is a key issue to understand the biophysical principles governing interactome networks, such as the relationship between protein-binding domains and the consensus motifs they recognize. Accordingly, we have presented a method to indentify unnoticed domain-motif interactions (DMIs) among high-resolution 3D structures, which not only provides information on the consensus motif and binding domain, but also allows ready identification of the key residues on both the motif and the domain side. Applying this methodology to all currently available 3D structures has revealed 152 DMIs, 127 of which have not been described previously. Moreover, 64 of the motifs have been found to be significantly enriched in proteins interacting with those containing the respective binding domain. In a *leave-one-domain-out* benchmark on the 3D structures of known ELMs [16], our method could rediscover and compute consensus motifs for 2/3 of the known cases. In addition, it is very precise as none of the random motif-domain pairs we tested as negative control cases were accepted. As far as we know, no other method for *de novo* motif discovery can provide such details for a novel DMI. Indeed, few other methods exploit the information encoded in 3D structures, although the importance of the 3D structure of motifs and their flanking regions for functional analysis has recently been highlighted [31]. The information that a peptide adopts a particular structure in interaction, even though it may be unstructured on its own, is currently only available from high-resolution 3D structures. While sequence-based methods for motif discovery have the advantage that they are applicable to larger datasets, they cannot necessarily reveal the binding domain of a suggested motif [28], or the atomic details of the interface. Knowing the interface of an interaction, however, is critical for functional studies [17] as well as the development of interfering elements, be it drug-like compounds [1,2], novel binding proteins [3] or engineered peptides in synthetic circuits or networks [4,5]. The transient interactions and small interfaces in DMIs make them interesting candidates for both applications. With previous large-scale methods, gaining this knowledge in one step was not possible. Only once a pattern and its binding domain have been identified, the recently published method by Petsalaki *et al.* [40] can be used to search structures of this domain for surface patches that are complementary to the pattern, although it has some difficulties with helical and beta-strand-forming peptides. In addition, without further information, that method cannot show which residues in the domain interact with the key residues in the motif, and would

thus benefit form a combined strategy with the approach presented here.

A different issue with the sequence-based quest for motifs is that the “shared feature” may be too loosely defined. For example, there are many kinases that phosphorylate [ST] or Y residues, yet it would not be possible to derive a meaningful motif from the set of phosphorylated sequences alone. Our method instead focuses on the atomic interface between peptide and domain, which includes a concise definition of the environment in which the motif is bound. This, in turn, ensures that all peptides in one group for pattern derivation use the same interface, which is a rather strict definition of “shared feature”. An additional advantage is the direct possibility of visual inspection of the suggested DMIs. The main bottleneck is the availability of domain-peptide interactions in 3D structures in large enough numbers and with sufficient diversity to allow for the derivation of a consensus motif. However, if this information is available, the results are highly specific and contain a level of detail that cannot be provided by other techniques. Furthermore, our method successfully detected helical peptides or those acting through beta-sheet addition, which present difficult cases for the other 3D-structure-based method for DMI interface detection [40]. Yet, while our method successfully identifies DMIs in helical peptides, we do not fully exploit the information provided in those structures -- for the peptide to be helical, it should not contain proline residues [65]. However, even though no prolines occur in the sequences, SLiMFinder cannot determine that this residue is “forbidden” because the amount of information encoded in the sequences used for training is much too small for such conclusions. Only studies on all possible sequence variations, such as phage display data, could allow the derivation of forbidden residues in certain positions. However, one might consider manually modifying the patterns of helical peptides to reflect this, by replacing all arbitrary positions (.) by [^P], as it has been done e.g. for the hormone receptor ligands in ELM (pattern [^P](L)[^P][^P](L)(L)[^P], parentheses indicate that the leucines are key residues).

We chose strict thresholds on contacts and interface size to limit the occurrence of false positives, which are often problematic when dealing with so few key residues as in motif-mediated interactions. While our high precision shows the advantages of those strict thresholds, we do miss some true motifs as described in ELM, in particular due to the exclusion of regions assigned to Pfam domains, which is responsible for almost half of the true motifs we do not recover. Yet the inclusion of these regions would disproportionately

increase the computation time as well as the risk of false positives, since it has been shown that motifs usually occur outside of domains [21]. Nevertheless, it may be possible to create a fine-tuned version of our method that is able to also detect motifs located in domains. The fact that we do not find a pattern for unspecific cleavage sites (e.g. for the Clp proteases) shows that the method is also capable of separating random peptides from functional ones at the stage of motif derivation, should random peptides have been accepted by the SVM. Some motifs that were only detected in interactions between a peptide and domain in the same protein (intrachain interaction) could not be confirmed as true DMIs upon visual inspection and thus, we cannot assume to derive motifs accurately from intrachain data alone. This issue may improve if intrachain DMI data would be included in the training data, which is currently not the case as no reliable collection of intrachain peptide-mediated interactions is available.

It should also be noted that our final list of DMIs (Table 1 and Fig. 5) only includes those cases that could be confirmed in the interactome cross-validation, even though we know that some real cases, like the cytoskeleton-associated proteins domain (CAP_GLY) and many other known ELM motifs, are not significantly enriched in the current interactomes. This issue may improve with growth of protein interaction databases. Likewise, newly solved 3D structures may contain new DMIs, or raise information content for existing ones above the threshold required for application for our method. We cannot expect to recover the exact patterns described in ELM, which are manually curated and often exploit dedicated experiments to the relevance of a particular position or residue. Yet both our patterns and those from ELM score significantly in the datasets derived from 3D structures, and manual comparison shows that they are often similar. A potential problem is that the motifs we derive can only take sequences into account that occur in 3D structures, which may introduce a bias that would not be present in studies on all possible binding peptides. This might be addressed by applying methods such as iSPOT on the 3D structures identified here, combined with data from phage display scans [42]. On the other hand, we can include information on modified residues and particular spacing patterns in motif derivation, which are usually characteristic for a domain family and not just for individual instances. Recent studies have shown that the binding preferences of individual domains are probably too complex to be captured in regular expressions but that more complex models will be required [19,37]. In addition to the physicochemical binding preferences of the domain, contextual factors will govern which interactions happen *in vivo*

and which do not [23,66]. The importance of the context may vary for different recognition domains and biological processes; for example, phosphorylation networks appear to heavily rely on contextual information [35], while a few (phosphorylation independent) domains have been shown to very specifically recognise the amino acid sequence of their binding partners [18,34]. Again, more complex models will probably be required to integrate all this information that leads to the *in vivo* specificity of any given protein. Nevertheless, consensus motifs can be very helpful in studying commonalities among and differences between peptide-binding domains. A side effect of the large-scale derivation of motifs is that peptides are always given together with their binding domain in a way that can easily be transferred to other sorts of data, which is not always the case for manually curated DMIs. For example, by matching the pattern in one sequence and the binding domain in another may explain the mechanism behind some of the many protein-protein interactions which are currently being discovered in high-throughput interaction discovery experiments. This, in turn, may make them amenable to tinkering with this part of the network, either by designed peptides that might have fine-tuned affinity and specificity for the binding domain [4,67] or by drug-like compounds that can interfere with the interaction [1,2,38], a long-standing dream of the pharmaceutical industry [39,68,69].

The information we exploit here to identify novel DMIs among the set of interactions of known 3D structure (i.e. well-defined structure upon binding), is of different nature than the one used by more traditional sequence and disorder-based methods applied by many other tools [28,29,30]. We think that our approach presents an extension to the currently available techniques, and should be regarded as complementary to them. Sequence-based discovery methods have access to much more data, especially with current high-throughput interaction and other functional association studies. The surface-searching method also accesses a larger pool of data, because more structures of individual domains are available than of DMI. Yet neither method can provide the level of detail we access here. The combination of these two kinds of data, with 3D structures to define motif interfaces and large sequence databases to establish evolutionary over-represented patterns, will certainly make a powerful predictor of linear motifs. On a limited scale, we already created such a hybrid method in the interactome cross-validation, yet more sophisticated implementations could include a wider variety of biological data and tackle the problem of capturing specificity, on the more abstract level of domain families as well as on the level of individual domains.

Methods

Control dataset

The control data is based on the Structural Classification of Proteins (SCOP) [47]. For each fold, we chose one representative structure and generated all possible peptides of 4-20 residues length. This dataset contains high overlaps among peptides from the same structure, which gives us a large variety of possible structural conformations of peptides to study. For the benchmark, we masked all peptides in regions covered by Pfam domains [49] and selected unmasked background peptides that had at least 4 contacts with a domain in a different protein chain as negative control cases. Thereby we ensured that negative cases had a chance to be identified as candidate peptides and would not be removed by our filters on peptides covered by domains and the minimum number of contacts.

Structural parameters

The *Elongation* or "length" of a peptide is the distance between the C_α of the first and last residue of a peptide in Angstroms (Å). Because flexibility increases with peptide length in residues, short peptides have a small range of possible elongation values, while it varies more for long peptides. The *Linearity* is computed by constructing a line through the first and last C_α of each peptide, then calculating the distance of each C_α in the peptide to this line, and returning the maximum distance. A low value indicates a very flat or linear peptide. We used DSSP [48] for *secondary structure* assignments to peptides from the set of known DMI and from the SCOP background. The assignment is done on a single protein, after removing other chains in the structure but before extraction of the peptides, as DSSP does not always perform well on small fragments. Each peptide is assigned the DSSP class most frequently found among its residues. Note that because SVMs work on numerical data, including a more detailed description of the secondary structure of a peptide, such as the order of DSSP classifications, would require a much more complex model. We also used *accessibility* data computed by DSSP, and assigned the average accessibility of its residues to each peptide.

Training of the Support Vector Machine (SVM)

We used the implementation "SVM-light" [70] and trained it on our data with a cost-factor of 10, meaning that errors in the classification of positive cases are 10 times worse than

errors in the classification of negative cases, a trade-off between training error and margin of 0.1, and a linear kernel. These parameters were selected after searching the parameter space for different combinations of values for cost-factor, trade-off and kernel function, and testing recovery of known positive and negative (SCOP control set) cases. The estimation of classification errors also takes the fact that our set of positive cases is much smaller than the negative set into consideration.

Interactions with neighbouring domains

To form a DMI, each peptide accepted by the SVM was also required to interact with a nearby domain, which may be part of the same protein or of another. Linear motifs usually form one connected interface with their binding domain, thus we excluded peptides in which there was a region of more than 4 residues that did not contact the domain, as well as those cases in which less than 60% of the peptide residues made contact with the domain.

Interface size and ratio

We used NACCESS [71] to compute the interface between domain and peptide, and the "full interface" between the domain-containing protein and the peptide-containing protein (interchain DMIs) or between the domain and the rest of the protein (intrachain DMIs). In general, to accept a peptide-mediated interaction, the domain-peptide interface has to constitute at least 50% of the full interface. To accommodate for different stoichiometries in domain-peptide interactions (multiple domains binding one peptide), the threshold for the interface ratio was set to $0.5/N$, where N is the number of domains involved in the interaction. If domains do not contribute roughly equally to an interaction or, more exactly, if any domain contributes less than $N/0.5$, they are removed in a filtering step. Since this changes the number of domains involved in the putative DMI and hence the minimum interface size, we repeated the filtering until each domain-peptide interface contributed appropriately for its stoichiometry and no domain was removed any more in the given step; in other words, we repeated the procedure until convergence. Note that, while all domain-peptide-interfaces have to contribute at least $0.5/N$, they do not necessarily contribute equally.

Clustering by topology

We first aligned all sequences for a given domain to the corresponding Pfam HMM profile

[49]. The aligned positions are used as a normalized numbering of the sequences, allowing easy comparison and mapping of corresponding positions. Next we computed the contacts for each domain-peptide interaction, and compared for each pair of domains how many of the corresponding domain positions contacted the peptide in both cases (c_b). The resulting distance score is $1 - \frac{2 \times c_b}{c_1 + c_2}$, with c_1 and c_2 being the number of domain positions involved in contacts for the two respective domains. If the interface sizes are vastly different (one has more than double the contacts of the other), we set the score to 1, as these interfaces are considerably different despite possible overlaps (visual inspection). After computing the distance matrix for all DMIs involving the given domain, we clustered the interaction topologies by complete linkage hierarchical clustering [72] and cut the resulting tree at a distance of 1, which corresponds to no shared contacts (or artificially separated cases with large differences in interface size, cf. above).

Computing peptide regions for overlapping peptides

Each continuous part of a protein that was covered by one or more peptides was designated as a peptide-containing region. These regions are non-overlapping by definition, and they represent parts of the protein that contain one or more peptides structurally similar to those found in known DMIs. The main motivation for this was that each pattern match in a given structure should only be counted once for “support” of that pattern, independent of how many accepted peptides include it (cf. Fig. 3).

Clustering by sequence

For the sequence-based clustering, domain and peptide alignments are computed individually, and then the pairwise similarities are combined into one score, which is then used for clustering. For domains, all sequences of a given family are aligned to the Pfam HMM profile, and the sequence identity is computed from this alignment, yielding a pairwise domain sequence identity score $s_{d_{ij}}$ for each pair of domains. The corresponding peptide-containing regions are aligned using the Needleman-Wunsch algorithm [73], yielding a pairwise peptide sequence identity score $s_{p_{ij}}$ for each pair of peptides. The distance score is then computed as $1 - \frac{s_{d_{ij}} + s_{p_{ij}}}{2}$ for each pair *ij*, where *i* and *j* are candidate domain-peptide interactions. Like for the topological clustering, we applied complete linkage hierarchical clustering [72]. Note that we use both domain and peptide sequences,

because for the diversity of a DMI it makes a difference whether a given peptide is always bound by the same domain or by different domains. As a cut-off here we chose 0.1, which corresponds to 90% sequence identity. Thus, all resulting clusters have a combined sequence identity of 90% considering both domains and peptides.

Motif derivation

We used SLiMFinder [29] to derive consensus motifs for the sets of peptide sequences bound in each topological cluster. We computed the amino acid frequencies from all sequences in the PDB, and disabled the 'termini' flag because beginnings and ends of our sequence fragments usually do not correspond to actual protein termini. "Unrelated protein clusters" (UPCs) were defined using the sequence-based clustering described above. Only topological clusters with 3 or more UPCs were searched for consensus motifs; the information content is too low in the other cases.

Enrichment of DMIs in interactome networks

We created interactomes for human, fly, worm and yeast by integrating protein-protein interaction data from the databases MINT, IntAct and HPRD [74,75,76] that are supported by peer-reviewed publications. To ensure species specificity, we excluded hybrid interactions observed between proteins from different species, resulting in networks with 53,290 (human), 19,260 (fly), 5,566 (worm) and 60,721 (yeast) interactions, respectively. As described above, we only considered interactions that could not be explained by domain-domain interactions as observed in 3D structures [15], which reduces the interactomes to 43,882, 18,113, 5,234 and 58,426 edges, respectively. These interactomes involve 7,808 human, 6,610 fly, 3,111 worm and 5,266 yeast proteins, respectively.

To calculate motif enrichments in the interactome networks, we assigned Pfam domains, via HMM profiles, to all proteins in the respective interactomes, and tested for motif hits by regular expression pattern matching, only considering regions outside domains as described above. We then created a contingency table for each motif and species stating how many proteins contained at least one motif match, and how many interact with a protein containing the motif's binding domain. The enrichment factor was computed as $\frac{p_{im}/p_i}{p_m/p}$, where p_{im} is the number of proteins that interact with another protein know to

contain the binding domain and also have a motif match, p_i is the number of proteins that interact with another containing the binding domain, p_m is the number of proteins with a motif match, and *p* is the total number of proteins in the interactome. The p-value was computed using Fisher's exact test on the contingency table, as implemented in R [77].

Acknowledgments

The authors would like to thank Richard J. Edwards (University of Southampton) for help with SLiMFinder and SLiMSearch, Bernat Serra (IRB Barcelona) for help with visual inspection of intermediate results, and Roland A. Pache (IRB Barcelona) for helpful discussions. PA acknowledges the financial support received from the Spanish Ministerio de Innovación y Ciencia through the grants BIO2007-62426 and PSS-010000-2009, and the European Commission under FP7 Grant Agreement 223101(AntiPathoGN).

References

1. Neduva V, Russell RB (2006) Peptides mediating interaction networks: new leads at last. Curr Opin Biotechnol 17: 465-471.
2. Russell RB, Gibson TJ (2008) A careful disorderliness in the proteome: sites for interaction and targets for future therapies. FEBS Lett 582: 1271-1275.
3. Taussig MJ, Stoevesandt O, Borrebaeck CA, Bradbury AR, Cahill D, et al. (2007) ProteomeBinders: planning a European resource of affinity reagents for analysis of the human proteome. Nat Methods 4: 13-17.
4. Remenyi A, Good MC, Lim WA (2006) Docking interactions in protein kinase and phosphatase networks. Curr Opin Struct Biol 16: 676-685.
5. Mandell DJ, Kortemme T (2009) Computer-aided design of functional protein interactions. Nat Chem Biol 5: 797-807.
6. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631-636.
7. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, et al. (2003) A protein interaction map of Drosophila melanogaster. Science 302: 1727-1736.
8. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98: 4569-4574.
9. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. Nature 437: 1173-1178.
10. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122: 957-968.
11. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403: 623-627.
12. Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, et al. (2000) The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol 7 Suppl: 957-959.
13. Vassilev LT, Vu BT, Graves B, Carvajal D, Podlaski F, et al. (2004) In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. Science 303: 844-848.
14. Aloy P, Russell RB (2006) Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol 7: 188-197.
15. Stein A, Panjkovich A, Aloy P (2009) 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. Nucleic Acids Res 37: D300-304.
16. Stein A, Aloy P (2008) Contextual specificity in peptide-mediated protein interactions. PLoS One 3: e2524.
17. Humphris EL, Kortemme T (2007) Design of multi-specificity in protein interfaces. PLoS Comput Biol 3: e164.
18. Yun SM, Moulaei T, Lim D, Bang JK, Park JE, et al. (2009) Structural and functional analyses of minimal phosphopeptides targeting the polo-box domain of polo-like kinase 1. Nat Struct Mol Biol 16: 876-882.
19. Chen JR, Chang BH, Allen JE, Stiffler MA, MacBeath G (2008) Predicting PDZ domain-peptide interactions from primary sequences. Nat Biotechnol 26: 1041-1045.
20. Li SS (2005) Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction. Biochem J 390: 641-653.
21. Fuxreiter M, Tompa P, Simon I (2007) Local structural disorder imparts plasticity on linear motifs. Bioinformatics 23: 950-956.
22. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, et al. (2006) Analysis of molecular recognition features (MoRFs). J Mol Biol 362: 1043-1059.
23. Stein A, Pache RA, Bernado P, Pons M, Aloy P (2009) Dynamic interactions of proteins in complex networks: a more structured view. Febs J 276: 5390-5405.
24. Heery DM, Kalkhoven E, Hoare S, Parker MG (1997) A signature motif in transcriptional co-activators mediates binding to nuclear receptors. Nature 387: 733-736.
25. Remaut H, Waksman G (2006) Protein-protein interaction through beta-strand addition. Trends Biochem Sci 31: 436-444.
26. Pawson T, Linding R (2005) Synthetic modular systems--reverse engineering of signal transduction. FEBS Lett 579: 1808-1814.
27. Neduva V, Russell RB (2005) Linear motifs: evolutionary interaction switches. FEBS Lett 579: 3342-3345.
28. Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, et al. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol 3: e405.

29. Edwards RJ, Davey NE, Shields DC (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. PLoS One 2: e967.
30. Davey NE, Shields DC, Edwards RJ (2009) Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. Bioinformatics 25: 443-450.
31. Chica C, Diella F, Gibson TJ (2009) Evidence for the concerted evolution between short linear protein motifs and their flanking regions. PLoS One 4: e6052.
32. Stiffler MA, Chen JR, Grantcharova VP, Lei Y, Fuchs D, et al. (2007) PDZ domain binding selectivity is optimized across the mouse proteome. Science 317: 364-369.
33. Landgraf C, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, et al. (2004) Protein interaction networks by proteome peptide scanning. PLoS Biol 2: E14.
34. Zarrinpar A, Park SH, Lim WA (2003) Optimization of specificity in a cellular protein interaction network by negative selection. Nature 426: 676-680.
35. Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jorgensen C, et al. (2007) Systematic discovery of in vivo phosphorylation networks. Cell 129: 1415-1426.
36. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, et al. (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 31: 3625-3630.
37. Tonikian R, Zhang Y, Sazinsky SL, Currell B, Yeh JH, et al. (2008) A specificity map for the PDZ domain family. PLoS Biol 6: e239.
38. Parthasarathi L, Casey F, Stein A, Aloy P, Shields DC (2008) Approved drug mimics of short peptide ligands from protein interaction motifs. J Chem Inf Model 48: 1943-1948.
39. Wells JA, McClendon CL (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature 450: 1001-1009.
40. Petsalaki E, Russell RB (2008) Peptide-mediated interactions in biological systems: new discoveries and applications. Curr Opin Biotechnol 19: 344-350.
41. Brannetti B, Via A, Cestra G, Cesareni G, Helmer-Citterich M (2000) SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family. J Mol Biol 298: 313-328.
42. Brannetti B, Helmer-Citterich M (2003) iSPOT: A web tool to infer the interaction specificity of families of protein modules. Nucleic Acids Res 31: 3709-3711.
43. Diella F, Haslam N, Chica C, Budd A, Michael S, et al. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci 13: 6580-6603.
44. Xu W, Harrison SC, Eck MJ (1997) Three-dimensional structure of the tyrosine kinase c-Src. Nature 385: 595-602.
45. Aloy P, Russell RB (2004) Ten thousand interactions for the molecular biologist. Nat Biotechnol 22: 1317-1321.
46. Oliver AW, Swift S, Lord CJ, Ashworth A, Pearl LH (2009) Structural basis for recruitment of BRCA2 by PALB2. EMBO Rep 10: 990-996.
47. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, et al. (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36: D419-425.
48. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577-2637.
49. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The Pfam protein families database. Nucleic Acids Res 36: D281-288.
50. Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, et al. (2009) The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res 37: D310-314.
51. Aloy P, Ceulemans H, Stark A, Russell RB (2003) The relationship between sequence and interaction divergence in proteins. J Mol Biol 332: 989-998.
52. Kim HY, Ahn BY, Cho Y (2001) Structural basis for the inactivation of retinoblastoma tumor suppressor by SV40 large T antigen. Embo J 20: 295-304.
53. Lodowski DT, Pitcher JA, Capel WD, Lefkowitz RJ, Tesmer JJ (2003) Keeping G proteins at bay: a complex between G protein-coupled receptor kinase 2 and Gbetagamma. Science 300: 1256-1262.
54. Teyra J, Paszkowski-Rogacz M, Anders G, Pisabarro MT (2008) SCOWLP classification: structural comparison and analysis of protein binding regions. BMC Bioinformatics 9: 9.
55. Pawson T (2004) Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. Cell 116: 191-203.
56. Yaffe MB, Smerdon SJ (2004) The use of in vitro peptide-library screens in the analysis of phosphoserine/threonine-binding domain structure and function. Annu Rev Biophys Biomol Struct 33: 225-244.
57. Edwards RJ (2008) SLiMSearch.
58. Manke IA, Lowery DM, Nguyen A, Yaffe MB (2003) BRCT repeats as phosphopeptide-binding modules involved in protein targeting. Science 302: 636-639.

59. Galjart N (2005) CLIPs and CLASPs and cellular dynamics. Nat Rev Mol Cell Biol 6: 487-498.
60. Bouillet P, O'Reilly LA (2009) CD95, BIM and T cell homeostasis. Nat Rev Immunol 9: 514-519.
61. Cotter TG (2009) Apoptosis and cancer: the genesis of a research field. Nat Rev Cancer 9: 501-507.
62. Lessene G, Czabotar PE, Colman PM (2008) BCL-2 family antagonists for cancer therapy. Nat Rev Drug Discov 7: 989-1000.
63. Pawson T (2009) Protein Interaction Domains.
64. Wang J, Hartling JA, Flanagan JM (1997) The structure of ClpP at 2.3 A resolution suggests a model for ATP-dependent proteolysis. Cell 91: 447-456.
65. Chou PY, Fasman GD (1978) Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 47: 45-148.
66. Freund C, Kuhne R, Yang H, Park S, Reinherz EL, et al. (2002) Dynamic interaction of CD2 with the GYF and the SH3 domain of compartmentalized effector molecules. Embo J 21: 5985-5995.
67. Li C, Pazgier M, Liu M, Lu WY, Lu W (2009) Apamin as a Template for Structure-Based Rational Design of Potent Peptide Activators of p53. Angew Chem Int Ed Engl.
68. Russell RB, Aloy P (2008) Targeting and tinkering with interaction networks. Nat Chem Biol 4: 666-673.
69. Zanzoni A, Soler-Lopez M, Aloy P (2009) A network medicine approach to human disease. FEBS Lett 583: 1759-1765.
70. Joachims T (1999) Making large-Scale SVM Learning Practical. In: Schölkopf B, Burges, C, Smola, A, editor. Advances in Kernel Methods - Support Vector Learning: MIT Press.
71. Hubbard SJ, Thornton, J.M. (1993) NACCESS.
72. de Hoon MJ, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics 20: 1453-1454.
73. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443-453.
74. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, et al. (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res 35: D572-574.
75. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, et al. (2007) IntAct--open source resource for molecular interaction data. Nucleic Acids Res 35: D561-565.
76. Prasad TS, Kandasamy K, Pandey A (2009) Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. Methods Mol Biol 577: 67-79.
77. R-Development-Core-Team (2009) R: A Language and Environment for Statistical Computing. Vienna, Austria.

Table and Table legends

Table 1: *DMIs significantly enriched in the interactome cross-validation.*
List of domains and motif patterns determined in this study and found to be significantly enriched in the interactome cross-validation. The top-ranked SLIMFinder motif and the pattern with the highest significant enrichment are shown for each topological cluster (TC), along with the respective enrichment and p-values. If available, ELM ligands of the domain are given below the patterns derived in this work. A table with all DMI, including those not enriched in the interactome cross-validation as well as the 3 top-ranked patterns for all candidate DMIs is given in Supplementary Table 1.

Table 1

Domain	TC	top-ranked pattern	Enrichment	p-value	most enriched pattern	Enrichment	p-value
14-3-3	0	[HR]S.P	1.31	0	SHSY	3.63	0.001
14-3-3	2	LDL	1.19	0.007	LD.{0,1}L	1.22	0
	R[SFYW].S.P (LIG_14-3-3_1), R.[SYFWTQAD].[ST].[PLM] (LIG_14-3-3_2), [RHK][STALV].[ST].[PESRDIF] (LIG_14-3-3_3)						
Arm	0	KK[KR]K	1.45	0.043	KKRK	2.25	0.001
Arm	1	KKRKV	2.13	0.383	K[KR].K[LV][DE]	3.04	0.001
Arm	2	KK[KR]K	1.45	0.043	KKRK	2.25	0.001
Asp	2	HPFH	0	1	VV.A	2.58	0
Bcl-2	2	QL..I[AG]D	0	1	R.I[AG]D.[LV]	13.56	0.009
BIR	0	AVP[FI]	8.87	0.107	[IV].[FY][FY].P	25.16	0
Borealin	0	L.EFL	33.95	0	L.EFL	33.95	0
BRCT	0	S..FP.A	1.2	0.501	D..QVF.F	23.55	0.002
BRCT	1	SPTF	2.14	0.116	S.TF	1.56	0.01
Bromodomain	2	[GS].GG	1.36	0	GKG.{0,1}GK	8.8	0
Chromo	2	ARK[ST]	2.27	0.07	T.{0,2}ARKS	9.62	0.003
Cullin	0	WN.V..W.W	86.76	0	W..V..W..DI	86.76	0.012
Cyclin_N	0	K.{0,1}RRL	1.13	0.458	KR.L..E	3.01	0.003
	[RK].L.{0,1}[FYLIVMP] (LIG_CYCLIN_1)						
DUF618	0	PSYSP	0	1	PSY.P	56.58	0.001
Dynein_light	0	K.TQT	6.2	0.15	TQT	4.11	0
	[KR].TQT (LIG_Dynein_DLC8_1)						
FHA	0	EVTE.D	25.86	0.039	LE.TE	5.32	0
	T..[ILA] (LIG_FHA_1)						
Fibrinogen_C	0	HRP	2.37	0.085	GPR	2.25	0.014
Filamin	0	[KR]S[AS]	1.2	0.066	[ST]..[ST][ST]	1.17	0.021
Focal_AT	0	R.L.E	1.56	0.022	LSE	1.81	0.002
	[LV][DE].[LM][LM]..L (LIG_PXL)						
Histone	0	Q.RT.Y.F	0	1	QG.TL.G	40.89	0.001
Hormone_recep	2	[IL]L[HR].LL	0.7	0.806	[IL]L[HR].L	1.36	0.02
	L[^P]{2}[HI]I[^P]{2}[IAV][IL] (LIG_CORNRBOX), [^P](L)[^P][^P](L)(L)[^P] (LIG_NRBOX)						
IF4E	0	YDR.FL	156.16	0	YDR.FL	156.16	0
IRS	0	[FI]..[KR].[FY]	1.48	0.003	[FI]..[KR].[FY]	1.48	0.003
MAP1_LC3	0	D.WTH.S	108.44	0.009	D..THLS	108.44	0.009
MBT	0	HRK..RD	56.58	0.018	RKV.RD	339.48	0.003
PCNA_C	0	K.{0,2}QATL	34.22	0.029	K.{0,2}Q.T	1.6	0.021
PCNA_N	1	K.{0,2}QATL	34.22	0.029	K.{0,2}Q.T	1.6	0.021
	(^.{0,3}|Q).[^FHWY][ILM][^P][^FHILVWYP][DHFM][FMY].. (LIG_PCNA)						
PDZ	0	RETQV	0	1	R.ET.V	2.89	0
	.[ST].[VIL]$ (LIG_PDZ_1), .[VYF].[VIL]$ (LIG_PDZ_2), .[DE].[IVL] (LIG_PDZ_3)						

Peptidase_C14	1	D.SD	1.36	0.028	DE.D	2.18	0
Peptidase_C14	4	DE.D	2.18	0	DEVD	4.1	0
PHD	0	RTKQT	11.52	0.011	A.TK..AR	17.29	0
PID	2	Y.NP.YK	0	1	GY.N.TY	68.15	0.015
Pkinase	3	RRRHP	1.02	0.675	RR.HPS	4.08	0.015
Pkinase	6	T.NL	1.14	0.024	T.NL	1.14	0.024
Pkinase_Tyr	2	EIF..FE	0	1	E.FG..E	2.43	0.021
Profilin	0	PPP.{0,1}PP	4.48	0	PPP..P.P	8.6	0
Proteasome_A_N	0	K.EDN.G	0	1	KEE..L	3.1	0.007
RNA_pol_L	0	T.R..QF..R	32.47	0.031	R.VQF.A	21.64	0.003
SET	0	AR.{0,1}K.T	2.28	0.068	ARKST	20.55	0
SH2	0	YVNV	3.85	0	HIYDE	8.18	0.015
SH2	1	S.TIYA	4.09	0.229	IY.QVQ	8.18	0.015
	Y.N. (LIG_SH2_GRB2), Y[IV].[VILP] (LIG_SH2_PTP2), Y[QDEVAIL][DENPYHI][IPVGAHS] (LIG_SH2_SRC), Y..Q (LIG_SH2_STAT3), Y[VLTFIC].. (LIG_SH2_STAT5)						
SH3_1	0	P.{0,1}P.{0,2}P.{0,2}P	2.2	0	P.PV.{0,1}PP	9.22	0.019
SH3_1	1	P.{0,1}P..P	1.8	0	P.{0,1}PP.{1,2}P	3	0
SH3_1	2	DR.TKP	1.72	0.468	DR.T	1.44	0
	[RKY]..P..P (LIG_SH3_1), P..P.[KR] (LIG_SH3_2), ...[PV]..P (LIG_SH3_3)						
SIR2	0	HKKLM	0	1	[KR][HR].[KR]	1.46	0.008
TPR_1	0	EEVD	2.01	0.027	ME.VD	4.02	0.007
	EEVD$ (LIG_TPR)						
TRF	0	[FY].L.P[LV]	2.67	0.172	FN.A..GR	244	0.004
Trypsin	1	PG.Y	1.87	0.001	PG.Y	1.87	0.001
Trypsin	3	CGK	0.66	0.925	CG..T	1.86	0.015
Trypsin	5	PAIQP	0	1	P.IQ	1.59	0.018
Trypsin	7	CT..IPP	0	1	CT..I.P	8	0.024
Trypsin	11	CG.[KR]	1.24	0.189	CG..T	1.86	0.015
Trypsin	12	[FY]E.IP.E	0	1	DF..IP.{0,1}E	14.39	0.007
Tyr-DNA_phospho	0	KLNY	177.45	0.006	KLNY	177.45	0.006
V-set	3	Q.DPAF	15.93	0.062	K..[HK].G	1.38	0.008
V-set	10	E.DKW	5.31	0.053	A.FRHD	15.93	0.006
V-set	11	WF..T..LW	0	1	QE..D..RE	10.62	0.014
V-set	15	D.PDY.S	0	1	P.Y.S	1.79	0.001
V-set	16	E.DKW	5.31	0.053	L.FGYP	31.87	0.001
VHS	0	D..LL	1.12	0.314	DL..I	1.72	0.012
WD40	0	RTKQT	6.71	0.006	TKQTA	8.39	0.003
	F.[IV][^WFY][^WFY][IL][ILM] (LIG_EH1)						
WW	0	P.{0,2}PP.{0,2}P	1.64	0	PPPY	6.24	0
PP.Y (LIG_WW_1), PPLP (LIG_WW_2), ...[ST]P. (LIG_WW_4)							

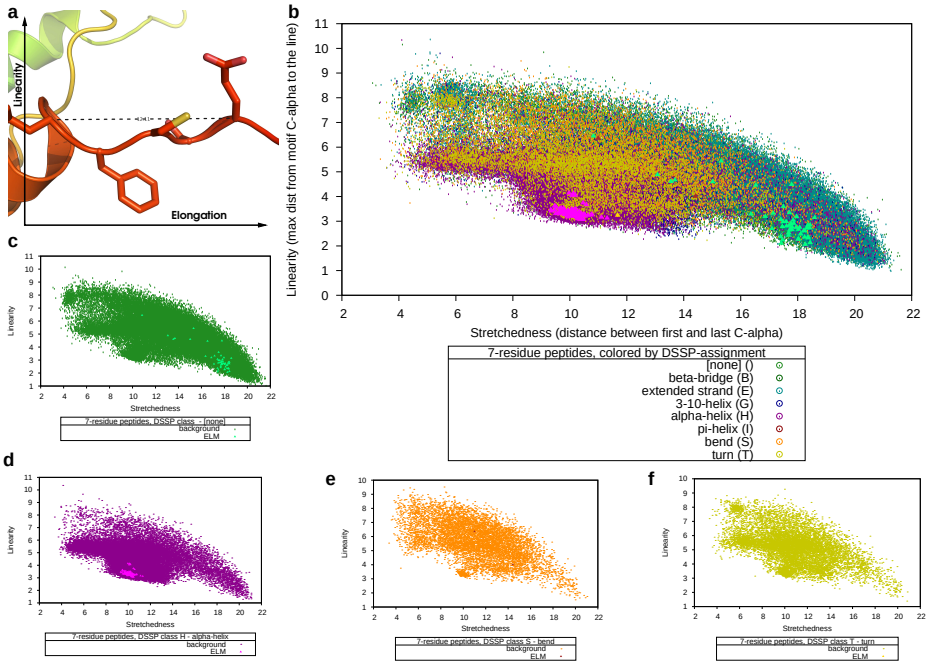

Figure 3.5: Novel DMIs manuscript – Figure 1: Linearity and elongation of linear motifs. (a) The Retinoblastoma-associated protein B domain (RB_B)-binding peptide shows the typical linear and elongated form found in 3D structures of many motifs (PDB ID 1gh6). The concepts of linearity (the maximum deviation of any C_α in the motif from the line through the first and last C_α) and elongation (the distance between the first and last C_α of a motif) are illustrated in this structure. (b) A slice of the data used for SVM training: linearity, elongation and secondary structure classification for 7-residue-peptides, with data from the SCOP background shown as dots and the data for known DMI shown as solid triangles, using one colour per DSSP classification. Panels (c) to (f) show the distribution of linearity:elongation values for those secondary structure classifications for which we had known 7-residue-peptides (none, alpha-helix, bend, and turn). These data slices illustrate how known linear motifs fall into distinct regions of the parameter space.

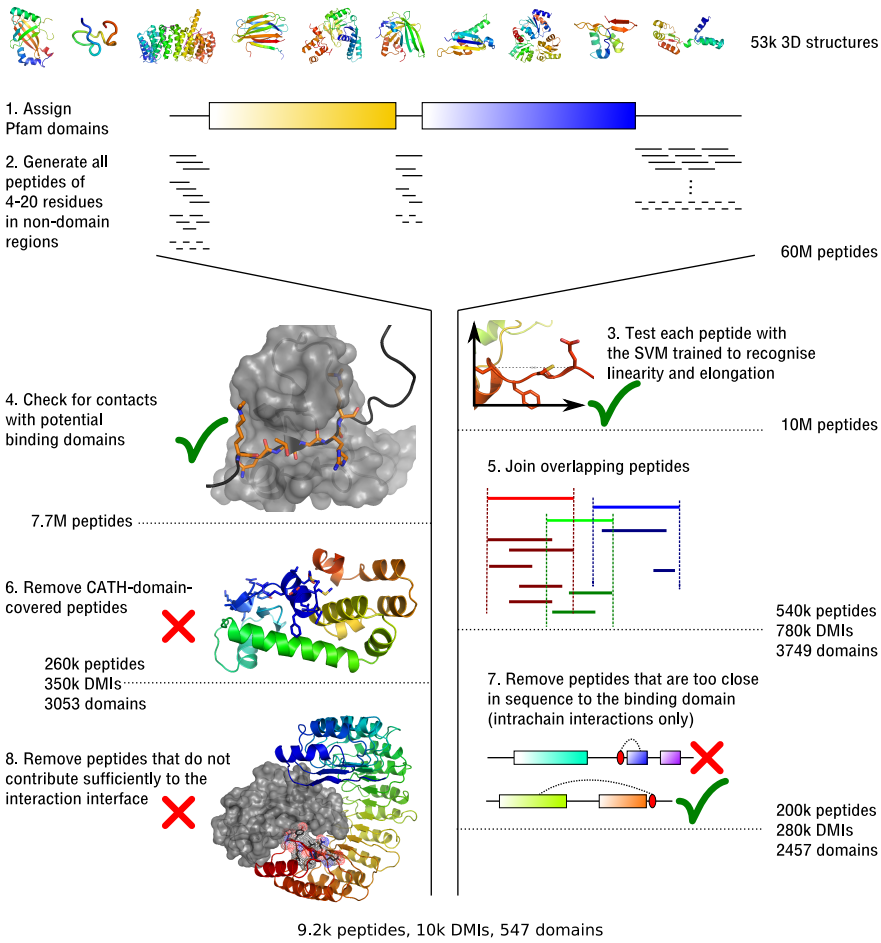

Figure 3.6: Novel DMIs manuscript – Figure 2: Overview of the generation and filtering of motif-like peptides. (1, 2) We generated all possible peptides of 4-20 residues from regions of 3D structures that did not match Pfam domains. (3, 4) For peptides accepted by the SVM trained on linearity and elongation (cf. Figure 1) we computed whether there were sufficient contacts with domains in the same structure, which may be in the same or in another protein chain. (5) Peptides that are completely covered by other (longer) peptides are removed, so that the largest accepted peptide represents shorter candidates binding to the same region. (6) Peptides in intrachain interactions that are part of CATH domains are often artifacts of differences between structure- and sequence-based domain assignment and are therefore excluded. (7) Peptides in intrachain interactions that are sequentially directly next to the binding domain are often artifacts and thus removed, though in general peptides close to domains are allowed, as long as they have a sufficient sequential distance from their binding domain. (8) Exclude candidate DMI in which the interface is smaller than 150 $\AA$, or in which the interface between domain and peptide is less than 50% of the total interface between the proteins.

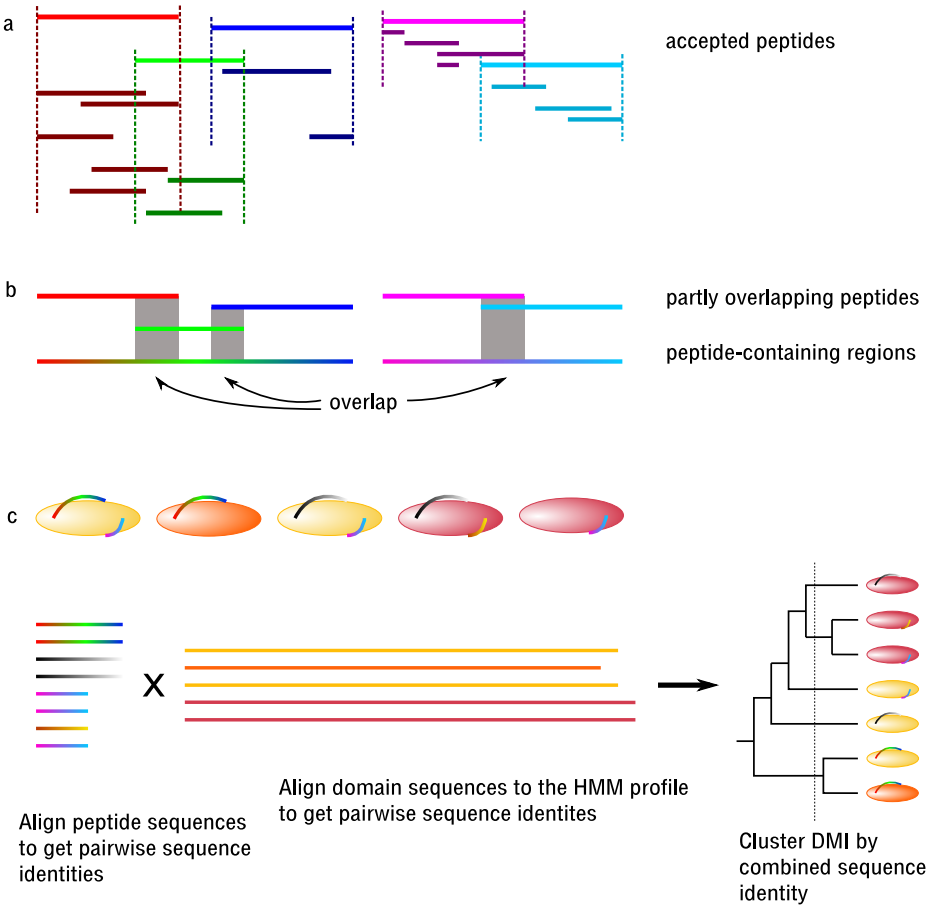

Figure 3.7: Novel DMIs manuscript – Figure 3: Joining of partially overlapping peptides for sequence-based clustering. (a) Partially overlapping peptides cannot be represented by either one, as both may contribute to an interface in ways not covered by the other. Yet to improve the quality of peptide alignments, and to ensure that motif matches in the overlapping regions (shown in gray) are only counted once for motif support, we need to create a construct that holds unique, non-overlapping regions of one or more peptides accepted by the SVM and having a sufficient interface with a domain. (b) Thus, for each continuous stretch of a protein that is covered by one or more peptides, we built a peptide-containing region. (c) These regions are then aligned to generate non-redundant sets of peptides binding to a given domain, and each motif match in a peptide-containing region only qualifies for motif support once. The 90% sequence clustering of the DMIs is computed from a combination of the sequence identities of peptide-containing regions and those of the binding domains.

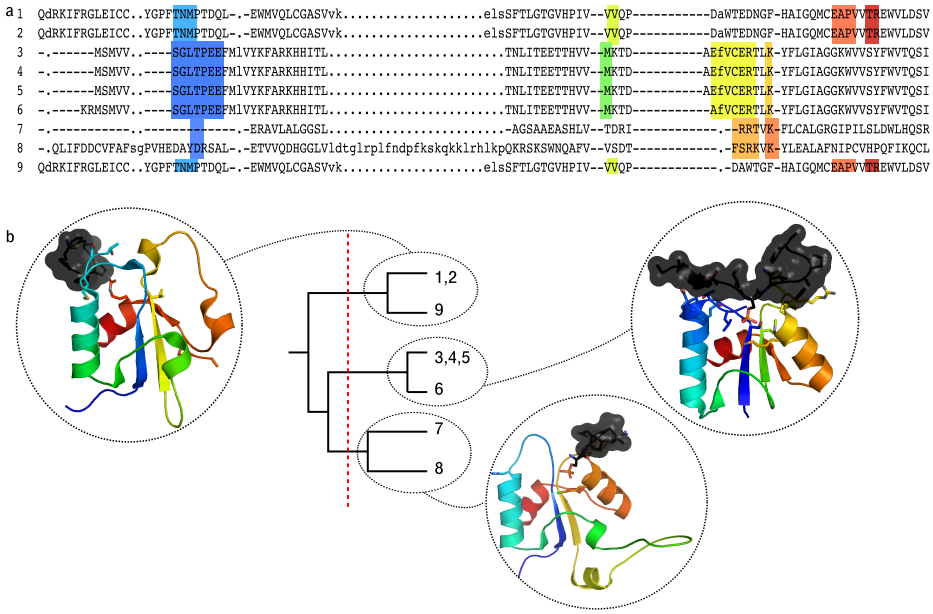

Figure 3.8: Novel DMIs manuscript – Figure 4: Topological clustering of peptide-mediated interactions. (a) Alignment of BRCA1 C Terminus (BRCT) sequences to the domain's HMM profile; interface residues are highlighted. The colour corresponds to the "rainbow" colouring scheme used for the domain visualisation in (b). Lowercase letters refer to amino acids that do not match the domain's profile, - to positions in the profile that do not occur in the given sequence. (b) Clustering of the interaction topologies, based on shared interface residues. Domains with the same or highly similar topologies are grouped together. In the structural representation, all three BRCT domains have the same orientation. Note that the BRCT domain usually forms dimers that bind the peptide, using the interfaces from clusters (3,4,5,6) and (1,2,9), respectively (cf. Figure 5).

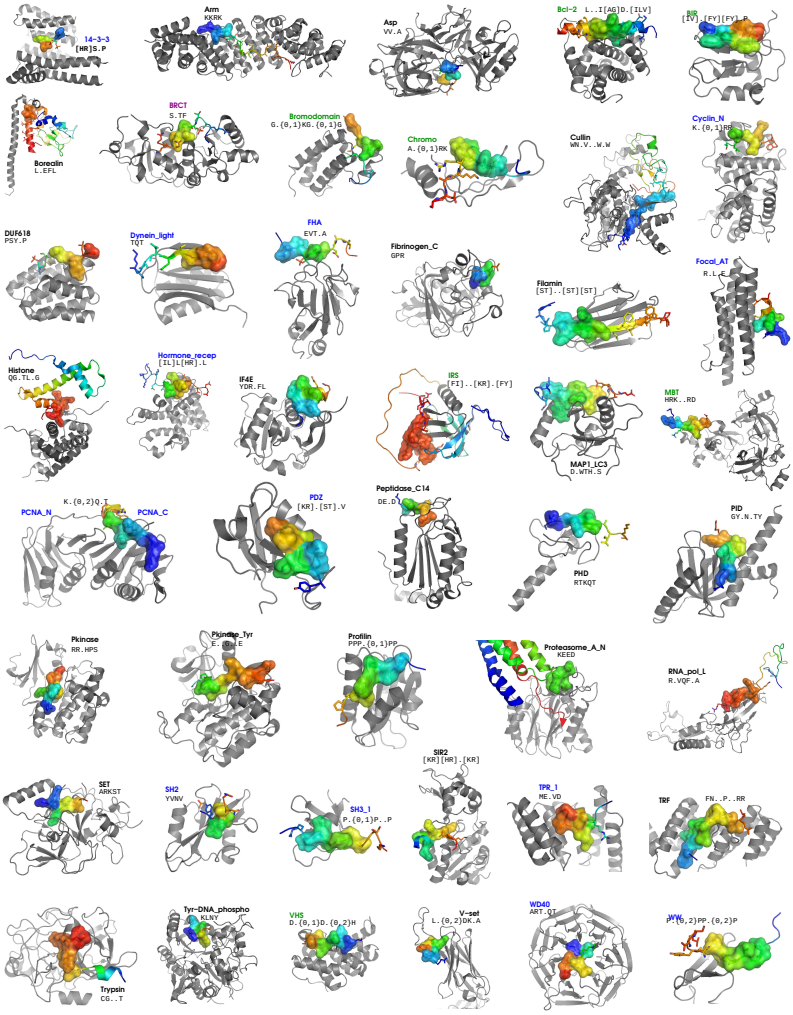

Figure 3.9: Novel DMIs manuscript – Figure 5: DMIs significantly enriched in the interactomes. Significantly enriched motifs were found for 46 distinct domains (shown in gray; PCNA_N and PCNA_C are shown in the same structure). Binding peptides are given in a rainbow colour scheme, with the SVM-accepted part in sticks representation and the consensus motif in surface representation. In most cases, differences between the interaction types for a given domain are subtle, thus only one is shown in this representative figure. However, for domains that form repeats to bind peptides (Arm, BRCT (cf. Fig. 4 and main text), TPR_1, TRF, WD40), we have visualized all domains required to bind one peptide; these usually employ different interaction types. Blue domain names indicate those that were described in the ELM training dataset (Stein and Aloy, 2008), violet names mark additions to ELM since 2007 (Gould et al., 2010), which were not in our training set, and green names indicate DMIs that are described on the Pawson lab web site (Pawson, 2010) but not in ELM.

3.2.a) Supplementary Material

Supplementary Table 1

Due to its excessive length, the table showing all DMI candidates derived from high-resolution 3D structures, specifying the binding domain, topology or interaction type ID, the consensus motif, enrichment and p-value in the interactome cross-validation (if applicable) and the ELM pattern and name (if available), with up to 3 top-ranked motifs per interaction type is not printed here. It is available at `www-aloy.irbbarcelona.org/publications/supplementary/novel_DMIs_suppl.xls`.

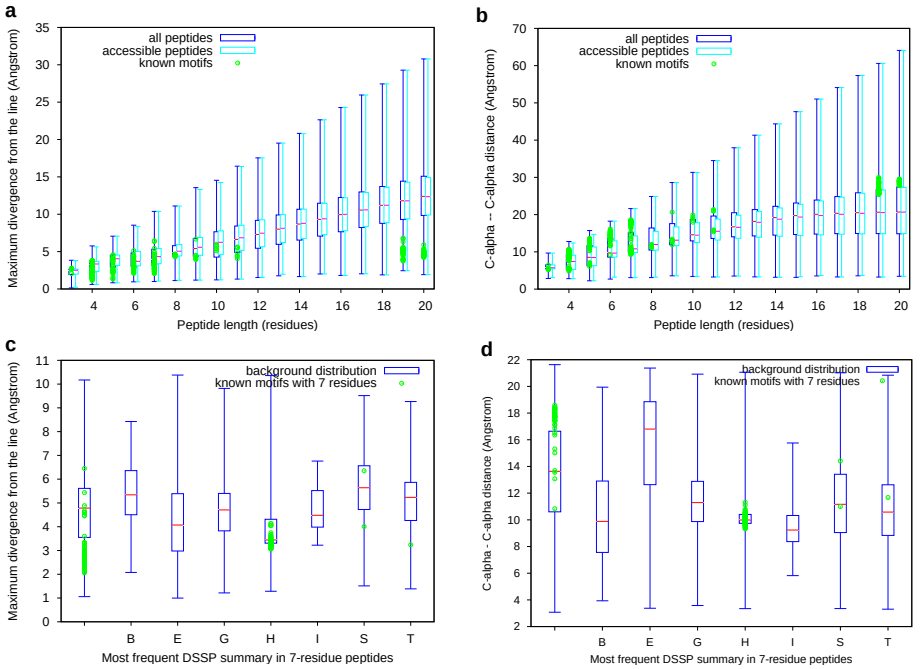

Figure 3.10: Novel DMIs manuscript – supplementary figure 1: Linearity and Elongation of known motifs in comparison to a background sampling. (a) Linearity, (b) Elongation, (c) Linearity for 7-residue-peptides, split by DSSP-classification, (d) Elongation for 7-residue-peptides, split by DSSP-classification. The distribution of values varies greatly across and within classes of secondary structure.

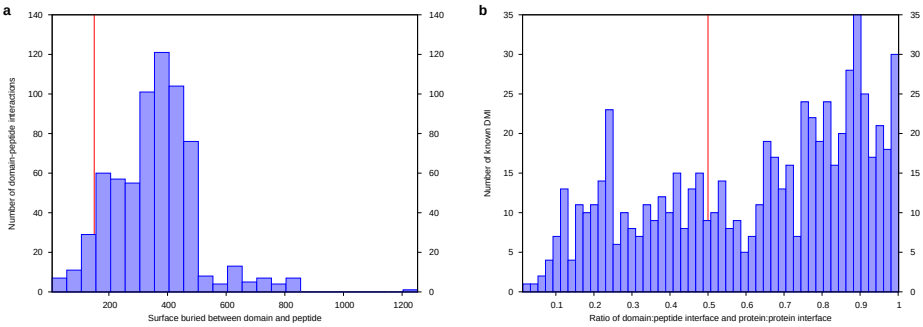

Figure 3.11: Novel DMIs manuscript – supplementary figure 2: Interface size and ratio. (a) Interface size for known domain-peptide interactions. (b) Ratio of the interface between domain and peptide to the full protein-protein interface for the known DMI. Both are computed as described in the Methods section of the main manuscript.

3.3 Uncovering novel targets for Aurora A kinase

A special group among motif-binding domains are kinases, which recognise specific patterns and phosphorylate a serine, threonine or tyrosine residue in them, which may lead to a change in the activity, structure and/or localization of the substrate. It has been shown that contextual information is particularly important for substrate specificity in phosphorylation networks (Linding et al., 2007), in part because the consensus motifs are often degenerate, with only one or two fixed positions in some cases, thus matching a plethora of proteins most of which are likely not biological substrates of the kinase. Aurora kinases are a family of closely related serine/threonine protein kinases involved in the regulation of mitosis. Each of the three Aurora kinases (A, B & C) has different roles, which are tightly coupled to their characteristic localization at critical cell-cycle times (Vader and Lens, 2008). Aurora A participates in cell cycle progression and has a well-characterized role in centrosome maturation (Barr and Gergely, 2007) and spindle assembly (Sardon et al., 2008), while Aurora B localizes to the nucleus in interphase and to the kinetochores and the spindle midzone in anaphase (see Fig. 3.12, page 109). For the human Aurora A kinase, less than 20 phosphorylation targets are known. The consensus pattern has been derived in the orthologous yeast kinase as `[KR].[ST][IVL]` (Cheeseman et al., 2002). Given that this pattern is degenerate and very short, searching for it in all human proteins is not likely to predict potential substrates well. Furthermore, several of the human substrates have other residues than `[IVL]` in the $+1$ position (directly C-terminal to the phosphorylation site), which requires an even more lenient pattern. By integrating various types of contextual information, we can greatly increase the chances of successful substrate prediction. Sub-cellular localization reduces the candidate substrate set from over 75,000 human proteins to merely 300 spindle proteins. The localization of a potential substrate to the kinetochore may indicate that it is phosphorylated by Aurora B rather than Aurora A – the consensus motifs of the two kinases are highly similar and do not allow separation. As no 3D structure of Aurora and a substrate at the active site is available, we cannot use the kind of molecular contextual information described in Stein and Aloy (2008). However, we did incorporate information from 3D structures when searching for candidate sites in substrates: we only consider motif matches in regions that are either outside globular domains,

or on their surface, because regions in the core of a domain are unlikely to be accessible for phosphorylation. We scored proteins with occurrences of the consensus motif by whether interaction with Aurora A has been observed, if the motif site was found phosphorylated in mass spectrometry experiments, and whether the site was conserved in other vertebrates. The more of these criteria apply, the higher a protein is ranked, thus creating a list of likely physiological substrates. Using this strategy, we have generated a ranked list of 90 potential Aurora substrates in human, of which 76 are novel. Experimental validation on a randomly selected group of candidates, using *in vitro* kinase assays and mass spectrometry analyses, suggests a prediction accuracy of about 80%. We estimate that our approach can be readily applied to more that 30 human kinases, offering an efficient possibility to identify new substrates of a kinase, to deepen the understanding of its function, but also to possibly suggest unexpected roles.

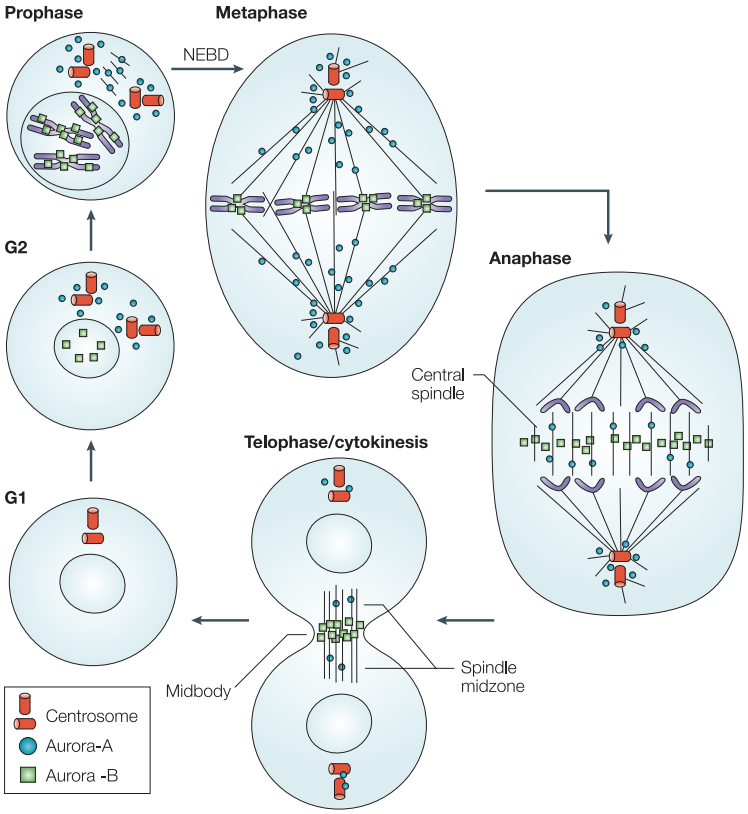

Figure 3.12: Localization of Aurora A and B during the cell cycle (taken from Marumoto et al. (2005)). Aurora A (blue circles) is first found at centrosomes (red cylinders), in metaphase also at spindle poles. Aurora B (green squares) is found on chromosome arms, then localizes to the inner centromere region in metaphase and subsequently to the central spindle in anaphase and telophase. NEBD, nuclear envelope breakdown.

Uncovering novel substrates for Aurora A kinase

Teresa Sardon[1,*], Roland A. Pache[2,*], Amelie Stein[2,*], Henrik Molina[1], Isabelle Vernos[1,3,†] and Patrick Aloy[2,3,†]

1. Centre for Genomic Regulation (CRG) and Universitat Pompeu Fabra (UPF). c/ Dr Aiguader 88, 08003 Barcelona, Spain.

2. Institute for Research in Biomedicine (IRB) and Barcelona Supercomputing Center (BSC). c/ Baldiri i Reixac 10-12, 08028 Barcelona, Spain.

3. Institució Catalana de Recerca i Estudis Avançats (ICREA)

*These authors contributed equally to this work.

†Corresponding authors:
Isabelle Vernos. Tel: +34 933160275; Email: isabelle.vernos@crg.es
Patrick Aloy. Tel: +34 934039690; Email: patrick.aloy@irbbarcelona.org

Running title: Novel substrates for Aurora A

Abstract

Aurora A is a serine/threonine kinase essential for cell cycle progression, centrosome maturation and spindle assembly. Although the participation of Aurora A in these events is well-established, its mechanism of action is poorly understood in most cases. Moreover, the relatively small number of known substrates for this kinase does not account for its many different roles. In this work, we present and validate a novel strategy to identify Aurora A substrates, along with their specific phosphorylation sites. We have developed a computational approach that integrates distinct types of biological information to generate a ranked list of 90 potential Aurora substrates of which 76 are novel. Experimental validation on a randomly selected group of candidates, using *in vitro* kinase assays and mass spectrometry analyses, suggests a prediction accuracy of about 80%. Our results open the way to getting a better understanding of Aurora A function during cell division and suggest novel unexpected roles for the Aurora kinase family. We estimate that our approach can be readily applied to more that 30 human kinases.

Keywords: Aurora substrates / pattern discovery / phosphorylation motifs

Background

Aurora kinases are a family of closely related serine/threonine protein kinases that includes three different members in metazoans (A, B & C). All of them play important roles in cell division and genome stability maintenance, with distinct functions for each family member that are tightly coupled to their characteristic localization at critical cell-cycle times [1]. Aurora A (AurA) is a centrosomal kinase that participates in cell cycle progression and has a well-characterized role in centrosome maturation during late G2 and prophase [2] and in spindle assembly [3]. Aurora B (AurB) localizes to the nucleus in interphase and to the kinetochores and the spindle midzone in anaphase. It is required for chromosome bi-orientation and cytokinesis [1]. Finally, the third member of the family, Aurora C (AurC), performs its function more specifically during meiosis, as its expression appears to be restricted to the mammalian testis. Although the study of Aurora C has been lagging behind, the

current data suggest that its function is closely related to that of Aurora B in cytokinesis [4].

The three Aurora kinases have a similar protein domain organization with a conserved catalytic domain and divergent N-terminal and C-terminal domains. The high sequence similarity of the catalytic domains has important consequences on the specificity of the different Aurora family members towards their substrates. Indeed, AurA and B are known to share several substrates *in vitro* (e.g., HH3 and MBP), and they both phosphorylate the widely accepted consensus motif that was first determined for Ipl1, the only Aurora kinase in the budding yeast *S. cerevisiae* [5]. However, just a few amino acid changes in their catalytic domains determine the specificity of interaction of each kinase with its partners, and thereby define their localization and function. Indeed, AurA is targeted to the spindle microtubules through its specific interaction with TPX2 [6, 7], whereas AurB is targeted to the kinetochores and the spindle midzone through its interaction with INCENP [8]. All these data suggest that the substrate selectivity of AurA and B relies more on their interactions with specific partners and on their localization rather than on a direct discrimination of the substrates themselves. Recently, Fu *et al.* (2009) showed that introducing a point mutation in AurA to change its specificity of interaction from TPX2 to INCENP also changed its localization, directing it to the centromeres and the spindle midzone like AurB. Interestingly, the mistargeted kinase was able to rescue the cellular phenotype resulting from AurB knockdown, suggesting that localization is the major determinant for the specificity of substrate selection by the two Aurora family members [9].

A growing interest in the Aurora kinases has been fueled over the last years by multiple studies suggesting links with different types of cancer [10, 11]. In human, the gene for AurA maps to the chromosomal region 20q13, which is often amplified in primary tumors and cancer cell lines [12]. Moreover, AurA and B are overexpressed in several tumor types, and their overexpression is associated with poor prognosis [13-16]. It has also been shown that AurA exogenous overexpression promotes the tumorigenic transformation of human and rodent cells both *in vitro* and *in vivo* [12, 17]. Interestingly, AurA amplification has been reported to induce resistance to taxol, a cytotoxic drug commonly used in cancer therapy [18, 19]. The great hopes placed

in personalized anti-cancer therapies have promoted a strong interest in the development of small-molecule inhibitors against the Aurora kinases [20-22]. Indeed, the promising tumor growth inhibitory activity achieved by some of these compounds in nude mice has strengthened the idea that they could lead to novel therapeutical approaches in anticancer therapy, either on their own or in combination with other types of drugs [23, 24]. Notably, some of them are currently being tested in clinical trials [25].

The potential of Aurora kinase inhibitors as putative therapeutic drugs highlights the importance of determining all the pathways in which these kinases may participate, to eventually know which and how cellular processes may be affected by their inhibition. Although it is now clearly established that Aurora kinases are involved in several mitotic events, their mechanism of action is not well understood in most cases. In addition, the number of known substrates for these kinases is still relatively small and, although it has noticeably increased in the last years, it can still not account for all the many different roles attributed to them. Furthermore, recent studies suggest that Aurora kinases may also have roles during interphase and in non-dividing cells [26, 27]. The identification of Aurora substrates is therefore paramount to contextualize their functions in a global manner and to predict the effects of their inhibition in different cell types and whole organisms.

A growing number of methods to identify substrates of protein kinases are currently available. The most commonly used approaches are based on the identification of kinase-substrate interactions by co-precipitation from cell lysates, followed by mass-spectrometry-based determination of the phosphorylated sequences [28, 29]. However, this method has the limitation of a low sensitivity due to the very transient nature of the phosphorylation reaction. More recently, the development of *in vitro* phosphorylation-based screening platforms using purified components has allowed the high-throughput analysis of substrate phosphorylation by selected kinases [30]. Although these types of high-throughput screenings may inherently give a high number of false negative and false positive results, they can undoubtedly be useful for the discovery of novel substrates. Another approach that is currently used combines the characterization of protein phosphorylation by mass spectrometry with the use of specific kinase inhibitors (Mitocheck, http://www.mitocheck.org/). The

success of this approach relies strongly on the specificity of the inhibitors and the sensitivity of the phospho-mass spectrometry methodology. In addition, bioinformatics strategies have been developed for the *de novo* discovery of functional motifs by analysis of sets of sequences sharing some common property like the interaction with a common partner [31], the same subcellular localisation or a particular post-translational modification, including the presence of similar phosphorylation sites [32, 33]. These methods exploit the convergent evolution of linear motifs by looking for patterns that are over-represented among unrelated sequences in the query set. However, without the help of other biological data to contextualize the motifs, computational methods suffer from the fact that functional motifs are often very short (typically 4-10 residues) and degenerate (i.e., they only contain a few fixed positions), and can therefore be found in virtually every single protein, potentially leading to many false positive predictions [34]. Indeed, the importance of incorporating contextual information when predicting kinase-substrate relationships has been demonstrated by Linding *et al.* [35]. Recently, a combined bioinformatics and experimental approach has been successfully used to predict targets of the Polo-like kinase in yeast [36], while another hybrid approach identified PKA substrates in this model organism [37].

In this work, we present and validate a novel strategy to identify new substrates of Aurora A kinase. By analysing the available data on known Aurora A substrates and their phosphorylation sites, we have developed a computational approach that profits from distinct types of biological information, such as subcellular localization, interaction partners, high-resolution three-dimensional structures, evolutionary conservation of functional motifs and/or *in vivo* phosphorylation data, to generate a ranked list of potential Aurora substrates. We then validated our predictions experimentally on a randomly selected group of candidates using *in vitro* kinase assays and mass spectrometry analyses. Our results not only provide a mean to better understand the function of the Aurora kinases during cell division but also suggest that they may have additional novel unexpected functions in other specific processes.

Results and Discussion

Definition of the Aurora A consensus motif

The consensus motif of the Aurora kinase was first characterised in the budding yeast *Saccharomyces cerevisiae* by Cheeseman *et al.* [5] as [KR].[ST][ILV]. This motif is now widely accepted [2] and we will refer to it as the '*yeast motif*'. More recently, two slightly different patterns were then proposed for the human AurA kinase: [KNR]R.[ST][AFILMV] [38] and R.[ST][ILV] [39], the latter being a more restricted form of the yeast motif.

To check the general validity of these motifs, we conducted a literature search to gather all the known human substrates of the Aurora A kinase. Many of the sites were compiled in the Phospho.ELM database [40], but we also extended our search looking in the literature for specific studies reporting Aurora sites [38, 39, 41-47]. Altogether, we gathered a total of 18 phosphorylation sites and one autophosphorylation site in 17 different proteins.

We then examined whether these sites conformed to the consensus patterns and found that the known patterns had a relatively low prediction potential with only 8 out of the 19 known phosphorylation sites following the yeast pattern, and only 5 and 7 of them following the two human motifs, respectively. The first step in our study was therefore an attempt to broaden the coverage of the pattern to reduce the possibility of missing putative candidate substrates due to an excessively stringent motif definition. We aligned the site-containing regions from our list for the assessment of sequence conservation to try to derive a broader recognition pattern (see *Materials and Methods*). All three previously proposed patterns included the autophosphorylation site RTTL in Aurora A and required a hydrophobic residue in the +1 position (the residue directly C-terminal to the phosphorylation site). However, since many known substrates had other residues in that position and therefore did not fulfil this requirement, we defined a more lenient pattern, [KR].[ST][^P], which is a generalisation of the yeast motif with any residue but proline allowed at the +1 position (we will refer to it as the '*notP motif*'). This restriction was motivated by a report from Ferrari *et al.*, showing that proline at +1 abolishes phosphorylation by Aurora A [38]. The notP motif matched 13 of the 19 sites in our list of known Aurora A phosphorylation sites, including the autophosphorylation site and all those

conforming to the patterns described above. The remaining 6 sites could not fit the pattern unless it was made even more lenient. Since this would certainly increase the number of false positives significantly [48], we decided to use the notP motif for further analysis. When investigating the presence of the notP motif in the human proteome, we found that it matched 432,312 sites in 68,115 out of 77,683 non-redundant protein sequences extracted from public databases (see *Materials and Methods*). As many of these hits were most likely false positives, we focused on the development of a filtering strategy to increase the specificity of our analysis. In the following sections, we describe how we narrowed down the list of biologically functional substrates of Aurora A by integrating several different types of contextual information, including subcellular localisation, *in vivo* phosphorylation data, interactions with the kinase, as well as site accessibility and motif conservation across vertebrate species (Figure 1). It is worth noting that the consensus phosphorylation motif for Aurora A also matches most of the known substrates of Aurora B (15 out of 18 sites compiled in the Phospho.ELM database). This is not surprising since this motif is an extension of the yeast pattern, and Ipl1, the only Aurora kinase in this organism, shares functions and substrates with both Aurora A and B. In fact, as mentioned in the introduction, recently published data suggest that the substrate specificity of these two kinases is primarily a consequence of their different subcellular localizations and/or interaction with different cofactors [9, 49].

Collecting the list of spindle proteins

Aurora A specifically localizes to the centrosomes, spindle poles and spindle microtubules [50-52], and its localized kinase activity is essential for cell cycle progression, centrosome maturation and separation, as well as for bipolar spindle assembly [2, 3, 51, 52]. We therefore decided to restrict our analysis to the subset of proteins with known or confirmed spindle and/or centrosomal localization that were identified in the large-scale proteomic analysis of the mitotic spindle and centrosomes in human cells [53-55]. In addition, we also included proteins annotated as localising to the spindle in the GeneOntology (GO) database [56] (Figure 1). This group of proteins also includes some localising to the kinetochores, which could well be substrates of AurB. By reducing the number of potential target substrates of Aurora from the 77,683 human proteins to only 308, we restricted our analysis to a specific biological context (the spindle and the centrosomes) relevant for the Aurora

kinase, thereby greatly reducing the possibility of getting false positive hits. However it is important to keep in mind that by doing this we may also eliminate the possibility of identifying some real candidate substrates because they were not found in these proteomic analyses. Calculating the fraction of known substrates that were not present in our subset of centrosome/spindle proteins, we can infer that approximately one third of all the Aurora substrate proteins are likely to be present in our spindle/centrosome set (see *Supplementary Information*). To evaluate the sensibility of our approach, we investigated wether there was an increase in the number of proteins containing the different consensus motifs and in the number of motif hits found per protein in the subset of 308 spindle/centrosome proteins. We found a more than two-fold significant enrichment for these two parameters in our subset, indicating that, indeed, we were considerably increasing the specificity of our predictions (see *Supplementary Information*).

Checking structural accessibility of phosphorylation sites

Domains are well-structured, independently folding building blocks of proteins. They are often excluded in the search for small motifs, such as phosphorylation sites, because large parts of domains are buried and thus inaccessible to other proteins, and because motifs are frequently found in hinge and loop regions outside of domains [31, 57-59]. Consequently, we found it is reasonable to exclude protein regions that are highly unlikely to contain relevant sites, although some well-known phosphorylation sites do lie inside domains, such as the autophosphorylation site of AurA itself. However, this site is on the surface of the protein domain and thus accessible for phosphorylation. Therefore, completely excluding protein domains from our analysis would remove candidate sites that may turn out to be real sites, such as the AurA autophosphorylation site from our final candidate list. Furthermore, Gnad *et al.* have shown that the accessibility of phosphorylation sites is significantly higher than that of non-phosphorylated S/T/Y residues [59]. Thus, whenever available, we used information from high-resolution 3-dimensional (3D) structures to predict whether a site in a given domain could be accessible, and we kept the sites within domains lying on the protein surface (see *Materials and Methods*). If no reliable structural information was available, we excluded the sites found to be inside domains to get a high confidence set of novel substrate candidates even at the expense of potentially discarding a few real ones. As another filtering method, we

used conservation data to assess whether a site was present in the different orthologs of the corresponding protein (see below). As domains tend to be more conserved than other regions, a good conservation of a site within a domain was not considered as particularly relevant since this did not necessarily imply that the conservation of the putative target residue for phosphorylation had a functional relevance. Sites outside of domains were always considered accessible (Figure 1).

Incorporating evolutionary information

Functionally relevant phosphorylation motifs are expected to be better conserved during evolution than the motif hits that just appear randomly. For instance, Budovskaya *et al.* [37] used information on the conservation of phosphorylation motifs to predict substrates of the cAMP-dependent protein kinase (PKA). They found that the presence of a highly conserved PKA phosphorylation motif was a strong predictor of phosphorylation by PKA *in vitro*, and suggested that substrate candidates with the most highly conserved sites might be targets of PKA *in vivo* [37]. More recently Malik *et al.* found that phosphorylation sites with experimentally verified biological functions are significantly better conserved than other phosphorylation sites [60]. Moreover, an analysis of the overlap between phosphoproteomics datasets of different species revealed it to be higher than expected by chance [61]. We therefore decided to include this level of analysis in our filtering procedure to eliminate biologically irrelevant hits.

In order to assess the conservation of a candidate site, we created sets of orthologs for all proteins in our spindle/centrosome set from the 30 vertebrate species covered by the Ensembl database [62]. We limited the conservation analysis to vertebrates, because short motifs usually evolve quickly and can thus only be expected to be conserved in closely related species [31]. The first step was to check whether the orthologous sequences collected for each protein were diverse enough to distinguish between weakly and highly conserved sites. With this purpose, we assessed the similarity of all orthologs of a given protein through pairwise sequence comparison. We found that they were sufficiently different to distinguish conserved from non-conserved regions, with average sequence identities ranging from 46% to 97%.

The standard approach to determine site conservation is building a multiple sequence alignment (MSA) of all orthologous sequences, followed by either measuring the entropy (or variance) in each column of the MSA [63], or performing an evolutionary trace analysis [64]. That approach, however, depends on the quality of the MSA which itself is dominated by the domains of the given proteins, because they are usually more conserved than other regions, due to their structural constraints [65]. As phosphorylation sites are often located in unstructured or intrinsically disordered regions outside of domains [59, 66] and, as they are fast-evolving functional motifs [31], we could not expect them to be properly aligned in an MSA [65, 66] and thus devised two alternative ways of assessing motif conservation.

Our first strategy was to examine whether the motif was found at roughly the same position in the orthologous sequences as in the corresponding human protein (we will refer to it as the '*presence-based method*'). We first computed the relative position of the phosphorylation site in the human protein, and then checked for its relative matching position in each ortholog, allowing a deviation of 1% in the position. As relative positions of sites may not be preserved across domain insertions/deletions or gene fusion events, this method could not be applied when the orthologs differed too much in size. We therefore restricted this approach to the set of orthologs having a similar length to the corresponding human protein (see *Materials and Methods*).

Our second strategy to assess motif conservation used BLASTP [67] to search for locally matching sequence stretches in vertebrate orthologs containing the phosphorylation motif (we will refer to it as the '*BLAST-based method*'). As BLASTP produces local alignments of the residues surrounding the given phosphorylation sites, this approach could also work in the case of non-uniform changes in sequence length in some species, and thus did not require any restriction for the selection of the set of orthologs (see *Materials and Methods*). Similar approaches have recently been used by Budovskaya *et al.* [37], Malik *et al.* [60] and Boekhorst *et al*. [61] to predict substrates of PKA, identify functionally relevant phosphorylation sites and determine the overlap between phosphoproteomes of different species, respectively.

As conservation values can vary considerably between proteins, we ranked the sites within each protein by conservation. In the substrate candidate selection, we only

considered sites found in the top 5 of the ranking for each substrate candidate. Among the known AurA substrates, 12 out of 19 phosphorylation sites were ranked in the top 5 according to the notP motif conservation, by at least one of our two conservation assessment methods. Concerning the absolute level of conservation, we chose a very strict threshold, only considering motifs conserved in at least 90% of the orthologs by at least one of our two methods. Only 4 out of 19 phosphorylation sites in the known substrates fulfilled this criterion based on the notP motif. However, our rationale here was to develop a method that could identify biologically relevant phosphorylation sites with high specificity, even at the risk of loosing some real substrates. The complete filtering procedure could indeed vastly reduce the number of candidate substrates to 90, with slightly over 1,500 potential phosphorylation sites (Figure 1).

Prioritising substrate candidates

Our final list of 90 candidate substrates for Aurora kinases contained proteins that passed the four filters shown in Figure 1. This is, they must localize in the spindle or centrosome and contain, at least, one accessible and conserved phosphorylation motif. In addition, they fulfilled at least one of the following criteria: (i) the predicted motif had been found to be phosphorylated *in vivo* (ii) a direct or indirect interaction, mediated by only one other protein, with Aurora A had been described in the literature, (iii) at least one consensus site was found to be conserved according to both conservation assessment methods. Using all this information we decided to devise a ranking strategy that could indicate the degree of likelihood for the candidates to be real Aurora substrates (based on available experimental data). Accordingly, we ordered the candidates based on a simple scheme that awards one point (i) to proteins interacting directly or indirectly with AurA, (ii) to proteins having conserved potential phosphorylation site(s) according to the presence-based method or (iii) the BLAST-based method. Two additional points were attributed to proteins having a notP motif coinciding with a known *in vivo* phosphorylated site, therefore providing a very strong support for being a real substrate. Finally, to rank the substrate candidates having the same number of points, we took into consideration the number of sites phosphorylated *in vivo* and conserved according to both the presence- and BLAST-based method (‘3-fold overlap’), as well as the number of sites fulfilling at least two of these three criteria (‘2-fold overlap’), favouring proteins with a

short sequence length, because they are easier to handle experimentally, in case of score ties. The final ranked list of candidate substrates for Aurora kinases and the predicted phosphorylation sites are shown in Table 1.

Experimental validation of the predictions

To validate our predictive method we decided to test whether a randomly selected group of the predicted candidate substrates could be phosphorylated *in vitro* by AurA. The candidate proteins to be tested were selected using as unique criterion the availability of the full-length clone in an appropriate vector for expression in mammalian cells (see *Materials and Methods*). The 10 selected candidates were a good random representation of the candidates as they fell in different positions throughout our ranked list (Table 1). To set up our *in vitro* kinase assay, we first tested the conditions for the detection of phosphorylation by AurA on the previously known substrates TPX2 and TACC3, each with a different tag (TPX2-FLAG and GST-TACC3). As shown in Figure 2A, both proteins were strongly phosphorylated by AurA under our experimental conditions (see *Materials and Methods*).

To assess the specificity of the kinase for proteins containing the consensus motif, we performed *in vitro* kinase assays in the same conditions as above, using as substrate a protein present in the spindle/centrosome proteome but lacking the consensus motif for Aurora kinase phosphorylation (DYNL1). As shown in Figure 2A, AurA did not phosphorylate this protein in any of the two tagged versions, strongly suggesting that the presence of a consensus motif is essential for phosphorylaton by this kinase and confirming that this protein can be used as a negative control in our assays.

We then performed the kinase assay on the ten selected candidate substrates. Depending on the availability of the clones, the candidates were expressed either as FLAG or GST tagged fusion proteins in HEK293T cells and pulled down from cell lysates through their tag. The *in vitro* phosphorylation assays were then performed in parallel with the corresponding positive and negative controls (see *Materials and Methods*). Figures 2B and 2C show representative autoradiographies with the corresponding Coomassie stained gels of these experiments. Since the amount of purified protein and the signal intensities varied widely in each case, the time of gel

exposure to obtain the autoradiographies was adapted to each protein to get the best resolution. As shown in Figure 2B, eight of the ten tested candidate substrates showed incorporation of ^{32}P upon incubation with AurA. Only two of the tested proteins (YWHAG (14-3-3γ) and YWHAE (14-3-3ε)) did not incorporate ^{32}P, indicating that they were not substrates of AurA *in vitro* (Figure 2C).

Since the notP consensus motif matches also Aurora B phosphorylated sites, we then tested whether Aurora B could also phosphorylate any of the selected proteins *in vitro*. As shown in Figure 2C, Aurora B phosphorylated FLAG-RACGAP and GST-TACC3 *in vitro*, showing that the kinase is active in our experimental conditions. Consistently, it did not promote ^{32}P incorporation in our negative control, DYNL1. We found that AurB did indeed give a weak phosphorylation signal in all the positive substrates identified for AurA (Figure 2B). However, in the same conditions, it did not phosphorylate either YWHAG or YWHAE. We further confirmed these negative results, using three different buffers for both kinase reactions, discarding the possibility that the lack of phosphorylation could be due to buffer conditions (data not shown). In addition, to avoid the interference of AurA autophosphorylation signal in the autoradiography (as the kinase and these substrates run at similar positions in the gels), we repeated the kinase assay using another recombinant kinase with a larger tag (GFP-AurA) and therefore running at a different position. Figure 2C presents the full gel and autoradiography of a representative experiment that clearly shows that no signal is detected in the autoradiography at the level of any of the two proteins YWHAG and YWHAE.

These two negative hits, YWHAG and YWHAE (14-3-3γ and 14-3-3ε, respectively) are adapter proteins implicated in the regulation of a large spectrum of both general and specialized signaling pathways. They bind to a large number of partners, usually through the recognition of a phosphoserine or phosphothreonine. In fact, their inclusion in the list was based on data suggesting their indirect interaction with Aurora, probably mediated by a real substrate of the kinase, and conserved notP motif sites. However, as the 14-3-3 family itself is highly conserved, the evolutionary conservation of the motifs does not necessarily imply functional relevance. In such cases, it may be worth considering relative conservation instead, which has recently

been used successfully in the detection of functional motifs [68]. It is possible that a similar situation applies to some of the other candidates present in our list.

To analyze our data in a semiquantitative way, we quantified each autoradiograph signal and normalized the obtained value with the time of film exposure (hours) and the amount of protein loaded (micrograms) (Figure 3A). In agreement with the visual evaluation of the autoradiographies, incubation of CENT1, DYNC1LI1, TUBB4, TUBG1, MAP7, NUSAP and SPIN with AurA resulted in ^{32}P incorporation. Although APC7 also seemed to be phosphorylated, the level of ^{32}P incorporation was much lower than in the other proteins. To determine whether APC7 was really a substrate of AurA *in vitro,* we used a mass spectrometry approach to directly check the putative phosphorylation at Ser85, the predicted target for Aurora's phosphorylation according to our method. The analysis was performed in parallel on APC7 incubated with AurA or only buffer (control). As shown in Figure 3B, the ion assigned to the phosphorylated peptide, $VRP_{phos}\mathbf{S}^{85}$TGNSASTPQSQCLPSEIEVK (Panel A), was not present in measurable amounts in the control sample (Figure 3C, left). This was in sharp contrast to the protein incubated with AurA, which showed a clear signal (Figure 3C, middle). No significant differences were observed when comparing the signals of the non-phosphorylated peptide VRP$\mathbf{S}^{85}$TGNSASTPQSQCLPSEIEVK in the control and the AurA treated samples (Figure 3D) as well as loading controls (three peptides not expected to be affected by the kinase) (Suppl Fig. 3, panels A, B and C) . We conclude that APC7 is a substrate of AurA *in vitro*.

It is worth noting that while we were performing the validation experiments, three reports described the identification of two novel substrates of the human Aurora kinases: Polo-like kinase 1 (Plk1) [69, 70] and Kntc2 (Ndc80) [71]. Both of them were included in our list, further validating our prediction method. Moreover, we also correctly identified the positions of some of the phosphorylated residues.

Our experimental kinase validation assays confirmed that eight out of ten proteins are Aurora substrates *in vitro*, therefore giving an overall prediction success rate of our method of around 80%. Although this rate may vary when increasing the number of tested proteins, our random sample suggests that many of the candidates could be real substrates.

Uncovering novel functions for Aurora kinases

To explore the different roles played by the suggested Aurora substrates, listed in Table 1, we grouped them according to their cellular functions, using the information available at the UniprotKB database (Figure 3). This analysis showed that several of the new candidates clustered together into functionally relevant clusters that were previously found to be under the regulation of Aurora A or B kinases (e.g., centrosome maturation and chromosome alignment). Interestingly, the clustering of several potential substrates of these kinases suggests that there may be a coordinated regulation of different proteins for any given function. Notably, some of the candidate substrates are involved in functions that so far have not been described as regulated by any Aurora kinase in the literature. This is the case for the proteins involved in sumoylation or protein degradation. This last functional group seems particularly interesting since it includes five different proteins of the Anaphase-promoting complex (APC)-dependent degradation machinery. One of them, APC7, is in fact, one of the proteins that we have validated in this work as a true AurA substrate *in vitro*. Another interesting functional cluster contains proteins related to cilia. Indeed, recent work has shown that AurA is essential for cilia reabsorption prior to the entry of quiescent cells into mitosis [26]. The putative phosphorylation of two cilia proteins, CROCC and DNAHC5 (axonemal dynein heavy chain 5) by AurA suggests a wider role function of this kinase in the regulation of cilia function.

The identification of novel putative substrates for the Aurora kinases opens the way to further investigate their known functions but also to examine their potential regulatory role in other as yet uncharacterized cellular processes.

Conclusions

In this work we described an approach to unveil novel candidate substrates for the Aurora kinases. Our *in vitro* experimental validation results suggest that our method has a true predictive potential, with an estimated precision of about 80%. Our approach was to strongly enforce specificity by applying stringent filtering criteria,

and it is likely that we filtered out many other potential candidates, suggesting that the list of Aurora substrates could reach more than a hundred.

Considering the selectivity of the two kinases for its substrates, we observed that with the exception of TPX2 (known to interact specifically with Aurora A but not with B) [7, 9], the other substrates appeared to be phosphorylated by both kinases *in vitro*, with AurA being more efficient in the reaction, even when AurB was five times more concentrated than AurA. This perfectly agrees with previously published reports indicating that AurB activity *in vitro* is much lower than the activity of AurA [72]. This difference between the kinases might be explained by the fact that phosphorylation of the activation loop of AurB is not enough to transform the kinase into a fully active form, as it requires the interaction with cofactors (like INCENP) [73] that naturally interact with the kinase in the cell. Comparison of Aurora A and B activity *in vitro* on a specific substrate is therefore irrelevant at this level. It is also important to remember that Aurora kinases share the basic consensus motif for phosphorylation and that recently published data indicate that functional substrate selectivity among the Aurora family members is probably based on their interaction with specific partners/cofactors and their localization [9, 49]. Therefore we assume that, in general, *in vitro* kinase assays cannot differentiate between the substrate specificity of Aurora kinase family members. Instead, the localization of a substrate in the cell will determine which Aurora kinase phosphorylates it.

From a methodological perspective, we have introduced a new approach to identify likely substrates of the Aurora kinases from all those protein sequences that match the consensus phosphorylation motif for Aurora kinases, so that we can vastly remove false positive predictions. To do this, we have developed two sequence comparison strategies to assess the conservation of phosphorylation sites in other vertebrate species. Although we have only applied our method to Aurora, it can easily be adapted to identify target proteins for other kinases. To estimate the general applicability of the method, with the data that is currently available, we have collected a comprehensive list of 57 known mitotic protein kinases (excluding isoforms), based on recent reviews [50, 74], and have checked for how many of them there is data on phosphorylation sites for at least 10 different substrates. With this criterion, which we consider sufficient for deriving reliable phosphorylation patterns

for each kinase, we found that it would be possible to apply our strategy to 14 of them, representing 25% of the total. If we extend the analysis to the whole proteome, we could identify 113 human protein kinases (excluding isoforms), and for 32 of them there is enough information available to automatically employ our strategy, increasing the potential applicability to roughly 30%.

Overall, we anticipate that the general strategy presented here will help desciphering the many cellular functions regulated by kinases including some that may yet to be discovered. In particular, we hope that our contribution to uncovering the functional roles of Aurora kinases will improve our understanding of the existing link between these kinases and cancer, leading to a better prediction of potential off-target events derived from novel therapeutical approaches.

Materials and methods

Set of human protein sequences and interactions

We collected a comprehensive set of human protein sequences by integrating the 17,317 human proteins from UniProtKB/Swiss-Prot, the 54,362 human protein sequences stored in UniProtKB/TrEMBL and the 11,013 annotated human splice variants associated with those UniProt entries [75]. This resulted in a non-redundant set of 77,683 human protein sequences. To check for direct and indirect interactions (i.e. mediated by one additional protein) between AurA and the known substrates, as well as between AurA and the 308 spindle proteins, we assembled a human interactome by integrating all human protein-protein interactions reported in IntAct [76], MINT [77] and HPRD [78], resulting in a non-redundant set of 48,356 interactions among 11,074 proteins.

Motif assessment and derivation

We used M-Coffee [79] to align the known phosphorylation sites identified in our literature search. We assessed the AurA consensus motifs found in the literature using the alignment of the phosphorylation site ±10 residues and derived the notP motif by comparing conserved and non-conserved positions.

List of spindle proteins

We extracted the lists of proteins whose localization to the centrosome, the spindle apparatus or the kinetochores was known and/or has been confirmed through the large-scale proteomics studies conducted by Andersen *et al.* [53], Sauer *et al.* [54] and Nousiainen *et al.* [55]. We integrated the 114 proteins with confirmed centrosomal localization which Andersen *et al.* found by mass-spectrometry analysis of isolated human centrosomes [53], with the 157 and 72 known spindle components which Sauer *et al.* [54] and Nousiainen *et al.* [55] identified from purified human mitotic spindles, respectively, by mapping the given protein IDs to UniProt Accession Codes (UniProt ACs) [75]. We also considered spindle proteins found in small-scale studies, by including the 80 proteins annotated with 'spindle' for 'cellular component' in GO [56], resulting in a final set of 308 non-redundant spindle proteins.

Exclusion of inaccessible sites

We assembled a set of template sequences from structures in the PDB [80] solved by X-Ray crystallography with a resolution of ≤3.5Å or by nuclear magnetic resonance (NMR). We assigned domains to human spindle proteins using the hidden Markov model (HMM) profiles from Pfam 22 [81, 82]. For each domain that contained any yeast or notP motifs, we used BLASTP [67] to find the best matching template sequence in the PDB based on sequence identity, considering only hits with at least 80% coverage and an E-value $\leq 10^{-4}$. We then cut out the template domain, computed its surface accessibility using NACCESS [83] and mapped it to the given query domain via the Pfam HMM alignment of the two domains. Sites in which all four motif residues had an accessibility above 10 (absolute total side chain value) were considered accessible. We excluded all phosphorylation sites inside domains, except those which were deemed accessible.

Motif conservation based on motif presence

From all vertebrate orthologs provided by Ensembl 49 [62], we chose those with a length within ±5% of the human substrate candidate in order to ensure comparability of relative positions. For a given human candidate protein, we identified all notP motif hits in each ortholog and computed their relative position (with respect to the length of the protein). A candidate site was considered conserved if at least 90% of the orthologs also contained a hit to the motif in about the same relative position (±1%).

Motif conservation based on local sequence matches

To assess the conservation of motif hits based on local sequence matches, we created local alignments of the given query protein sequence to all its orthologous sequences in vertebrate species downloaded from Ensembl 49 [62], using BLASTP [67] and considering only hits with an E-value $\leq 10^{-4}$. We then considered those candidate phosphorylation sites as conserved for which at least 90% of the orthologous sequences matched the given motif at the same position in the alignment.

In vivo phosphorylation data

We integrated the set of 736 *in vivo* phosphorylation sites in 260 proteins, which Nousiainen *et al.* identified from purified human mitotic spindles [55], and the list of 16,989 *in vivo* phosphorylation sites in 3,181 proteins which Dephoure *et al.* extracted from lysates of M-phase arrested cells [84]. Then, we mapped the given protein IDs to UniProt ACs [75] and checked for occurrence of the respective phosphopeptide, identified by tandem mass spectrometry, at about the same position (maximal offset of 20 residues) in the UniProt protein sequence. This resulted in a non-redundant set of 16,316 *in vivo* phosphorylation sites in 3,890 proteins.

DNA constructs

FLAG-tagged full length sequences of human TPX2 (Q9ULWO), APC7 (Q9UJX3), MAP7 (Q14244), RACGAP1 (Q9H0H5), SPIN1 (Q9Y657), NUSAP1 (Q9BXS6) and DYNLL1 (P63167) inserted into pCMV6 entry vector were obtained form Origene. GST-tagged full length sequences of CETN1 (Q12798), YWHAG (P61981), YWHAE (P62258), DYNC1LI1 (Q9Y6G9), TBB4 (P04350), TUBG1 (P23258) and DYNLL1 (P63167) inserted in a pOPINJ vector were obtained from the cloning facility of the Centrosome 3D Consortium (M. Coll; IRB Barcelona). His-AurB clone was prepared by PCR, amplifying the cDNA from a pGEX-4T-1 construct (present from E. Conti; MPI) and inserted into a pET28a vector. To obtain the EGFP version of hAurA, hAurA sequence was PCR amplified from a hAurA-pET28a clone [85] and inserted into a pHAT2-EGFP vector [3].

Protein expression and purification

Aurora A and B proteins were expressed from pET28a-hAurA, pHAT2-EGFP-hAurA and pET28a-hAurB in E. coli and purified through their His tag as previously described [85]. To express the different substrate candidates, HEK293 cells were transfected with the DNA constructs following the FuGENE6 protocol (Roche). After 48 hours, cells were collected by trypsinization, incubated with lysis buffer (50 mM Tris/HCl pH 7.4, 100 mM NaCl, 50 mM NaF, 1% Triton X-100, 1 mM EDTA, 1 mM EGTA, 1 mM DTT and protease inhibitors) on ice for 30 min and spun down (16000 g) for 20 min at 4°C. Protein expression was checked in the supernatant fractions by western blotting. FLAG-tagged recombinant proteins were purified from the cell lysates by immunoprecipitation with anti-FLAG-M2 agarose (Sigma). After incubating the agarose beads in the lysate for 1 h, beads were washed twice with lysis buffer, twice with PBS-NaCl (PBS, 0.1 % triton X-100, 0.5 M NaCl) and twice with MOPS kinase buffer (50 mM MOPS pH 7.4, 5 mM $MgCl_2$, 1 mM EGTA, 1 mM EDTA, 10 mM β-glycerophosphate). To purify the GST-tagged proteins, Glutathione Sepharose 4B (Amersham) aliquots were incubated in the cell lysates for 1 h, washed twice with lysis buffer, twice with PBS-NaCl (PBS, 0.1 % triton X-100, 0.5 M NaCl) and twice with MOPS kinase buffer.

In Vitro kinase assay

Directly after protein purification, aliquots of beads covered with the candidate proteins were resuspended in 10 μl MOPS kinase buffer mixed with hAurA (0.2 μM), hAurB (1 μM) or buffer (to reduce the autophosphorylation signal from the kinases, both hAurA and hAurB were preincubated with 100 μM cold ATP at 37°C for 30 min before mixing with the candidate substrates). ^{32}P-ATP was then added to the reaction and samples were incubated at 37°C for 40 min. Reaction was stopped by adding SDS-page loading buffer, and proteins were separated by electrophoresis. The Coomassie-stained gels were scanned with the Odyssey imaging system (Li-cor). Autoradiographs were obtained by exposing the gel to a PhosphoImager film (Fuji-Film) that was later scanned with a Typhoon Trio Imager (Amersham Biosciences). The intensity of the bands of the autoradiography was measured using ImageQuant 5.2 software. In parallel, the amount of protein loaded for each sample was determined by quantifying the Coomassie Brilliant Blue intensity with Odyssey software using different amounts of bovine serum albumin as standards. To compare

the ^{32}P incorporation between the different samples, autoradiography measurements were normalized by the μg of substrate protein present in the gel.

Mass spectrometry analysis

Proteins treated with buffer of hAurA were hereafter separated by 1D-gel electrophoresis and excised bands were trypsinized (Promega, Madison, WI, USA) following a previously described protocol [86]. Extracted peptides were analyzed by reversed phase HPLC (Agilent 1200 nano flow pump, Agilent Technologies, CA) coupled to a mass spectrometer (Orbitrap XL, ThermoFisher, Bremen, Germany). MS/MS data were extracted and searched against the IPI Human database (Date) using Mascot version 2.2 (Matrix Science Inc., London, UK). 10 ppm and 0.5 Da was used as MS and MS/MS mass accuracies, respectively. N-terminal acetylation, oxidation of methionine and phosphorylation of serine, threonine and tyrosine residues were allowed as variable modifications. To identify phosphorylation events specific to the hAurA kinase, a label-free approach was chosen as previously described [87]. In short: Extracted ion chromatograms were created for the identified phosphopeptides for both treated and control samples. These chromatograms were compared to the chromatograms of (i) the corresponding unmodified peptide sequence and (ii) peptides not expected to be modified. MS/MS spectra of phosphopeptides and extracted ion chromatograms for the phosphopeptides, corresponding non-phosphorylated peptides and reference peptides are shown in Figure 3 and Supp. Fig 3. By comparing the extracted ion chromatograms of the corresponding non-phosphorylated peptides in treated and control samples and using peptides not expected to be modified, we estimated the degree of phosphorylation.

Acknowledgments

We thank A. Zanzoni (IRB Barcelona) for helpful discussions, and M. Coll (IRB Barcelona) and E. Conti (MPI of Biochemistry) for providing some of the clones used in the study. PA acknowledges the financial support received from the Spanish Ministerio de Innovación y Ciencia through the grants BIO2007-62426 and PSE-010000-2009, and the European Commission under FP7 Grant Agreement

223101(AntiPathoGN). Work in the Vernos laboratory is supported by Ministerio de Innovación y Ciencia grants BFU2006-04694, BFU2005-24990-E, and CSD2006-00023. RAP is a recipient of the Spanish FPU fellowship.

References

1. Vader G, Lens SM: **The Aurora kinase family in cell division and cancer.** *Biochim Biophys Acta* 2008, **1786:**60-72.
2. Barr AR, Gergely F: **Aurora-A: the maker and breaker of spindle poles.** *J Cell Sci* 2007, **120:**2987-2996.
3. Sardon T, Peset I, Petrova B, Vernos I: **Dissecting the role of Aurora A during spindle assembly.** *Embo J* 2008, **27:**2567-2579.
4. Slattery SD, Mancini MA, Brinkley BR, Hall RM: **Aurora-C kinase supports mitotic progression in the absence of Aurora-B.** *Cell Cycle* 2009, **8:**2984-2994.
5. Cheeseman IM, Anderson S, Jwa M, Green EM, Kang J, Yates JR, 3rd, Chan CS, Drubin DG, Barnes G: **Phospho-regulation of kinetochore-microtubule attachments by the Aurora kinase Ipl1p.** *Cell* 2002, **111:**163-172.
6. Kufer TA, Sillje HH, Korner R, Gruss OJ, Meraldi P, Nigg EA: **Human TPX2 is required for targeting Aurora-A kinase to the spindle.** *J Cell Biol* 2002, **158:**617-623.
7. Bayliss R, Sardon T, Ebert J, Lindner D, Vernos I, Conti E: **Determinants for Aurora-A activation and Aurora-B discrimination by TPX2.** *Cell Cycle* 2004, **3:**404-407.
8. Adams RR, Wheatley SP, Gouldsworthy AM, Kandels-Lewis SE, Carmena M, Smythe C, Gerloff DL, Earnshaw WC: **INCENP binds the Aurora-related kinase AIRK2 and is required to target it to chromosomes, the central spindle and cleavage furrow.** *Curr Biol* 2000, **10:**1075-1078.
9. Fu J, Bian M, Liu J, Jiang Q, Zhang C: **A single amino acid change converts Aurora-A into Aurora-B-like kinase in terms of partner specificity and cellular function.** *Proc Natl Acad Sci U S A* 2009, **106:**6939-6944.
10. Keen N, Taylor S: **Aurora-kinase inhibitors as anticancer agents.** *Nat Rev Cancer* 2004, **4:**927-936.
11. Mahadevan D, Beeck S: *Expert Opinion Drug Discovery* 2007, **2:**15.
12. Bischoff JR, Anderson L, Zhu Y, Mossie K, Ng L, Souza B, Schryver B, Flanagan P, Clairvoyant F, Ginther C, et al: **A homologue of Drosophila aurora kinase is oncogenic and amplified in human colorectal cancers.** *Embo J* 1998, **17:**3052-3065.
13. Chen HL, Tang CJ, Chen CY, Tang TK: **Overexpression of an Aurora-C kinase-deficient mutant disrupts the Aurora-B/INCENP complex and induces polyploidy.** *J Biomed Sci* 2005, **12:**297-310.

14. Naruganahalli KS, Lakshmanan M, Dastidar SG, Ray A: **Therapeutic potential of Aurora kinase inhibitors in cancer.** *Curr Opin Investig Drugs* 2006, **7:**1044-1051.
15. Zeng WF, Navaratne K, Prayson RA, Weil RJ: **Aurora B expression correlates with aggressive behaviour in glioblastoma multiforme.** *J Clin Pathol* 2007, **60:**218-221.
16. Kurai M, Shiozawa T, Shih HC, Miyamoto T, Feng YZ, Kashima H, Suzuki A, Konishi I: **Expression of Aurora kinases A and B in normal, hyperplastic, and malignant human endometrium: Aurora B as a predictor for poor prognosis in endometrial carcinoma.** *Hum Pathol* 2005, **36:**1281-1288.
17. Dutertre S, Descamps S, Prigent C: **On the role of aurora-A in centrosome function.** *Oncogene* 2002, **21:**6175-6183.
18. Anand S, Penrhyn-Lowe S, Venkitaraman AR: **AURORA-A amplification overrides the mitotic spindle assembly checkpoint, inducing resistance to Taxol.** *Cancer Cell* 2003, **3:**51-62.
19. McGrogan BT, Gilmartin B, Carney DN, McCann A: *Biochim Biophys Acta Rev Cancer* 2008, **1789:**36.
20. Taylor S, Peters JM: **Polo and Aurora kinases: lessons derived from chemical biology.** *Curr Opin Cell Biol* 2008, **20:**77-84.
21. Hughes TV, Emanuel SL, O'Grady HR, Connolly PJ, Rugg C, Fuentes-Pesquera AR, Karnachi P, Alexander R, Middleton SA: **7-[1H-Indol-2-yl]-2,3-dihydro-isoindol-1-ones as dual Aurora-A/VEGF-R2 kinase inhibitors: design, synthesis, and biological activity.** *Bioorg Med Chem Lett* 2008, **18:**5130-5133.
22. Pollard JR, Mortimore M: **Discovery and development of aurora kinase inhibitors as anticancer agents.** *J Med Chem* 2009, **52:**2629-2651.
23. Lin YG, Immaneni A, Merritt WM, Mangala LS, Kim SW, Shahzad MM, Tsang YT, Armaiz-Pena GN, Lu C, Kamat AA, et al: **Targeting aurora kinase with MK-0457 inhibits ovarian cancer growth.** *Clin Cancer Res* 2008, **14:**5437-5446.
24. Yang J, Ikezoe T, Nishioka C, Tasaka T, Taniguchi A, Kuwayama Y, Komatsu N, Bandobashi K, Togitani K, Koeffler HP, et al: **AZD1152, a novel and selective aurora B kinase inhibitor, induces growth arrest, apoptosis, and sensitization for tubulin depolymerizing agent or topoisomerase II inhibitor in human acute leukemia cells in vitro and in vivo.** *Blood* 2007, **110:**2034-2040.
25. Boss DS, Beijnen JH, Schellens JH: **Clinical experience with aurora kinase inhibitors: a review.** *Oncologist* 2009, **14:**780-793.
26. Pugacheva EN, Jablonski SA, Hartman TR, Henske EP, Golemis EA: **HEF1-dependent Aurora A activation induces disassembly of the primary cilium.** *Cell* 2007, **129:**1351-1363.
27. Mori D, Yamada M, Mimori-Kiyosue Y, Shirai Y, Suzuki A, Ohno S, Saya H, Wynshaw-Boris A, Hirotsune S: **An essential role of the aPKC-Aurora A-NDEL1 pathway in neurite elongation by modulation of microtubule dynamics.** *Nat Cell Biol* 2009, **11:**1057-1068.
28. Knebel A, Morrice N, Cohen P: **A novel method to identify protein kinase substrates: eEF2 kinase is phosphorylated and inhibited by SAPK4/p38delta.** *EMBO J* 2001, **20:**4360-4369.
29. Neville DC, Rozanas CR, Price EM, Gruis DB, Verkman AS, Townsend RR: **Evidence for phosphorylation of serine 753 in CFTR using a novel metal-**

ion affinity resin and matrix-assisted laser desorption mass spectrometry. *Protein Sci* 1997, **6:**2436-2445.
30. Meng L, Michaud GA, Merkel JS, Zhou F, Huang J, Mattoon DR, Schweitzer B: **Protein kinase substrate identification on functional protein arrays.** *BMC Biotechnol* 2008, **8:**22.
31. Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB: **Systematic discovery of new recognition peptides mediating protein interaction networks.** *PLoS Biol* 2005, **3:**e405.
32. Edwards RJ, Davey NE, Shields DC: **SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins.** *PLoS One* 2007, **2:**e967.
33. Edwards RJ, Moran N, Devocelle M, Kiernan A, Meade G, Signac W, Foy M, Park SD, Dunne E, Kenny D, Shields DC: **Bioinformatic discovery of novel bioactive peptides.** *Nat Chem Biol* 2007, **3:**108-112.
34. Stein A, Aloy P: **Contextual specificity in peptide-mediated protein interactions.** *PLoS One* 2008, **3:**e2524.
35. Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jorgensen C, Miron IM, Diella F, Colwill K, Taylor L, Elder K, et al: **Systematic discovery of in vivo phosphorylation networks.** *Cell* 2007, **129:**1415-1426.
36. Snead JL, Sullivan M, Lowery DM, Cohen MS, Zhang C, Randle DH, Taunton J, Yaffe MB, Morgan DO, Shokat KM: **A coupled chemical-genetic and bioinformatic approach to Polo-like kinase pathway exploration.** *Chem Biol* 2007, **14:**1261-1272.
37. Budovskaya YV, Stephan JS, Deminoff SJ, Herman PK: **An evolutionary proteomics approach identifies substrates of the cAMP-dependent protein kinase.** *Proc Natl Acad Sci U S A* 2005, **102:**13933-13938.
38. Ferrari S, Marin O, Pagano MA, Meggio F, Hess D, El-Shemerly M, Krystyniak A, Pinna LA: **Aurora-A site specificity: a study with synthetic peptide substrates.** *Biochem J* 2005, **390:**293-302.
39. Ohashi S, Sakashita G, Ban R, Nagasawa M, Matsuzaki H, Murata Y, Taniguchi H, Shima H, Furukawa K, Urano T: **Phospho-regulation of human protein kinase Aurora-A: analysis using anti-phospho-Thr288 monoclonal antibodies.** *Oncogene* 2006, **25:**7691-7702.
40. Diella F, Gould CM, Chica C, Via A, Gibson TJ: **Phospho.ELM: a database of phosphorylation sites--update 2008.** *Nucleic Acids Res* 2008, **36:**D240-244.
41. Meraldi P, Honda R, Nigg EA: **Aurora kinases link chromosome segregation and cell division to cancer susceptibility.** *Curr Opin Genet Dev* 2004, **14:**29-36.
42. Barros TP, Kinoshita K, Hyman AA, Raff JW: **Aurora A activates D-TACC-Msps complexes exclusively at centrosomes to stabilize centrosomal microtubules.** *J Cell Biol* 2005, **170:**1039-1046.
43. Troiani S, Uggeri M, Moll J, Isacchi A, Kalisz HM, Rusconi L, Valsasina B: **Searching for biomarkers of Aurora-A kinase activity: identification of in vitro substrates through a modified KESTREL approach.** *J Proteome Res* 2005, **4:**1296-1303.
44. Wu JC, Chen TY, Yu CT, Tsai SJ, Hsu JM, Tang MJ, Chou CK, Lin WJ, Yuan CJ, Huang CY: **Identification of V23RalA-Ser194 as a critical mediator for Aurora-A-induced cellular motility and transformation by small pool expression screening.** *J Biol Chem* 2005, **280:**9013-9022.

45. Briassouli P, Chan F, Savage K, Reis-Filho JS, Linardopoulos S: **Aurora-A regulation of nuclear factor-kappaB signaling by phosphorylation of IkappaBalpha.** *Cancer Res* 2007, **67:**1689-1695.
46. Wynshaw-Boris A: **Lissencephaly and LIS1: insights into the molecular mechanisms of neuronal migration and development.** *Clin Genet* 2007, **72:**296-304.
47. Venoux M, Basbous J, Berthenet C, Prigent C, Fernandez A, Lamb NJ, Rouquier S: **ASAP is a novel substrate of the oncogenic mitotic kinase Aurora-A: phosphorylation on Ser625 is essential to spindle formation and mitosis.** *Hum Mol Genet* 2008, **17:**215-224.
48. Turk BE: **Understanding and exploiting substrate recognition by protein kinases.** *Curr Opin Chem Biol* 2008, **12:**4-10.
49. Hans F, Skoufias DA, Dimitrov S, Margolis RL: **Molecular distinctions between Aurora A and B: a single residue change transforms Aurora A into correctly localized and functional Aurora B.** *Mol Biol Cell* 2009, **20:**3491-3502.
50. Nigg EA: **Mitotic kinases as regulators of cell division and its checkpoints.** *Nat Rev Mol Cell Biol* 2001, **2:**21-32.
51. Carmena M, Earnshaw WC: **The cellular geography of aurora kinases.** *Nat Rev Mol Cell Biol* 2003, **4:**842-854.
52. Marumoto T, Zhang D, Saya H: **Aurora-A - a guardian of poles.** *Nat Rev Cancer* 2005, **5:**42-50.
53. Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, Mann M: **Proteomic characterization of the human centrosome by protein correlation profiling.** *Nature* 2003, **426:**570-574.
54. Sauer G, Korner R, Hanisch A, Ries A, Nigg EA, Sillje HH: **Proteome analysis of the human mitotic spindle.** *Mol Cell Proteomics* 2005, **4:**35-43.
55. Nousiainen M, Sillje HH, Sauer G, Nigg EA, Korner R: **Phosphoproteome analysis of the human mitotic spindle.** *Proc Natl Acad Sci U S A* 2006, **103:**5391-5396.
56. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
57. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, et al: **ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins.** *Nucleic Acids Res* 2003, **31:**3625-3630.
58. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK: **The importance of intrinsic disorder for protein phosphorylation.** *Nucleic Acids Res* 2004, **32:**1037-1049.
59. Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, Mann M: **PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites.** *Genome Biol* 2007, **8:**R250.
60. Malik R, Nigg EA, Korner R: **Comparative conservation analysis of the human mitotic phosphoproteome.** *Bioinformatics* 2008, **24:**1426-1432.
61. Boekhorst J, van Breukelen B, Heck AJ, Snel B: **Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes.** *Genome Biol* 2008, **9:**R144.

62. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, et al: **Ensembl 2008.** *Nucleic Acids Res* 2008, **36:**D707-714.
63. Pei J, Grishin NV: **AL2CO: calculation of positional conservation in a protein sequence alignment.** *Bioinformatics* 2001, **17:**700-712.
64. Mihalek I, Res I, Lichtarge O: **A family of evolution-entropy hybrid methods for ranking protein residues by importance.** *J Mol Biol* 2004, **336:**1265-1282.
65. Chica C, Labarga A, Gould CM, Lopez R, Gibson TJ: **A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences.** *BMC Bioinformatics* 2008, **9:**229.
66. Perrodou E, Chica C, Poch O, Gibson TJ, Thompson JD: **A new protein linear motif benchmark for multiple sequence alignment software.** *BMC Bioinformatics* 2008, **9:**213.
67. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
68. Davey NE, Shields DC, Edwards RJ: **Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery.** *Bioinformatics* 2009, **25:**443-450.
69. Seki A, Coppinger JA, Jang CY, Yates JR, Fang G: **Bora and the kinase Aurora a cooperatively activate the kinase Plk1 and control mitotic entry.** *Science* 2008, **320:**1655-1658.
70. Macurek L, Lindqvist A, Lim D, Lampson MA, Klompmaker R, Freire R, Clouin C, Taylor SS, Yaffe MB, Medema RH: **Polo-like kinase-1 is activated by aurora A to promote checkpoint recovery.** *Nature* 2008, **455:**119-123.
71. Ciferri C, Pasqualato S, Screpanti E, Varetti G, Santaguida S, Dos Reis G, Maiolica A, Polka J, De Luca JG, De Wulf P, et al: **Implications for kinetochore-microtubule attachment from the structure of an engineered Ndc80 complex.** *Cell* 2008, **133:**427-439.
72. Eyers PA, Churchill ME, Maller JL: **The Aurora A and Aurora B protein kinases: a single amino acid difference controls intrinsic activity and activation by TPX2.** *Cell Cycle* 2005, **4:**784-789.
73. Yang J, Zappacosta F, Annan RS, Nurse K, Tummino PJ, Copeland RA, Lai Z: **The catalytic role of INCENP in Aurora B activation and the kinetic mechanism of Aurora B/INCENP.** *Biochem J* 2009, **417:**355-360.
74. Schmit TL, Ahmad N: **Regulation of mitosis via mitotic kinases: new opportunities for cancer management.** *Mol Cancer Ther* 2007, **6:**1920-1931.
75. UniProt-Consortium: **The universal protein resource (UniProt).** *Nucleic Acids Res* 2008, **36:**D190-195.
76. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, et al: **IntAct--open source resource for molecular interaction data.** *Nucleic Acids Res* 2007, **35:**D561-565.
77. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular INTeraction database.** *Nucleic Acids Res* 2007, **35:**D572-574.

78. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, et al: **Human protein reference database--2006 update.** *Nucleic Acids Res* 2006, **34:**D411-414.
79. Wallace IM, O'Sullivan O, Higgins DG, Notredame C: **M-Coffee: combining multiple sequence alignment methods with T-Coffee.** *Nucleic Acids Res* 2006, **34:**1692-1699.
80. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-242.
81. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2008, **36:**D281-288.
82. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
83. Hubbard SJ, Thornton JM: **NACCESS.** In *Book NACCESS* (Editor ed.^eds.). pp. Computer Program. City; 1993:Computer Program.
84. Dephoure N, Zhou C, Villen J, Beausoleil SA, Bakalarski CE, Elledge SJ, Gygi SP: **A quantitative atlas of mitotic phosphorylation.** *Proc Natl Acad Sci U S A* 2008, **105:**10762-10767.
85. Bayliss R, Sardon T, Vernos I, Conti E: **Structural basis of Aurora-A activation by TPX2 at the mitotic spindle.** *Mol Cell* 2003, **12:**851-862.
86. Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M: **In-gel digestion for mass spectrometric characterization of proteins and proteomes.** *Nat Protoc* 2006, **1:**2856-2860.
87. Steen H, Jebanathirajah JA, Springer M, Kirschner MW: **Stable isotope-free relative and absolute quantitation of protein phosphorylation stoichiometry by MS.** *Proc Natl Acad Sci U S A* 2005, **102:**3948-3953.

Rank	Prot name	UniProt ID	Phosphorylation sites
1	TOP2B	Q02880	S50, S129, T1371, **S1550**, **S1576**
2	CENPE	Q02224	S201, **T422**, S434, **S454**, T626, S789, S1320, T1505, S1562, T1671, T1712, S2048, S2119, S2314, T2450, S2468
3	TPX2	Q9ULW0	**S121**, **S125**, **S359**, S729, S742
4	INCE	Q9NQS7	S72, S87, S91, S94, S452, **T897**, **S898**
5	APC7	Q9UJX3	S80, **S85**, S322
6	AURKB	Q96GD4	**T232**
7	PLK1	P53350	**T210**, T259
8	2A5D	Q14738	**S88**, S454, T550, **S573**
9	NEDD1	Q8NHV4	S171, S176, S293, **S423**, S637
10	MAP7	Q14244	T146, S161, S181, **S200**, S705
11	KIF11	P52732	T222, T223, T370, **S931**
12	CENPC*	Q03188	T183, **S261**, **S763**
13	Q8IW27*	Q8IW27	T183, **S261**, **S763**
14	KIF4A*	O95239	**T799**, **S1028**, **S1126**
15	CND1	Q15021	S442, T454, **S585**, T761, T1203, **T1388**, S1395
16	CE170	Q5SW79	T116, S126, **S312**, S472, **T501**, **T628**, S829, S952, **S958**, S968, S980, T1074, T1090, T1098, S1101, S1123, T1156, S1210, T1405
17	KIF1B	O60333	T1122, T1295, S1347, T1439, **S1658**
18	SMC1A*	Q14683	**S970**
19	TOP2A	P11388	**S1387**
20	RGAP1	Q9H0H5	**S144**, T472, **S573**, **T601**
21	HASP	Q8TF76	**S93**, **S143**, T660
22	RBP2	P49792	S352, T714, S778, S786, S1912, T2192, T2398, **S2447**
23	SPIN1	Q9Y657	T31, **S124**, S241
24	KI20A*	O95235	**S244**, S632, S677
25	AZI1	Q9UPN4	**S76**, **T93**, S202
26	EDC4	Q6P2E9	**S871**, T1075
27	CLAP1	Q7Z460	**S646**, S687, T730, S741
28	CK5P2	Q96SN8	**S547**, S1804
29	NUMA1	Q14980	T1879, S1883, S1887, S1901, **S1945**, S1969, S1991, T2007, S2026, S2047

Rank	Prot name	UniProt ID	Phosphorylation sites
30	ASPM	Q8IZT6	S149, S160, S170, **S270**, T503, T640, **S1103**, S3003, T3383, S3387, T3406, S3425
31	MACF1	Q9UPN3	S57, T362, S647, T671, S814, S1268, T1393, **S5009**, S5086, T5158, T5208, S5264, S5267, S5277, S5324, S5350, S5367, S5372, S5409, T5418, T5419
32	NIN	Q8N4C6	S109, T811, S812
33	CETN1	Q12798	S85, T102, S122, S170
34	1433G	P61981	S58, S59, S71, T145
35	HSP71	P08107	T425, T495, T502, S537
36	1433E	P62258	S59
37	2AAA	P30153	S146, T268, T568
38	CBX3	Q13185	**S93**
39	Q9BUB1	Q9BUB1	**S58**, **S99**
40	NUSAP	Q9BXS6	**S240**
41	TBB2C	P68371	**S322**, **S338**
42	DC1L2	O43237	**T441**
43	DC1L1	Q9Y6G9	**T213**, **T456**, **S487**
44	KNTC2*	O14777	**S55**, **S62**
45	CND2	Q15003	**S70**
46	Q86T11	Q86T11	**S717**, **S752**
47	KIF23	Q02241	**S160**, **S911**, **S912**
48	MAP4	P27816-2	**T354**, **S863**
49	Q9BTE9	Q9BTE9	**T668**
50	MAP4	P27816	**S636**
51	SPAG5	Q96R06	**S362**
52	NUP98*	P52948	**S618**
53	CASC5*	Q8NG31	**S60**, **S765**
54	STK6	O14965	S155, T288
55	KAPCA	P17612	S339
56	CDC20	Q12834	S487, S492
57	CKAP5	Q14008	T1813, S1825, S1861, S1872, S1983, S1988
58	LATS1	O95835	S4, S37, S45, S58, S84, S142, S416, S629, S690
59	CTRO	O14578	S25, T119, S480, S496, S520, S582, S592, S652, S888, S942, T1018, T1174, S1568
60	TBB4	P04350	S176, T218
61	DCTN1	Q14203	S19, S351, T460, T481, T753, T950, S958, S974, T1044, T1152, S1153, T1210, T1249

Rank	Prot name	UniProt ID	Phosphorylation sites
62	CETN2	P41208	T102, S170
63	TP4A1	Q93096	S163
64	CBX1	P83916	S141, T169
65	DPOLB	P06746	S104
66	SEPT2*	Q15019	S31, S234
67	WDR8	Q9P2S5	S10, T117
68	SEPT9*	Q9UHD8	S324, T518
69	CDC23	Q9UJX2	S257
70	KIF22	Q14807	S244, T284, S389
71	KI18A	Q8NI77	S266
72	GCP3	Q96CW5	S92, S896, S904, S905
73	ANLN	Q9NQW6	T684, S753, S803, S937
74	CROCC	Q5TZA2	S676, S1069, S1084, T1109, S1187, S1204, S1218, S1569, S1575, S1591, S1619, S1624, S1784, T1881, S1900, S2004, S2005
75	CEP35	Q5VT06	S4, S17, T22, S67, S68, S76, T96, S232, S239, T255, S557, T577, S814, T831, T956, S974, S1143, T1307, T1658, T1680, S2161, S2737, S2744, S2752, T2766, T3100
76	DYH5	Q8TE73	T948, T974, T977, S996, S1013, S2942, S3335, T3474, T3504, T3769, T3791, S3824
77	SUMO1	P63165	S9
78	CDC2	P06493	T26
79	KC1A	P48729	T109, S237, S242
80	KC1D	P48730	S101, S229
81	TBG1	P23258	T288
82	TBB5	P07437	S48, S176, T218
83	NEK2	P51955	S29
84	FBX5	Q9UKT4	S94, S354, T395, S440
85	TCPA	P17987	T13
86	HS90B	P08238	S206
87	HS90A	P07900	S211
88	CDC27	P30260	T171
89	PP1G	P36873	S22
90	Q07161	Q07161	S336

Table 3.1: Aurora manuscript – Table 1: Ranking of the 90 substrate candidates for Aurora kinases. Known substrates of Aurora A or B are shown in purple, and proteins known to localize to the kinetochores or the spindle midzone are marked with a star as potential substrates of AurB. Experimentally tested candidates are shown in gray with new confirmed substrates highlighted in green. Substrate candidates known to interact with AurA either directly or indirectly (i.e. mediated by another protein) are shown in orange and yellow, respectively. Finally, all accessible and conserved matches to the notP motif ([KR].[ST][^ P]) are listed, and known in vivo phosphorylation sites are marked in bold.

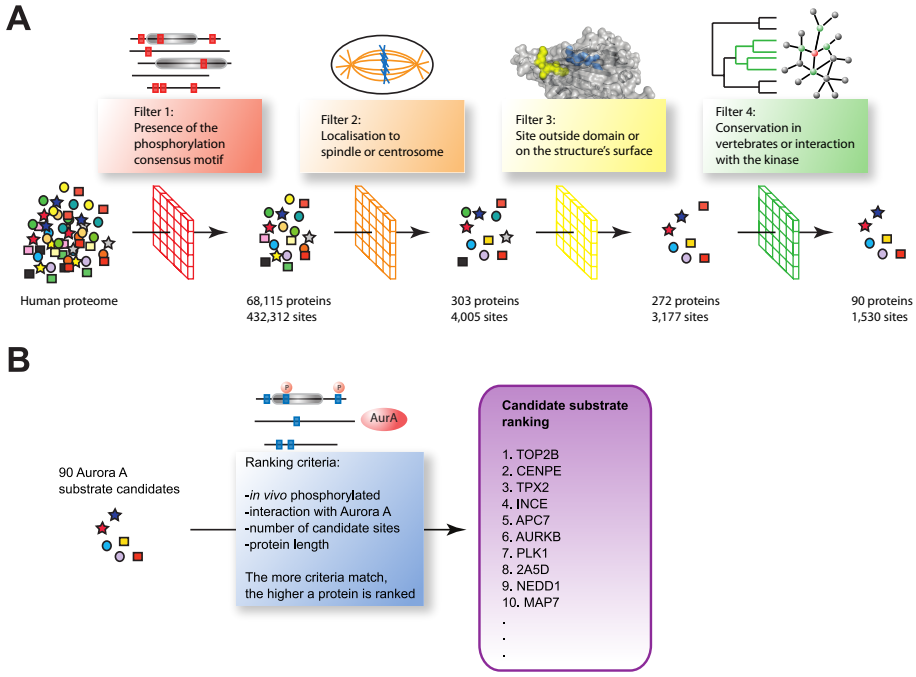

Figure 3.13: Aurora manuscript – Figure 1: Schematic representation of the bioinformatics approach developed to uncover new Aurora kinase substrates. (A) Candidate substrates selection. Aurora substrate candidates were selected based on a series of filters applied to the whole human proteome: presence of an Aurora phosphorylation motif in the sequence, localization to the centrosome or the spindle, accessibility of the consensus motif and conservation of the potential phosphorylation site among vertebrates. (B) Ranking of candidate Aurora substrates. The 90 proposed Aurora substrates were ranked according to different criteria (see the main text for details).

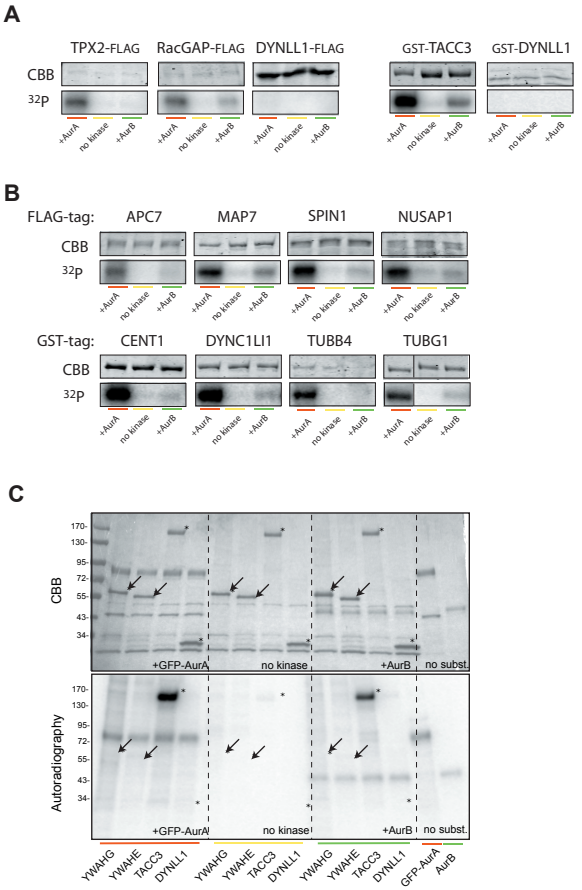

Figure 3.14: Aurora manuscript – Figure 2: Validation of candidate substrates by in vitro phosphorylation assays. (A) The specificity of the *in vitro* Aurora kinase assay was first tested on known AurA substrates (positive controls) and on a protein not containing the consensus motif for Aurora phosphorylation (DYNLL1, negative control). Substrates were expressed in human cells, and pulled down by their FLAG or GST tag. Precipitated proteins were mixed with Aurora A (AurA), Aurora B (AurB) or buffer (no kinase) in the presence of 32P-ATP, then separated by SDS-page and stained with Coomassie brilliant blue (CBB). Incorporated 32P was visualized by autoradiography (32P). Shown gel fragments belong to representative experiments repeated at least three times. (B) AurA phosphorylation assays were performed as in (A) on unknown substrate candidates randomly selected from the candidate list. Shown gel fragments belong to representative experiments repeated at least twice. (C) Autoradiography (32P) and Coomassie staining (CBB) of an *in vitro* kinase assay gel where YWHAG (14-3-3γ) and YWHAE (14-3-3ϵ) were tested in comparison with a positive (TACC3) and a negative control (DYNLL1). An EGFP-tagged version of AurA was used to avoid interference of the autophosphorylation signal.

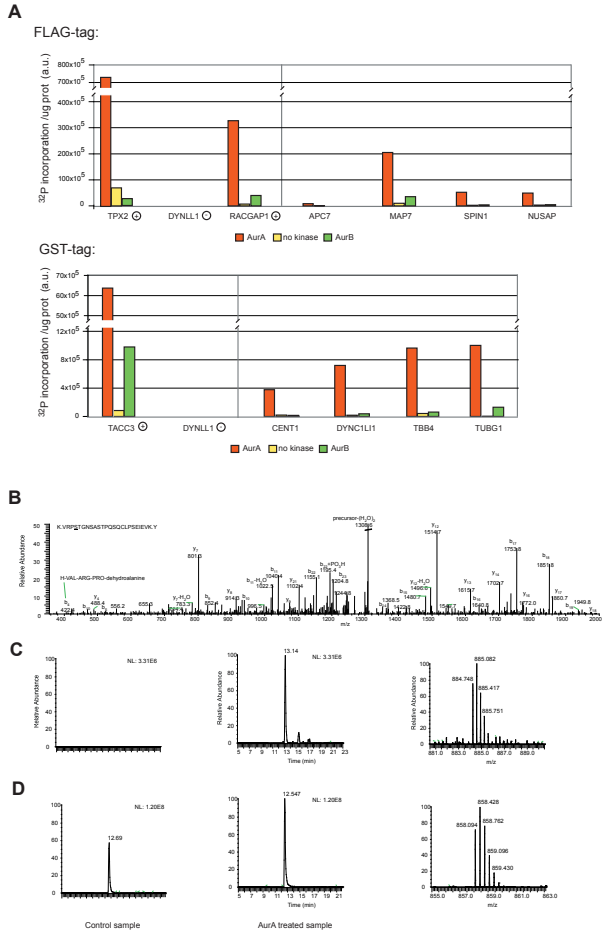

Figure 3.15: Aurora manuscript – Figure 3: Quantification of the incorporation of 32P by the different substrates and MS analysis of APC7. (A) Semi-quantitative comparison of the 32P incorporated into the different substrates after incubation with AurA, buffer or AurB. Values quantified from the autoradiography were normalized by the amount of substrate (μg) loaded on the gel. Graphs show results from a representative experiment repeated at least twice. (B) Tandem MS spectrum for APC7. Graph shows the assigned MS/MS spectrum of the tryptic phosphopeptide VRPphosS85TGNSASTPQSQCLPSEIEVK (m/z 1326.622, Mascot score 53, Expect 1.1.10-3). (C) Extracted ion chromatograms for the phosphorylated tryptic peptide of APC7 in control sample (left) and treated sample (center). MS/MS spectrum of the respective peptide is shown on the right. (D) Extracted ion chromatograms for the non-phosphorylated VRPS85TGNSASTPQSQCLPSEIEVK tryptic peptide of APC7 in control sample (left) and treated sample (center). On the right, MS/MS spectrum of the respective peptide. Scales for each of the two sets of extracted ion chromatograms are kept constant.

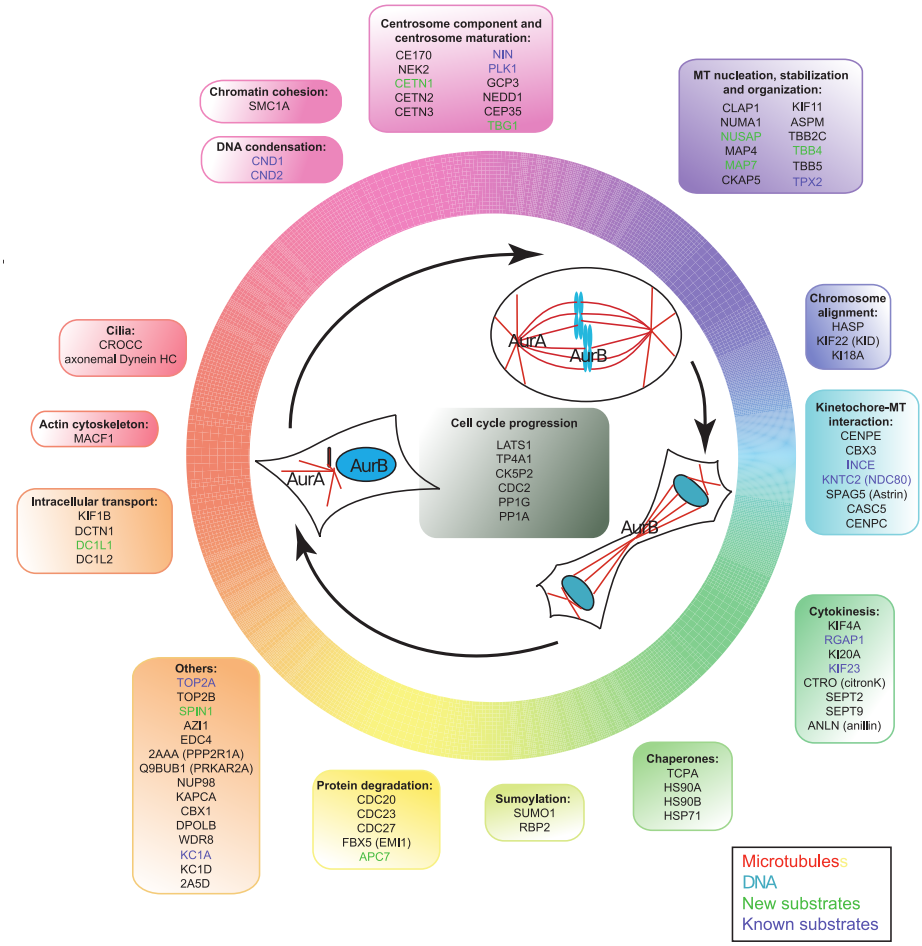

Figure 3.16: Aurora manuscript – Figure 4: Functional distribution of the predicted Aurora substrates. The known and candidate substrates predicted by our method were classified according to their functions throughout the cell cycle. Proteins without well-defined attributed function and proteins affecting several pathways were classified as "others". In the cell schemes, microtubules appear in red and DNA in blue. In the text boxes, blue, red and black indicate known, validated and proposed Aurora substrates, respectively.

3.3.a) Supplementary Material

Calculation of motif enrichment in the set of spindle/centrosomal proteins

To assess the specificity gained by integrating sub-cellular localisation data and thus limiting the set of potential substrate candidates to only those proteins which localise to the mitotic spindle and/or the centrosome, we calculated the enrichment of proteins with an Aurora A phosphorylation motif in the set of 308 spindle/centrosomal proteins (Supplementary Figure 3.17). We computed the enrichment of proteins with a yeast and notP motif, respectively, in the set of 308 spindle/centrosomal proteins by taking the ratio of the fraction of spindle proteins containing the given motif and the fraction of all human proteins with that motif. We also determined the enrichment of the number of motif hits per protein in the set of spindle proteins by calculating the ratio of the average number of motifs per spindle protein and the average number of motifs per human protein. To assess the statistical significance of the enrichments, we performed Monte-Carlo permutation tests by constructing a background of 10,000 random sets of 308 proteins each. As the length distribution of all human proteins is significantly different from the one of the spindle proteins (p-value$=1.96 \cdot 10^{-46}$, two-sided Mann-Whitney U test), and sequence length correlates with the number of motif occurrences (R^2=0.85), we controlled for protein length by considering only random human proteins with sequence lengths similar to the ones in the original set of spindle proteins ($\pm 5\%$) when calculating the statistical significance of the enrichment.

We found that the set of spindle proteins is 1.49-fold enriched in proteins with the yeast motif [KR].[ST][ILV], which is significantly higher than the enrichment of equally-sized random subsets of human proteins (p-value$=1 \cdot 10^{-3}$, Monte-Carlo permutation test). As it had been shown for the cyclin-dependent kinase Cdk1 of the budding yeast *Saccharomyces cerevisiae* that phosphorylation motifs can be clustered within substrates (Moses et al., 2007), we also determined the enrichment of the number of motif hits per spindle protein. Again controlling for sequence length, we found a 2.57-fold enrichment over the number of motifs per human protein (p-value $< 10^{-4}$, Monte-Carlo permutation test), indicating that there are significantly more yeast motifs present in the 308 spindle proteins than in equally-sized random

subsets of human proteins. The same holds for proteins with the notP motif [KR].[ST][^ P], which are 1.12-fold enriched in the set of spindle proteins (p-value=$4.02 \cdot 10^{-2}$, Monte-Carlo permutation test), and for the number of motifs per protein, which is 2.34-fold enriched in the spindle proteins (p-value $< 10^{-4}$, Monte-Carlo permutation test). Thus, the integration of sub-cellular localisation data significantly increased the specificity of our predictions.

Supplementary Table 1

Due to its excessive length, this table is not printed here. It is available at `www-aloy.irbbarcelona.org/publications/supplementary/aurora_suppl_table_1.xls`
Detailed legend for the individual pages of the table:

Expression and interaction of known substrates. Indicates for each known substrate of Aurora A and B whether it is expressed in HeLa cells, T cells, or ubiquitously (expression data taken from HPRD (Prasad et al., 2009)). These cell types were the main source of the spindle data set used in this study (Andersen et al., 2003; Sauer et al., 2005; Nousiainen et al., 2006). In addition, we state whether the known substrates interact with Aurora A.

GO annotations. GO terms for cellular component and biological process for all 308 spindle proteins.

In vivo phosphorylation. Data for all spindle proteins that have been observed phosphorylated in vivo, derived from Nousiainen et al. (2006); Dephoure et al. (2008).

Interaction with Aurora A. List of proteins in the spindle dataset that interact with Aurora A, either directly or mediated by a third protein (indirect interaction).

Conservation of known ph.sites. Shows the conservation of known Aurora A phosphorylation sites as computed by the BLAST-based and the presence-based method, respectively. Both methods were developed for this study and are described in the main text of the manuscript.

Alternative protein names: UniProt AC and ID as well as other common names for the 308 spindle proteins used in this study. Known Aurora substrates are

highlighted in violet, experimentally tested proteins in gray, new substrates in green.

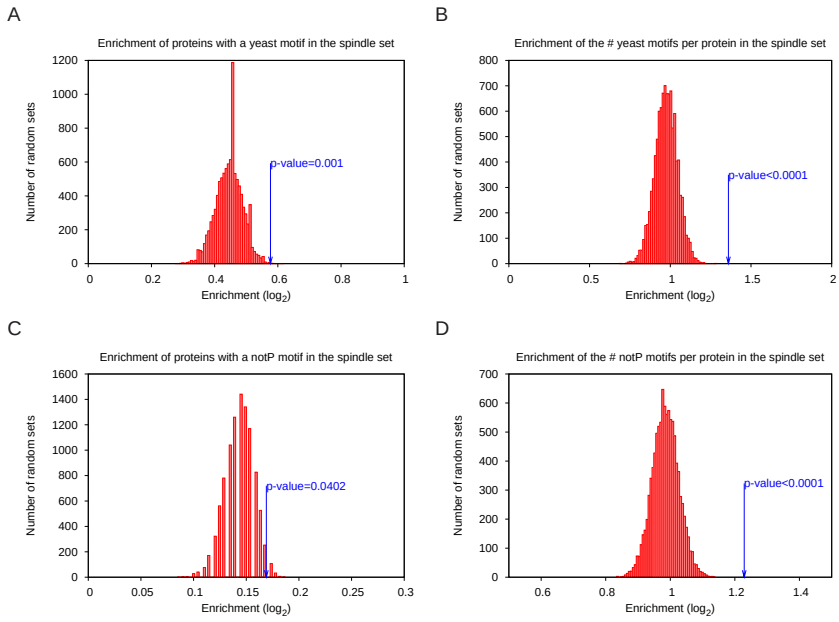

Figure 3.17: Aurora manuscript – supplementary figure 1: Enrichment of phosphorylation motifs and of proteins containing at least one of those motifs in the set of spindle proteins compared to the human proteome. A: Enrichment of proteins with at least one yeast motif ([KR].[ST][ILV]). B: Enrichment of the number of yeast motifs per protein. C: Enrichment of proteins with at least one notP motif ([KR].[ST][^ P]). D: Enrichment of the number of notP motifs per protein. Enrichments are given on a $\log_2$ scale, with those of the 10,000 equally-sized random sets of human proteins shown in red, and the respective enrichment in the set of spindle proteins marked with a blue arrow. We controlled the random background for protein length by considering only human proteins with sequence lengths similar to the ones in the set of spindle proteins ($\pm 5\%$).

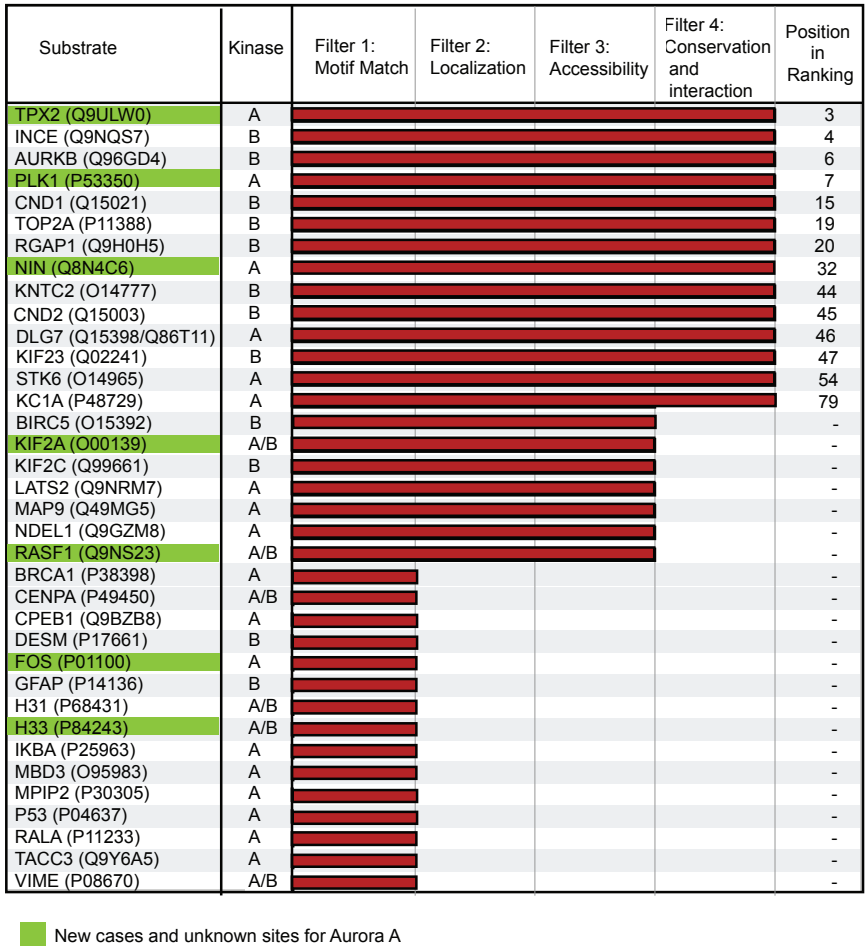

Figure 3.18: Aurora manuscript – supplementary figure 2: Filtering and ranking of known Aurora substrates. Proteins known to be phosphorylated by Aurora A or B, sorted by whether they contain a motif match (filter 1), localise to the spindle and/or centrosome (filter 2), whether motif hits are accessible (filter 3), and whether the proteins either contain conserved hits or interact with Aurora A (filter 4), and their position in the candidate ranking, if applicable.

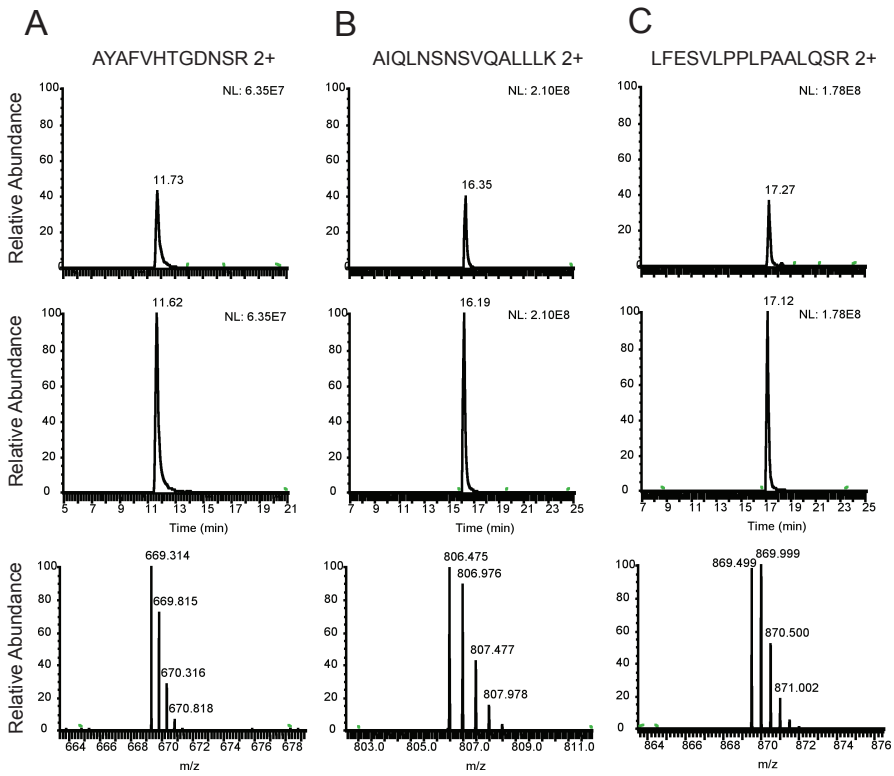

Figure 3.19: Aurora manuscript – supplementary figure 3: Normalization controls for the MS analysis of the APC7 phosphorylation at S85 by Aurora A. Extracted ion chromatograms created for the APC7 phosphopeptide VRPphosSTGNSASTPQSQCLPSEIEVK for both treated and control samples (Fig. 3.15) were compared to the chromatograms of peptides not expected to be modified (reference peptides). Panel A to C show the MS/MS spectra of peptides and the extracted ion chromatograms for sets (control vs. treatment) of selected reference tryptic APC7 peptides. Peptides originating from the control are shown in the upper ion chromatogram, while peptides from the Aurora A treated APC7 are shown in lower chromatogram. MS spectra of the respective peptides are shown below. The peptides shown in panel A through C: (A) AYAFVHTGDNSR 2+, (B) AIQLNSNSVQALLLK 2+ and (C) LFESVLPPLPAALQSR 2+, are peptides not expected to undergo modification and serve as normalization control. Scales for each of the sets of extracted ion chromatograms are kept constant.

3.4 Database of 3D interacting domains – 3did

The database of 3D interacting domains (3did) is a platform to make high-resolution structural data of domains in interaction available to the scientific community. 3did was initially created to provide a collection of pairs of interacting Pfam domains for which 3D structures are available (Stein et al., 2005). Domain-domain interactions (DDIs) are stable functional associations of proteins, with an interface size of around $2.000\mathring{A}^2$ on average. DDIs occur between domains of different proteins (interchain interactions), but also within one protein, i.e., between the domains of a multidomain protein (intrachain interactions). The atomic details available in high-resolution 3D structures (Berman et al., 2000) and collected in 3did allow for detailed studies of the role of individual regions or residues in these interactions, and provide crucial information for mutational studies as well as engineering of protein interactions.

More on DDIs
→ 1.4.b), page 30

Recently, we have introduced domain-motif interactions (DMIs) into 3did, to offer a more complete picture of the molecular details of the interactions a given domain may participate in (Stein et al., 2009b). In DMIs, a linear motif of 4-10 residues is bound by a globular domain, which again may happen within or between proteins, although data on intraprotein DMIs currently is scarce. Due to the shortness of these motifs, DMIs are usually transient and have much smaller interfaces than DDIs ($350\mathring{A}^2$ on average). They are often involved in signal transduction, where their transient nature allows for fast propagation of the information, and may depend on post-translational modifications (PTMs) such as phosphorylations or methylations. Another recent addition to 3did is the computation of interface topologies to group DDIs and DMIs into *interaction types*. This is based on the observation that homologous proteins usually interact in the same way, i.e., with the same relative orientation and using the same binding interfaces (Aloy and Russell, 2004). However, especially in the case of proteins or domains that interact with many different partners, there may be more than one interaction topology. To separate DDIs and DMIs by interaction type, we have developed a clustering procedure based on the fraction of shared interface residues, which is outlined here and described in more detail in a subsequent manuscript (see page 65). In the future, we plan to incorporate the domain-motif interactions identified in Stein and Aloy (2010) into 3did, comprising 3D structures as well

as regular expressions to describe the typical residues found in the peptides for a variety of domains. This will provide a much broader basis for structural studies of peptide-mediated interactions than what is currently available.

Stein A, Russell RB, Aloy P. 3did: interacting protein domains of known three-dimensional structure. Nucleic Acids Res. 2005; 33(Database issue): D413-7.

Stein A, Panjkovich A, Aloy P. 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. Nucleic Acids Res. 2009; 37 (Database issue): D300-4.

4 Discussion

The work presented in this thesis addresses two major topics: contextual specificity, or the information encoded by an interaction's environment that determines its specificity, and the value and availability of 3D structures in studying modular protein interactions. As a starting point, we identified numerous instances of peptide-mediated or domain-motif interactions (DMIs) in high-resolution 3D structures by searching for the consensus motifs provided in the Eukaryotic Linear Motif database (ELM, Puntervoll et al. (2003)) and the corresponding binding domain. This yielded hundreds of instances of 3D structures of linear motifs, involving 30 different recognition domains. We used this data to study the molecular basis of the high specificity of DMIs, and found that the context makes substantial contributions, and moreover is involved in unfavorable contacts in the flanking regions that are assumed to be critical in the determination of specificity, especially as these unfavorable contacts are more pronounced in constructed non-native interactions (Stein and Aloy, 2008). During the work with these structures of DMIs, we noticed that the peptides assume a particular flat and elongated structure, which is different from the shape of other peptides in random samples of 3D structures. We thus attempted to exploit this structural property to identify other DMIs in 3D structures, searching the spatial neighborhoods of globular domains for occurrence of such peptides. In combination with a number of filters regarding the interaction interface, we identified thousands of putative DMIs, which could be clustered into around 800 interaction types. For over 100 of them, sufficient data was available to determine a consensus motif, and 64 of those were also significantly enriched in the interactomes of four model organisms, which serves as an additional indicator that these may be true biological motifs (Stein and Aloy, 2010). In our database of 3D interacting domains (3did, Stein et al. (2005, 2009b)), we collect structures of domain-domain and domain-motif interactions (DDIs and DMIs, respectively). This offers a rich resource of atomic details to researchers analyzing

the molecular details of the interactions formed by these modular domains. It is used actively by the scientific community and cited in over 100 publications.[1] With the incorporation of more and more DMIs, we intend to provide as much high-resolution data as possible on this transient yet highly specific type of modular interaction, as to facilitate structure-based studies on their nature. In addition, 3D structures are excellent templates for predictions of interactions involving these domains, as it has been shown for DDIs (Aloy and Russell, 2002a). With the growing availability of structures for DMIs, such as those described in this thesis, the implementation of similar methods for peptide-mediated interactions should become possible in the near future. With the prediction of substrates for the Aurora A kinase, we have applied our knowledge of DMIs to their role in protein interaction networks. Since neither structural data of a substrate at the active site nor docking motifs for Aurora are known, we had only little data on the molecular context of this interaction. This, together with the rather degenerate motif, meant that it was particularly important to consider other contextual information. Thus we integrated data on sub-cellular localization and involvement in spindle formation, *in vivo* phosphorylation data and conservation rates in order to determine those sites and proteins that are most likely to actually be phosphorylated by Aurora A. With the experimental validation of 8 out of 10 candidate substrates, we have shown that our method has a high accuracy in substrate prediction (Sardon et al., 2010). More generally, this indicates that contextual information is often beneficial in pinpointing substrates or interaction partners of a specific protein.

4.1 Contextual Specificity

Despite their shortness, peptide-mediated interactions are known for their high specificity *in vivo*, and it had been shown previously that individual positions around the motif have a critical effect on this (Zarrinpar et al., 2003). Having collected a set of 3D structures with sufficient variety both within and across interaction types, we set out to analyze the molecular interfaces in order to find out how the high specificity is achieved. As expected, we observed that the motif is crucial for binding and contributes around 80% of

[1] citation count by Google Scholar

the interaction energy on average. Hence, the contribution of the context in our dataset is around 20%, although this is likely an underestimate as many structures of DMIs only contain truncated peptides, so that the full effect of the context might be stronger than observed here. To further assess specificity and cross-talk, we performed a peptide exchange experiment among instances of the same interaction type. For the motif itself, we observed that native as well as non-native peptides bind well according to our energy computations. However, we did observe a number of neutral and unfavorable contacts in the flanking regions, which are clearly present in the native DMIs and more pronounced in the non-native interactions (Stein and Aloy, 2008). We speculate that these unfavorable contacts serve to prevent undesired cross-talk. Given that many interactions involving DMIs are transient, a high affinity supported by lots of favorable contacts is not required. Thus, it may well be that these interactions have evolved to maximise specificity rather than affinity. However, to some extent cross-talk might actually be a desired side effect, as it would allow a cross-binding protein to take over a particular function in the cell in case of loss of the native participant (Kelley and Ideker, 2005; Ulitsky and Shamir, 2007). Estimating the possibility of cross-talk is thus an important aspect in the field of systems biology. When comparing the energetic assessment of cross-talk with sequence similarity, we observed that there is no clear correlation; some instances had high sequence similarity but were predicted not to bind, while other, sequentially distant pairs seemed to allow cross-talk (Fig. 3.4 on page 63). This implies that sequence data alone will likely not be sufficient to determine whether a given domain and peptide will bind, and that 3D structures contain important details which could be exploited in such a prediction.

The importance of the flanking regions in DMIs for specificity and motif structure has recently been emphasized in several other studies, based on sequence analysis and phage display data (Chica et al., 2009; Davey et al., 2009; Tonikian et al., 2009). Once a domain of interest is identified, phage display experiments with sufficiently long peptides can assess molecular specificity on a similar level as our approach. To complement structure-based studies, phage display data can provide broader coverage, though with lower resolution. Thus a combination of the two approaches would be well-suited for the detailed characterization of individual peptide-binding domains, as it has been shown for a few domains by the SPOT technique (Brannetti et al., 2000). In

particular, the broad coverage of phage display would allow the identification of "forbidden" residues, which have a strong negative effect on the interaction. Given data from 3D structures alone, it is possible to determine favorable contacts in the interface, but much harder to predict which mutations would abolish binding. However, in order to combine structural data with that from other experiments of lower resolution, a sufficient number of 3D structures first needs to be available for each peptide-domain interaction type. With the work presented in this thesis, we have identified such structures for dozens of domains, and made them available in our database 3did. We have also shown that some domains bind peptides in different orientations, which makes 3D templates even more important, as they can explain and potentially help classify different binding patterns. An advantage of having multiple structures of DMIs is that this allows studies on which parts of the interface are conserved and which are more variable, as it has been done for DDIs recently (Panjkovich and Aloy, 2010). Also, as shown here, the motif's position is much more fixed topologically than that of the flanking regions. This implies that changes in the motif region also have a stronger effect on the interaction, though further studies are required to confirm this.

Several of the DMIs studied here are not recognized based on a peptide alone, but require post-translational modifications (PTMs) like phosphorylations. Perhaps due to the large and often charged nature of such modifications, the consensus motifs of PTM-dependent DMIs are often even more degenerate than those of PTM-independent DMIs. For instance, the recognition motifs for SH2 in ELM are rather short, and structures of the corresponding DMI contain only few flanking residues, so that an analysis of the role of the molecular context is difficult. Instead, we assume that the biological context of phosphorylation networks provides important specificity information here – many PTMs indicate activation of a particular pathway or are characteristic for a specific state of the cell. The enzymes in charge of their addition and removal specifically recognize their substrates, which in turn at least partly ensures specificity of the PTM-dependent DMIs as it couples them to the respective states of the cell in which the modification is present (Linding et al., 2007). This could be referred to as "dynamic context", which exists on the molecular, e.g. with the PTMs described above, but also on the cellular level, for instance in the form of expression patterns or macromolecular assemblies. When analyzing the molecular specificity of the Tumor Necrosis Factor

(TNF)-receptor-associated factors (TRAF) peptides, we could not observe highly specific individual interactions. Their biological function is based on receptor trimerization, though, which is assumed to enhance specificity (Park et al., 1999). Another example is the Aurora A kinase, which has a crucial role in spindle assembly during the cell cycle. When searching for potential substrates of Aurora A, we therefore only considered proteins related to spindles, whether through co-purification or by functional associations. As indicated above, the scarcity of molecular contextual data made the use of cellular-level contextual information critical in this case. *In vivo* phosphorylation data from mass-spectrometry (MS) studies can pinpoint sites in proteins known to be modified, thus providing a further indicator for putative substrates. With the application of our prediction method to Aurora A in human, we created a list of 90 substrates, of which 10 were tested experimentally and 8 shown to be substrates of Aurora A *in vitro*, indicating a precision of 80%. Furthermore, MS analysis performed for 3 of these substrates confirmed that the phosphorylation site coincided with our prediction. This shows that our integration of different, orthogonal kinds of contextual information works well in identifying substrates and even phosphorylation sites, especially when considering the fact that the motif itself occurs multiple times in thousands of human proteins (Sardon et al., 2010). Based on the currently available phosphorylation data in resources such as phospho.ELM (Diella et al., 2004), we estimate that the methodology developed as part of this thesis and successfully applied to Aurora A can readily be transferred to more than 30 human kinases. Ongoing efforts like the characterization of phosphoproteins by MS and better interactome coverage make it likely that our approach will become applicable to an even wider range of kinases in the near future. Similarly, other enzymes for addition or removal of PTMs, such as phosphatases, should be amenable to substrate identification using adapted versions of this method.

Our studies presented in this thesis as well as other recent publications clearly show the importance of contextual information in determining specificity, especially in transient interactions such as DMIs. Therefore, we think that approaches integrating a variety of contextual data are most promising for studying and predicting biological interactions *in vivo*. The combination of data from different sources has an additional advantage: hardly any method for their detection is free from errors. By focusing on "cumulative evidence", we can minimize the risk of artifacts due to problems in individual datasets.

In principle, the contextual factors discussed here also apply to interaction specificity in designed proteins, synthetic circuits and organisms in the emerging field of synthetic biology. For instance, the set of putative substrates that will actually bind to designed interaction surfaces (e.g., Humphris and Kortemme (2007)) *in vivo* is affected by contextual factors such as sub-cellular localization and expression patterns. Nevertheless, in synthetic organisms, it may happen that a pair of proteins with a highly specific interaction in their native environment shows unexpected cross-talk when transferred to a different cell type, if it encounters competitors that the interaction has not been fine-tuned for (cf. Zarrinpar et al. (2003)). Knowledge of the detailed molecular specificity of the interactions involved may help prevent undesired cross-talk in those cases.

4.2 3D structures of modular protein interactions

High-resolution 3D structures are a great source of information for studies on the molecular specificity of protein interfaces. In the case of domain-domain interactions (DDIs), atomic-level details have been used successfully to predict whether two given domains will interact (Aloy and Russell, 2002a). The increased coverage of structures allows the application of this method to ever more pairs of domains (cf. Fig. 1.14 on page 37). We collect structural instances of DDIs in our database of 3D interacting domains (3did, Stein et al. (2005, 2009b)), which is widely used in the scientific community for studies of domain-domain interactions (e.g., Itzhaki et al. (2006)). For DMIs, however, such structural data was not easily accessible, because, unlike DDIs, DMIs usually cannot be reliably identified in sequences due to their often degenerate patterns. In addition, it had been assumed that there were only a few dozen instances of DMIs in 3D structures (Neduva and Russell, 2006). In the first project of this thesis, we set out to identify DMIs in 3D structures based on their recognition domains and an occurrence of the motif as described in ELM in the vicinity. Yet, not all 2,200 structures fulfilling these criteria did contain true DMIs. Visual inspection of each of them, and comparison to descriptions of DMIs in the literature were required in order to separate real DMIs from artifacts due to degenerate motifs. This approach has generated a highly reliable set of DMIs in structure, and research in several other groups has

already followed up on our findings (Parthasarathi et al., 2008; Chica et al., 2009; Davey et al., 2009; Rubinstein and Niv, 2009; London et al., 2010). The need for visual inspection in this approach, however, means that it is not applicable at larger scales, nor is it feasible to perform regular updates for newly published structures as we do it for DDIs in 3did. In spite of that, it was possible for us to develop a different, automated method for the recognition of DMIs in structure. While working with the DMIs described above, we observed a structural regularity of the peptide shape and derived structural parameters for their identification, which could be used for the automated detection of DMIs in 3D structures. We have successfully implemented a procedure for this in a subsequent project, which is also presented in this thesis (Stein and Aloy, 2010). Particular examples of the structural characteristics of peptides in DMIs had been reported previously, for example, with the SH3-binding linker region in the Src kinase that forms a PPII-helix even though it is not proline-rich (Pawson, 1997). To our knowledge, though, we are the first to exploit this characteristic on a general level, not limited to individual recognition domains, in order to identify novel instances of DMIs among all 3D structures. Our approach has strict thresholds regarding the interaction interface in order to minimize false positives, which, as our benchmark with a false-positive rate of 0 shows, works well. Nevertheless we successfully rediscover two thirds of the DMIs of known structure. The method has also been able to identify DMIs based on β-strand-addition and those involving peptides that form α-helices, which had posed difficulties to other structure-based methods (Petsalaki et al., 2009). While we do rediscover 67% of the known DMIs, the benchmark indicates that half of the undetected known DMIs (15% of all cases) are missed because the motifs are located in domains, which are currently excluded by our motif search procedure. This corresponds to previous estimates of motifs located in domains (e.g., Neduva et al. (2005)). However, due to the fact that these motifs are usually located in surface loops (Via et al., 2009), it should be possible to implement a structural filter that allows motifs on domain surfaces to be recognized. In fact, we are employing a similar strategy in our search for substrates for the Aurora kinase, where we only accept motifs inside domains if a structural template of the domain shows that they are accessible (Sardon et al., 2010). A recently published method for the identification of DMIs based on 3D structures does consider domain-domain interfaces as well (Willy et al., 2010). However, that method does not state any requirements for minimum motif contribution or interface

ratio, such that it is not clear how they differentiate interactions based on motifs inside domains from true DDIs.

In contrast to the structural characteristics of motifs exploited here, many sequence-based methods for the *de novo* identification of motifs focus on unstructured regions, in which motifs are frequently found. In other words, we do not focus on the fact that motif-containing regions may be unstructured on their own, but that they assume a particular conformation when bound to their recognition domain. The data pool available to sequence-based methods is much larger, yet these methods tend to suffer from a poor signal-to-noise ratio (Chica et al., 2009; Davey et al., 2009; Stein et al., 2009a). Methods based on structures of globular domains alone, such as that by Petsalaki et al. (2009), rely on smaller datasets than sequence-based methods, but can still be applied to more domains than our method, as more structures of individual domains are available than of DMIs. Nevertheless, neither of these methods is able to readily provide molecular details on both the domain and the motif side. These details, though, are crucial for mutational studies, to enhance our functional understanding of these interactions, as well as for the design of synthetic peptides and drug-like components (Reményi et al., 2006; Mandell and Kortemme, 2009; Rubinstein and Niv, 2009), to which DMIs, thanks to their small interface and transient nature, are especially amenable (Neduva and Russell, 2006; Russell and Gibson, 2008). A combination of all three types of methods should yield good accuracy and coverage in the *de novo* identification of motifs. As they are detected and classified by combining data from several instances, such motifs should be considered typical descriptors of the binding preferences of a particular domain, or perhaps a subset of domains of a given family, depending on the conservation of the domain in question. Regular expressions appear to be good descriptors of such preferences. However, for the detailed characterization of an individual peptide-binding domain, one should incorporate additional data from molecular computations on 3D structures and/or experiments such as phage display, which, as outlined above, can provide complementary information. In this case, PSSMs or other more complex models might be better suited for the description of the particular domain's binding characteristics (see also Tonikian et al. (2008, 2009); Chen et al. (2008)).

In the application of our method as presented in this thesis, we have identified motifs for almost 100 domains, around 80 of which which are not annotated in

ELM. 46 of these are significantly enriched in an interactome cross-validation, which is a good independent indicator of their biological significance. Identification of these hitherto unnoticed DMIs allows for the extraction of valuable biological and biophysical information, such as the relationship between peptide-binding domains and the consensus motifs they recognize. This data may then be used either in a predictive way, to scan protein sequences for occurrences of such motifs, or to provide molecular details like interface residues for interactions discovered in experiments with lower resolution, e.g., those from high-throughput interaction detection experiments. For instance, the motifs identified here that were significantly enriched in the interactome cross-validation may offer molecular explanations for over 6,000 PPIs in current model species interactomes (Table 4.1). Thanks to the automated nature of our approach, it can be applied regularly in order to mine newly published structures for occurrence of peptide-mediated interactions. Furthermore, with the increase of data in the PDB, more instances of putative DMIs for which there currently is not enough data to derive a motif may be identified, thus providing enough non-redundant information to make motif derivation possible. Similarly, with the improving coverage of PPI networks more of the motifs derived with our method could show significant enrichment in interactomes, which is an additional validation of their biological function. The inclusion and regular update of these DMIs in 3did means that hundreds of high-resolution 3D structures will become available for analyses of the molecular details in interactions, or for use as templates in structure-based interaction prediction methods. Given a sufficient number of templates both within and across interaction types, it should be possible to model many DMIs. In order to determine whether such a modelled interaction will actually take place, it is important to both understand the molecular binding preferences of the domains sufficiently to transfer this knowledge to modelled structures and to incorporated contextual information, particularly on the dynamic and cellular context that may not be evident from the interaction interface itself.

In addition to 3did, a number of other databases for structures of DMIs have recently become available, including ADAN (Encinar et al., 2009), DOMINO (Ceol et al., 2007), PeptiDB (London et al., 2010) and PepX (Vanhee et al., 2009). Some of these resources have specifically collected non-redundant sets of DMIs. While redundancy is often a problem, especially in structure-based datasets as structural data is not evenly distributed across the different se-

Organism	Number of interactions	with DDI models	with DMI models
yeast	60,000	2,300	950
worm	6,000	330	19
fly	19,000	1,150	160
human	53,000	9,400	5,200

Table 4.1: DMIs in interactome networks. While current interactome drafts are still incomplete (cf. section 1.1.b) and Table 1.1), modular interactions for which 3D structures are available can only provide molecular details for a fraction of them at the moment. Domain-domain interactions (DDIs) for which we have 3D structures in 3did (Stein et al., 2009b) could explain 4-18% of the given interactomes. The consensus patterns derived for the domain-motif interactions detected in this study (Stein and Aloy, 2010) are able to explain over 6,000 of the remaining PPIs in these four interaction networks (i.e., we are not considering those PPIs that can be explained by a DDI), which corresponds to up to 12%, depending on the species. As the table also shows, a large number of PPIs are currently lacking molecular models, so much further research on interaction interfaces is needed.

quence families, in our study on flanking regions we intentionally included all instances of any interaction type in order to be able to assess cross-talk and study the topological variability of DMIs (Stein and Aloy, 2008). Similarly, when searching for possible templates to model an interaction it is advantageous to access a pool with as much data as possible, in order to select the best-matching template. Thus, depending on the application or aim of a project, scientists may have different needs for the selection of the underlying dataset. As the multitude of resources and many other references cited in this thesis indicate, the interest in the field of motif-mediated interactions has substantially increased over the last years (Gould et al., 2010). Scientists have characterized DMIs *in vivo*, *in vitro* and *in silico*, have developed computational, experimental and hybrid methods for *de novo* motif detection and are tackling the different factors that determine their high specificity. We hope that our contributions to this field help to advance the understanding of peptide-mediated interactions.

5 Conclusions

In this thesis, we have studied peptide-mediated interactions from a structural point of view, focusing on molecular details and including complementary biological information to address their high specificity.

- High-resolution 3D structures of protein interactions are currently the only source of atomic details for interaction interfaces. We have identified hundreds of 3D structures of peptide-mediated interactions based on the consensus motifs described in the Eukaryotic Linear Motif database (ELM), followed by manual curation.
- By analyzing those 3D structures of peptide-mediated interactions, we showed that the molecular context, as defined by the contacts formed by the motif-flanking regions, contributes about 20% of the binding energy. Furthermore, we observed that this context forms many non-favorable contacts which are even more pronounced in non-native pairs of peptides and domains. These are assumed to account for the high specificity found in these interactions. Topologically, the motif's position is well-defined while the context is much more flexible.
- We noted a particular flat and elongated shape of the motif-containing peptides and exploited this structural characteristic to develop a method for a generic, motif-independent strategy of identifying peptide-mediated interactions in 3D structures. Application of this method to all known interactions of 3D structures lead to 64 interaction types with consensus motifs that were enriched in protein interaction networks. These peptide-mediated interactions might offer molecular explanations for over 6000 protein-protein interactions that have been observed experimentally and cannot be explained by domain-domain interactions.
- In our database of 3D interacting domains (3did), we make 3D struc-

tures and molecular interface details of domain-domain as well as domain-peptide interactions available to the scientific community. 3did is accessed and cited by a number of research groups across the world.

- In interactions that depend on post-translational modifications such as phosphorylations, the molecular context is not sufficient to explain their specificity. Higher level data such as sub-cellular localization, interaction networks and conservation can pinpoint those proteins that are likely to interact *in vivo*. We developed a method to integrate such contextual data for the human kinase Aurora A, and could predict novel substrates with high accuracy. With currently available data, this method is readily applicable to at least 30 other human kinases.

Bibliography

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Sali, A., and Rout, M. P. (2007). The molecular architecture of the nuclear pore complex. *Nature*, 450(7170):695–701.

Aloy, P. and Russell, R. B. (2002a). Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A*, 99(9):5896–5901.

Aloy, P. and Russell, R. B. (2002b). The third dimension for protein interactions and complexes. *Trends Biochem Sci*, 27(12):633–638.

Aloy, P. and Russell, R. B. (2003). InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, 19(1):161–162.

Aloy, P. and Russell, R. B. (2004). Ten thousand interactions for the molecular biologist. *Nat Biotechnol*, 22(10):1317–1321.

Aloy, P. and Russell, R. B. (2005). Structure-based systems biology: a zoom lens for the cell. *FEBS Lett*, 579(8):1854–1858.

Aloy, P. and Russell, R. B. (2006). Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol*, 7(3):188–197.

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402.

Alzari, P. M., Berglund, H., Berrow, N. S., Blagova, E., Busso, D., Cambillau, C., Campanacci, V., Christodoulou, E., Eiler, S., Fogg, M. J., Folkers, G., Geerlof, A., Hart, D., Haouz, A., Herman, M. D., Macieira, S., Nordlund, P., Perrakis, A., Quevillon-Cheruel, S., Tarandeau, F., van Tilbeurgh, H., Unger, T., Luna-Vargas, M. P. A., Velarde, M., Willmanns, M., and Owens, R. J. (2006). Implementation of semi-automated cloning and prokaryotic expression screening: the impact of SPINE. *Acta Crystallogr D Biol Crystallogr*, 62(Pt 10):1103–1113.

Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P., Nigg, E. A., and Mann, M. (2003). Proteomic characterization of the human centrosome by protein correlation profiling. *Nature*, 426(6966):570–574.

Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S. E., Hubbard, T. J. P., Chothia, C., and Murzin, A. G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res*, 36(Database issue):D419–D425.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A. T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S. N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K., and Hermjakob, H. (2010). The IntAct molecular interaction database in 2010. *Nucleic Acids Res*, 38(Database issue):D525–D531.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29.

Bader, G. D. and Hogue, C. W. V. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, 20(10):991–997.

Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science*, 289(5481):905–920.

Barr, A. R. and Gergely, F. (2007). Aurora-A: the maker and breaker of spindle poles. *J Cell Sci*, 120(Pt 17):2987–2996.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res*, 32 Database issue:D138–41.

Bax, A. and Torchia, D. A. (2007). Structural biology: molecular machinery in action. *Nature*, 445(7128):609.

Beck, M., Lučić, V., Förster, F., Baumeister, W., and Medalia, O. (2007). Snapshots of nuclear pore complexes in action captured by cryo-electron tomography. *Nature*.

Berman, H., Henrick, K., Nakamura, H., and Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res*, 35(Database issue):D301–D303.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000). The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–42.

Bhattacharyya, R. P., Reményi, A., Yeh, B. J., and Lim, W. A. (2006). Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. *Annu Rev Biochem*, 75:655–680.

Bowman, G. D., O'Donnell, M., and Kuriyan, J. (2004). Structural analysis of a eukaryotic sliding DNA clamp-clamp loader complex. *Nature*, 429(6993):724–730.

Boxem, M., Maliga, Z., Klitgord, N., Li, N., Lemmens, I., Mana, M., de Lichtervelde, L., Mul, J. D., van de Peut, D., Devos, M., Simonis, N., Yildirim, M. A., Cokol, M., Kao, H.-L., de Smet, A.-S., Wang, H., Schlaitz, A.-L., Hao, T., Milstein, S., Fan, C., Tipsword, M., Drew, K., Galli, M., Rhrissorrakrai, K., Drechsel, D., Koller, D., Roth, F. P., Iakoucheva, L. M., Dunker, A. K., Bonneau, R., Gunsalus, K. C., Hill, D. E., Piano, F., Tavernier, J., van den Heuvel, S., Hyman, A. A., and Vidal, M. (2008). A protein domain-based interactome network for C. elegans early embryogenesis. *Cell*, 134(3):534–545.

Bradley, P., Misura, K. M. S., and Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871.

Branden, C. and Tooze, J. (1999). *Introduction to Protein Structure*. Garland Publishing.

Brannetti, B. and Helmer-Citterich, M. (2003). iSPOT: A web tool to infer the interaction specificity of families of protein modules. *Nucleic Acids Res*, 31(13):3709–3711.

Brannetti, B., Via, A., Cestra, G., Cesareni, G., and Helmer-Citterich, M. (2000). SH3-SPOT: an algorithm to predict preferred ligands to different members of the SH3 gene family. *J Mol Biol*, 298(2):313–328.

Braun, W., Wider, G., Lee, K. H., and Wüthrich, K. (1983). Conformation of glucagon in a lipid-water interphase by 1H nuclear magnetic resonance. *J Mol Biol*, 169(4):921–948.

Carroll, S. B. (2005). Evolution at two levels: on genes and form. *PLoS Biol*, 3(7):e245.

Ceol, A., Aryamontri, A. C., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. (2010). MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res*, 38(Database issue):D532–D539.

Ceol, A., Chatr-aryamontri, A., Santonico, E., Sacco, R., Castagnoli, L., and Cesareni, G. (2007). DOMINO: a database of domain-peptide interactions. *Nucleic Acids Res*, 35(Database issue):D557–D560.

Chakrabarti, P. and Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins*, 47(3):334–343.

Cheeseman, I. M., Anderson, S., Jwa, M., Green, E. M., seog Kang, J., Yates, J. R., Chan, C. S. M., Drubin, D. G., and Barnes, G. (2002). Phospho-regulation of kinetochore-microtubule attachments by the Aurora kinase Ipl1p. *Cell*, 111(2):163–172.

Chen, J. R., Chang, B. H., Allen, J. E., Stiffler, M. A., and MacBeath, G. (2008). Predicting PDZ domain-peptide interactions from primary sequences. *Nat Biotechnol*, 26(9):1041–1045.

Chen, J. W., Romero, P., Uversky, V. N., and Dunker, A. K. (2006). Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J Proteome Res*, 5(4):879–887.

Chica, C., Diella, F., and Gibson, T. J. (2009). Evidence for the concerted evolution between short linear protein motifs and their flanking regions. *PLoS One*, 4(7):e6052.

Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature*, 357(6379):543–544.

Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J*, 5(4):823–826.

Cramer, P., Bushnell, D. A., and Kornberg, R. D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science*, 292(5523):1863–1876.

Cuff, A., Redfern, O. C., Greene, L., Sillitoe, I., Lewis, T., Dibley, M., Reid, A., Pearl, F., Dallman, T., Todd, A., Garratt, R., Thornton, J., and Orengo, C. (2009). The CATH hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space. *Structure*, 17(8):1051–1062.

Davey, N. E., Shields, D. C., and Edwards, R. J. (2009). Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. *Bioinformatics*.

de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724–727.

DeLano, W. (2008). The pymol molecular graphics system. http://www.pymol.org/. DeLano Scientific LLC, Palo Alto, CA, USA.

Demarest, S. J., Martinez-Yamout, M., Chung, J., Chen, H., Xu, W., Dyson, H. J., Evans, R. M., and Wright, P. E. (2002). Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature*, 415(6871):549–553.

Dephoure, N., Zhou, C., Villén, J., Beausoleil, S. A., Bakalarski, C. E., Elledge, S. J., and Gygi, S. P. (2008). A quantitative atlas of mitotic phosphorylation. *Proc Natl Acad Sci U S A*, 105(31):10762–10767.

Dessailly, B. H., Nair, R., Jaroszewski, L., Fajardo, J. E., Kouranov, A., Lee, D., Fiser, A., Godzik, A., Rost, B., and Orengo, C. (2009). PSI-2: structural genomics to cover protein domain family space. *Structure*, 17(6):869–881.

Diella, F., Cameron, S., Gemünd, C., Linding, R., Via, A., Kuster, B., Sicheritz-Pontén, T., Blom, N., and Gibson, T. J. (2004). Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, 5:79.

DiGiammarino, E. L., Filippov, I., Weber, J. D., Bothner, B., and Kriwacki, R. W. (2001). Solution structure of the p53 regulatory domain of the p19Arf tumor suppressor protein. *Biochemistry*, 40(8):2379–2386.

Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M., and Uversky, V. N. (2005). Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J*, 272(20):5129–5148.

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. (2001). Intrinsically disordered protein. *J Mol Graph Model*, 19(1):26–59.

Dunker, A. K. and Obradovic, Z. (2001). The protein trinity–linking function and disorder. *Nat Biotechnol*, 19(9):805–806.

Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform*, 11:161–171.

Edwards, R. J., Davey, N. E., and Shields, D. C. (2007). SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS ONE*, 2(10):e967.

Encinar, J. A., Fernandez-Ballester, G., Sánchez, I. E., Hurtado-Gomez, E., Stricher, F., Beltrao, P., and Serrano, L. (2009). ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. *Bioinformatics*.

Eswar, N., Eramian, D., Webb, B., Shen, M.-Y., and Sali, A. (2008). Protein structure modeling with MODELLER. *Methods Mol Biol*, 426:145–159.

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.-P., Duewel, H. S., Stewart, I. I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.-L., Moran, M. F., Morin, G. B., Topaloglou, T., and Figeys, D. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol*, 3:89.

Ferrari, S., Marin, O., Pagano, M. A., Meggio, F., Hess, D., El-Shemerly, M., Krystyniak, A., and Pinna, L. A. (2005). Aurora-A site specificity: a study with synthetic peptide substrates. *Biochem J*, 390(Pt 1):293–302.

Fersht, A. R. (2008). From the first protein structures to our current knowledge of protein folding: delights and scepticisms. *Nat Rev Mol Cell Biol*, 9(8):650–654.

Fiaux, J., Bertelsen, E. B., Horwich, A. L., and Wüthrich, K. (2002). NMR analysis of a 900K GroEL GroES complex. *Nature*, 418(6894):207–211.

Fields, S. (2009). Interactive learning: lessons from two hybrids over two decades. *Proteomics*, 9(23):5209–5213.

Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., and Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Res*, 38(Database issue):D211–D222.

Flicek, P., Aken, B. L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Gräf, S., Haider, S., Hammond, M., Howe, K., Jenkinson, A., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Koscielny, G., Kulesha, E., Lawson, D., Longden, I., Massingham, T., McLaren, W., Megy, K., Overduin, B., Pritchard, B., Rios, D., Ruffier, M., Schuster, M., Slater, G., Smedley, D., Spudich, G., Tang, Y. A., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S. P., Zadissa, A., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Smith, J., and Searle, S. M. J. (2010). Ensembl's 10th year. *Nucleic Acids Res*, 38(Database issue):D557–D562.

Frank, J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu Rev Biophys Biomol Struct*, 31:303–319.

Freund, C., Kühne, R., Yang, H., Park, S., Reinherz, E. L., and Wagner, G. (2002). Dynamic interaction of CD2 with the GYF and the SH3 domain of compartmentalized effector molecules. *EMBO J*, 21(22):5985–5995.

Fuxreiter, M., Tompa, P., and Simon, I. (2007). Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, 23(8):950–956.

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.-A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.-M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B., and Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636.

Gerstein, M. (1998). How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold Des*, 3(6):497–512.

Gnad, F., Ren, S., Cox, J., Olsen, J. V., Macek, B., Oroshi, M., and Mann, M. (2007). PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol*, 8(11):R250.

Goll, J. and Uetz, P. (2006). The elusive yeast interactome. *Genome Biol*, 7(6):223.

Gould, C. M., Diella, F., Via, A., Puntervoll, P., Gemünd, C., Chabanis-Davidson, S., Michael, S., Sayadi, A., Bryne, J. C., Chica, C., Seiler, M., Davey, N. E., Haslam, N., Weatheritt, R. J., Budd, A., Hughes, T., Pas, J., Rychlewski, L., Travé, G., Aasland, R., Helmer-Citterich, M., Linding, R., and Gibson, T. J. (2010). ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res*, 38(Database issue):D167–D180.

Groll, M., Ditzel, L., Löwe, J., Stock, D., Bochtler, M., Bartunik, H. D., and Huber, R. (1997). Structure of 20S proteasome from yeast at 2.4 A resolution. *Nature*, 386(6624):463–471.

Guerois, R., Nielsen, J. E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol*, 320(2):369–87.

Harada, K., Makino, M., Sugimoto, H., Hirota, S., Matsuo, T., Shiro, Y., Hisaeda, Y., and Hayashi, T. (2007). Structure and ligand binding properties of myoglobins reconstituted with monodepropionated heme: functional role of each heme propionate side chain. *Biochemistry*, 46(33):9406–9416.

Harris, B. Z. and Lim, W. A. (2001). Mechanism and role of PDZ domains in signaling complex assembly. *J Cell Sci*, 114(Pt 18):3219–3231.

Harris, T. W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W. J., Cruz, N. D. L., Davis, P., Duesbury, M., Fang, R., Fernandes, J., Han, M., Kishore, R., Lee, R., Müller, H.-M., Nakamura, C., Ozersky, P., Petcherski, A., Rangarajan, A., Rogers, A., Schindelman, G., Schwarz, E. M., Tuli, M. A., Auken, K. V., Wang, D., Wang, X., Williams, G., Yook, K., Durbin, R., Stein, L. D., Spieth, J., and Sternberg, P. W. (2010). WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res*, 38(Database issue):D463–D467.

Hernández, H. and Robinson, C. V. (2007). Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nat Protoc*, 2(3):715–726.

Hillier, B. J., Christopherson, K. S., Prehoda, K. E., Bredt, D. S., and Lim, W. A. (1999). Unexpected modes of PDZ domain scaffolding revealed by structure of nNOS-syntrophin complex. *Science*, 284(5415):812–815.

Huh, W.-K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003). Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686–691.

Humphris, E. L. and Kortemme, T. (2007). Design of multi-specificity in protein interfaces. *PLoS Comput Biol*, 3(8):e164.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–4574.

Itzhaki, Z., Akiva, E., Altuvia, Y., and Margalit, H. (2006). Evolutionary conservation of domain-domain interactions. *Genome Biol*, 7(12):R125.

Jeffrey, P. D., Russo, A. A., Polyak, K., Gibbs, E., Hurwitz, J., Massagué, J., and Pavletich, N. P. (1995). Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. *Nature*, 376(6538):313–320.

Jin, J., Xie, X., Chen, C., Park, J. G., Stark, C., James, D. A., Olhovsky, M., Linding, R., Mao, Y., and Pawson, T. (2009). Eukaryotic Protein Domains as Functional Units of Cellular Evolution. *Sci Signal*, 2(98):ra76.

Kelley, R. and Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol*, 23(5):561–566.

Kersey, P. J., Lawson, D., Birney, E., Derwent, P. S., Haimel, M., Herrero, J., Keenan, S., Kerhornou, A., Koscielny, G., Kähäri, A., Kinsella, R. J., Kulesha, E., Maheswari, U., Megy, K., Nuhn, M., Proctor, G., Staines, D., Valentin, F., Vilella, A. J., and Yates, A. (2010). Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Res*, 38(Database issue):D563–D569.

Kiel, C., Beltrao, P., and Serrano, L. (2008). Analyzing Protein Interaction Networks Using Structural Information. *Annu Rev Biochem*.

Kim, H. Y., Ahn, B. Y., and Cho, Y. (2001). Structural basis for the inactivation of retinoblastoma tumor suppressor by SV40 large T antigen. *EMBO J*, 20(1-2):295–304.

Kirschner, M. and Gerhart, J. (1998). Evolvability. *Proc Natl Acad Sci U S A*, 95(15):8420–8427.

Kittanakom, S., Chuk, M., Wong, V., Snyder, J., Edmonds, D., Lydakis, A., Zhang, Z., Auerbach, D., and Stagljar, I. (2009). Analysis of membrane protein complexes using the split-ubiquitin membrane yeast two-hybrid (MYTH) system. *Methods Mol Biol*, 548:247–271.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., Onge, P. S., Ghanny, S., Lam, M. H. Y., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A., and Greenblatt, J. F. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440(7084):637–643.

Landgraf, C., Panni, S., Montecchi-Palazzi, L., Castagnoli, L., Schneider-Mergener, J., Volkmer-Engert, R., and Cesareni, G. (2004). Protein interaction networks by proteome peptide scanning. *PLoS Biol*, 2(1):E14.

Letunic, I., Doerks, T., and Bork, P. (2009). SMART 6: recent updates and new developments. *Nucleic Acids Res*, 37(Database issue):D229–D232.

Levinthal, C. (1969). *Mossbauer Spectroscopy in Biological Systems*, pages 22–24. University of Illinois Press, Urbana, Illinois.

Li, S. S.-C. (2005). Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction. *Biochem J*, 390(Pt 3):641–653.

Li, W., Young, S. L., King, N., and Miller, W. T. (2008). Signaling properties of a non-metazoan Src kinase and the evolutionary history of Src negative regulation. *J Biol Chem*, 283(22):15491–15501.

Linding, R., Jensen, L. J., Ostheimer, G. J., van Vugt, M. A. T. M., Jørgensen, C., Miron, I. M., Diella, F., Colwill, K., Taylor, L., Elder, K., Metalnikov, P., Nguyen, V., Pasculescu, A., Jin, J., Park, J. G., Samson, L. D., Woodgett, J. R., Russell, R. B., Bork, P., Yaffe, M. B., and Pawson, T. (2007). Systematic discovery of in vivo phosphorylation networks. *Cell*, 129(7):1415–1426.

London, N., Movshovitz-Attias, D., and Schueler-Furman, O. (2010). The Structural Basis of Peptide-Protein Binding Strategies. *Structure*, 18(2):188–199.

Mandell, D. J. and Kortemme, T. (2009). Computer-aided design of functional protein interactions. *Nat Chem Biol*, 5(11):797–807.

Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934.

Manning, G., Young, S. L., Miller, W. T., and Zhai, Y. (2008). The protist, Monosiga brevicollis, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. *Proc Natl Acad Sci U S A*, 105(28):9674–9679.

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, 29:291–325.

Marumoto, T., Zhang, D., and Saya, H. (2005). Aurora-A - a guardian of poles. *Nat Rev Cancer*, 5(1):42–50.

Mayer, B. J. (2008). Clues to the evolution of complex signaling machinery. *Proc Natl Acad Sci U S A*, 105(28):9453–9454.

Moses, A. M., Hériché, J.-K., and Durbin, R. (2007). Clustering of phosphorylation site recognition motifs can be exploited to predict the targets of cyclin-dependent kinase. *Genome Biol*, 8(2):R23.

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T., and Tramontano, A. (2007). Critical assessment of methods of protein structure prediction-Round VII. *Proteins*, 69 Suppl 8:3–9.

Neduva, V., Linding, R., Su-Angrand, I., Stark, A., de Masi, F., Gibson, T. J., Lewis, J., Serrano, L., and Russell, R. B. (2005). Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol*, 3(12):e405.

Neduva, V. and Russell, R. B. (2005). Linear motifs: evolutionary interaction switches. *FEBS Lett*, 579(15):3342–3345.

Neduva, V. and Russell, R. B. (2006). Peptides mediating interaction networks: new leads at last. *Curr Opin Biotechnol*, 17(5):465–471.

Neylon, C. (2008). Small angle neutron and X-ray scattering in structural biology: recent examples from the literature. *Eur Biophys J*, 37(5):531–541.

Nickell, S., Kofler, C., Leis, A. P., and Baumeister, W. (2006). A visual approach to proteomics. *Nat Rev Mol Cell Biol*, 7(3):225–230.

Nourry, C., Grant, S. G. N., and Borg, J.-P. (2003). PDZ domain proteins: plug and play! *Sci STKE*, 2003(179):RE7.

Nousiainen, M., Silljé, H. H. W., Sauer, G., Nigg, E. A., and Körner, R. (2006). Phosphoproteome analysis of the human mitotic spindle. *Proc Natl Acad Sci U S A*, 103(14):5391–5396.

Obenauer, J. C., Cantley, L. C., and Yaffe, M. B. (2003). Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, 31(13):3635–3641.

Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., and Dunker, A. K. (2005). Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, 44(6):1989–2000.

Orchard, S., Kerrien, S., Jones, P., Ceol, A., Chatr-Aryamontri, A., Salwinski, L., Nerothin, J., and Hermjakob, H. (2007). Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. *Proteomics*, 7 Suppl 1:28–34.

Panjkovich, A. and Aloy, P. (2010). Predicting protein-protein interaction specificity through the integration of three-dimensional structural information and the evolutionary record of protein domains. *Mol Biosyst*, 6(4):741–749.

Park, Y. C., Burkitt, V., Villa, A. R., Tong, L., and Wu, H. (1999). Structural basis for self-association and receptor recognition of human TRAF2. *Nature*, 398(6727):533–538.

Parthasarathi, L., Casey, F., Stein, A., Aloy, P., and Shields, D. C. (2008). Approved drug mimics of short peptide ligands from protein interaction motifs. *J Chem Inf Model*, 48(10):1943–1948.

Pawson, T. (1997). New impressions of Src and Hck. *Nature*, 385(6617):582–3, 585.

Pawson, T. (2010). Protein interaction domains. http://pawsonlab.mshri.on.ca/index.php.

Pawson, T. and Kofler, M. (2009). Kinome signaling through regulated protein-protein interactions in normal and cancer cells. *Curr Opin Cell Biol*.

Pawson, T. and Linding, R. (2005). Synthetic modular systems–reverse engineering of signal transduction. *FEBS Lett*, 579(8):1808–1814.

Pawson, T. and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science*, 300(5618):445–452.

Pawson, T., Raina, M., and Nash, P. (2002). Interaction domains: from simple binding events to complex cellular behavior. *FEBS Lett*, 513(1):2–10.

Perrodou, E., Chica, C., Poch, O., Gibson, T., and Thompson, J. (2008). A new protein linear motif benchmark for multiple sequence alignment software. *BMC Bioinformatics*, 9(1):213.

Petsalaki, E., Stark, A., García-Urdiales, E., and Russell, R. B. (2009). Accurate prediction of Peptide binding sites on protein surfaces. *PLoS Comput Biol*, 5(3):e1000335.

Pincus, D., Letunic, I., Bork, P., and Lim, W. A. (2008). Evolution of the phospho-tyrosine signaling machinery in premetazoan lineages. *Proc Natl Acad Sci U S A*, 105(28):9680–9684.

Ponting, C. P. and Russell, R. R. (2002). The natural history of protein domains. *Annu Rev Biophys Biomol Struct*, 31:45–71.

Prasad, T. S. K., Kandasamy, K., and Pandey, A. (2009). Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol Biol*, 577:67–79.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Séraphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods*, 24(3):218–229.

Pujana, M. A., Han, J.-D. J., Starita, L. M., Stevens, K. N., Tewari, M., Ahn, J. S., Rennert, G., Moreno, V., Kirchhoff, T., Gold, B., Assmann, V., Elshamy, W. M., Rual, J.-F., Levine, D., Rozek, L. S., Gelman, R. S., Gunsalus, K. C., Greenberg, R. A., Sobhian, B., Bertin, N., Venkatesan, K., Ayivi-Guedehoussou, N., Solé, X., Hernández, P., Lázaro, C., Nathanson, K. L., Weber, B. L., Cusick, M. E., Hill, D. E., Offit, K., Livingston, D. M., Gruber, S. B., Parvin, J. D., and Vidal, M. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet*, 39(11):1338–1349.

Puntervoll, P., Linding, R., Gemünd, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D. M. A., Ausiello, G., Brannetti, B., Costantini, A., Ferrè, F., Maselli, V., Via, A., Cesareni, G., Diella, F., Superti-Furga, G., Wyrwicz, L., Ramu, C., McGuigan, C., Gudavalli, R., Letunic, I., Bork, P., Rychlewski, L., Küster, B., Helmer-Citterich, M., Hunter, W. N., Aasland, R., and Gibson, T. J. (2003). ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res*, 31(13):3625–30.

Remaut, H. and Waksman, G. (2006). Protein-protein interaction through beta-strand addition. *Trends Biochem Sci*, 31(8):436–444.

Reményi, A., Good, M. C., and Lim, W. A. (2006). Docking interactions in protein kinase and phosphatase networks. *Curr Opin Struct Biol*.

Rigoutsos, I. and Floratos, A. (1998). Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics*, 14(1):55–67.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178.

Rubinstein, M. and Niv, M. Y. (2009). Peptidic modulators of protein-protein interactions: progress and challenges in computational design. *Biopolymers*, 91(7):505–513.

Rueda, M., Ferrer-Costa, C., Meyer, T., Pérez, A., Camps, J., Hospital, A., Gelpí, J. L., and Orozco, M. (2007). A consensus view of protein dynamics. *Proc Natl Acad Sci U S A*, 104(3):796–801.

Ruff, M., Krishnaswamy, S., Boeglin, M., Poterszman, A., Mitschler, A., Podjarny, A., Rees, B., Thierry, J. C., and Moras, D. (1991). Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA(Asp). *Science*, 252(5013):1682–1689.

Russell, R. B. and Gibson, T. J. (2008). A careful disorderliness in the proteome: sites for interaction and targets for future therapies. *FEBS Lett*, 582(8):1271–1275.

Sardon, T., Pache, R. A., Stein, A., Molina, H., Vernos, I., and Aloy, P. (2010). Uncovering novel substrates for the Aurora A kinase. Submitted.

Sardon, T., Peset, I., Petrova, B., and Vernos, I. (2008). Dissecting the role of Aurora A during spindle assembly. *EMBO J*, 27(19):2567–2579.

Sauer, G., Körner, R., Hanisch, A., Ries, A., Nigg, E. A., and Silljé, H. H. W. (2005). Proteome analysis of the human mitotic spindle. *Mol Cell Proteomics*, 4(1):35–43.

Schwartz, M. A. and Madhani, H. D. (2004). Principles of MAP kinase signaling specificity in Saccharomyces cerevisiae. *Annu Rev Genet*, 38:725–748.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res*, 33(Web Server issue):W382–W388.

Seet, B. T., Dikic, I., Zhou, M.-M., and Pawson, T. (2006). Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol*, 7(7):473–483.

Service, R. F. (2008). Problem solved* (*sort of). *Science*, 321(5890):784–786.

Sims, R. J. and Reinberg, D. (2008). Is there a code embedded in proteins that is based on post-translational modifications? *Nat Rev Mol Cell Biol*, 9(10):815–820.

Smialowski, P., Pagel, P., Wong, P., Brauner, B., Dunger, I., Fobo, G., Frishman, G., Montrone, C., Rattei, T., Frishman, D., and Ruepp, A. (2010). The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res*, 38(Database issue):D540–D544.

Smith, G. P. and Petrenko, V. A. (1997). Phage Display. *Chem Rev*, 97(2):391–410.

Sprangers, R. and Kay, L. E. (2007). Quantitative dynamics and binding studies of the 20S proteasome by NMR. *Nature*, 445(7128):618–622.

Sprinzak, E. and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4):681–692.

Stein, A. and Aloy, P. (2008). Contextual specificity in peptide-mediated protein interactions. *PLoS ONE*, 3(7):e2524.

Stein, A. and Aloy, P. (2010). Novel peptide-mediated interactions derived from high-resolution 3D structures. Submitted.

Stein, A., Pache, R. A., Bernadó, P., Pons, M., and Aloy, P. (2009a). Dynamic interactions of proteins in complex networks: a more structured view. *FEBS J*, 276(19):5390–5405.

Stein, A., Panjkovich, A., and Aloy, P. (2009b). 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res*, 37(Database issue):D300–D304.

Stein, A., Rueda, M., Panjkovich, A., Orozco, M., and Aloy, P. (2010). A systematic study of the energetics involved in structural changes upon association and connectivity in protein-protein interaction networks. In preparation.

Stein, A., Russell, R. B., and Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res*, 33(Database issue):D413–D417.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968.

Stiffler, M. A., Chen, J. R., Grantcharova, V. P., Lei, Y., Fuchs, D., Allen, J. E., Zaslavskaia, L. A., and MacBeath, G. (2007). PDZ domain binding selectivity is optimized across the mouse proteome. *Science*, 317(5836):364–369.

Taussig, M. J., Stoevesandt, O., Borrebaeck, C. A. K., Bradbury, A. R., Cahill, D., Cambillau, C., de Daruvar, A., Dübel, S., Eichler, J., Frank, R., Gibson, T. J., Gloriam, D., Gold, L., Herberg, F. W., Hermjakob, H., Hoheisel, J. D., Joos, T. O., Kallioniemi, O., Koegl, M., Koegll, M., Konthur, Z., Korn, B., Kremmer, E., Krobitsch, S., Landegren, U., van der Maarel, S., McCafferty, J., Muyldermans, S., Nygren, P.-A., Palcy, S., Plückthun, A., Polic, B., Przybylski, M., Saviranta, P., Sawyer, A., Sherman, D. J., Skerra, A., Templin, M., Ueffing, M., and Uhlén, M. (2007). ProteomeBinders: planning a European resource of affinity reagents for analysis of the human proteome. *Nat Methods*, 4(1):13–17.

Thiru, A., Nietlispach, D., Mott, H. R., Okuwaki, M., Lyon, D., Nielsen, P. R., Hirshberg, M., Verreault, A., Murzina, N. V., and Laue, E. D. (2004). Structural basis of HP1/PXVXL motif peptide interactions and HP1 localisation to heterochromatin. *EMBO J*, 23(3):489–499.

Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem Sci*, 27(10):527–533.

Tong, A. H. Y., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C. W. V., Fields, S., Boone, C., and Cesareni, G. (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295(5553):321–324.

Tonikian, R., Xin, X., Toret, C. P., Gfeller, D., Landgraf, C., Panni, S., Paoluzi, S., Castagnoli, L., Currell, B., Seshagiri, S., Yu, H., Winsor, B., Vidal, M., Gerstein, M. B., Bader, G. D., Volkmer, R., Cesareni, G., Drubin, D. G., Kim, P. M., Sidhu, S. S., and Boone, C. (2009). Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. *PLoS Biol*, 7(10):e1000218.

Tonikian, R., Zhang, Y., Sazinsky, S. L., Currell, B., Yeh, J.-H., Reva, B., Held, H. A., Appleton, B. A., Evangelista, M., Wu, Y., Xin, X., Chan, A. C., Seshagiri, S., Lasky, L. A., Sander, C., Boone, C., Bader, G. D., and Sidhu, S. S. (2008). A specificity map for the PDZ domain family. *PLoS Biol*, 6(9):e239.

Turk, B. E. (2008). Understanding and exploiting substrate recognition by protein kinases. *Curr Opin Chem Biol*, 12(1):4–10.

Ubersax, J. A. and Ferrell, J. E. (2007). Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Biol*, 8(7):530–541.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–627.

Ulitsky, I. and Shamir, R. (2007). Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks. *Mol Syst Biol*, 3:104.

UniProt-Consortium (2009). The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res*, 37(Database issue):D169–D174.

Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2008). Intrinsically Disordered Proteins in Human Diseases: Introducing the D(2) Concept. *Annu Rev Biophys*, 37:215–246.

Vader, G. and Lens, S. M. A. (2008). The Aurora kinase family in cell division and cancer. *Biochim Biophys Acta*, 1786(1):60–72.

Vanhee, P., Reumers, J., Stricher, F., Baeten, L., Serrano, L., Schymkowitz, J., and Rousseau, F. (2009). PepX: a structural database of non-redundant protein-peptide complexes. *Nucleic Acids Res*.

Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I., Yildirim, M. A., Simonis, N., Heinzmann, K., Gebreab, F., Sahalie, J. M., Cevik, S., Simon, C., de Smet, A.-S., Dann, E., Smolyar, A., Vinayagam, A., Yu, H., Szeto, D., Borick, H., Dricot, A., Klitgord, N., Murray, R. R., Lin, C., Lalowski, M., Timm, J., Rau, K., Boone, C., Braun, P., Cusick, M. E., Roth, F. P., Hill, D. E., Tavernier, J., Wanker, E. E., Barabási, A.-L., and Vidal, M. (2009). An empirical framework for binary interactome mapping. *Nat Methods*, 6(1):83–90.

Via, A., Gould, C. M., Gemünd, C., Gibson, T. J., and Helmer-Citterich, M. (2009). A structure filter for the Eukaryotic Linear Motif Resource. *BMC Bioinformatics*, 10:351.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403.

Wells, J. A. and McClendon, C. L. (2007). Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, 450(7172):1001–1009.

Weng, S., Dong, Q., Balakrishnan, R., Christie, K., Costanzo, M., Dolinski, K., Dwight, S. S., Engel, S., Fisk, D. G., Hong, E., Issel-Tarver, L., Sethuraman, A., Theesfeld, C., Andrada, R., Binkley, G., Lane, C., Schroeder, M., Botstein, D., and Cherry, J. M. (2003). Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. *Nucleic Acids Res*, 31(1):216–218. downloaded data in September 2006.

Willy, H., Song, F., Aung, Z., Ng, S.-K., and Sung, W.-K. (2010). SLiM on DIet: Finding Short Linear Motifs on Domain Interaction Interfaces in PDB. *Bioinformatics*.

Wilson, R. J., Goodman, J. L., Strelets, V. B., and Consortium, F. (2008). FlyBase: integration and improvements to query tools. *Nucleic Acids Res*, 36(Database issue):D588–D593.

Wright, P. E. and Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol*, 293(2):321–331.

Xu, W., Harrison, S. C., and Eck, M. J. (1997). Three-dimensional structure of the tyrosine kinase c-Src. *Nature*, 385(6617):595–602.

Yaffe, M. B. and Smerdon, S. J. (2004). The use of in vitro peptide-library screens in the analysis of phosphoserine/threonine-binding domain structure and function. *Annu Rev Biophys Biomol Struct*, 33:225–244.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabási, A.-L., Tavernier, J., Hill, D. E., and Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110.

Yun, S.-M., Moulaei, T., Lim, D., Bang, J. K., Park, J.-E., Shenoy, S. R., Liu, F., Kang, Y. H., Liao, C., Soung, N.-K., Lee, S., Yoon, D.-Y., Lim, Y., Lee, D.-H., Otaka, A., Appella, E., McMahon, J. B., Nicklaus, M. C., Burke, T. R., Yaffe, M. B., Wlodawer, A., and Lee, K. S. (2009). Structural and functional analyses of minimal phosphopeptides targeting the polo-box domain of polo-like kinase 1. *Nat Struct Mol Biol*, 16(8):876–882.

Zanzoni, A., Soler-López, M., and Aloy, P. (2009). A network medicine approach to human disease. *FEBS Lett*, 583(11):1759–1765.

Zarrinpar, A., Park, S.-H., and Lim, W. A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. *Nature*, 426(6967):676–680.