[Frontiers in Bioscience 13, 6580-6603, May 1, 2008]

Understanding eukaryotic linear motifs and their role in cell signaling and regulation

Francesca Diella[1], Niall Haslam[1], Claudia Chica[1], Aidan Budd[1], Sushama Michael[1], Nigel P. Brown[2], Gilles Travé[3], Toby J. Gibson[1]

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany, [2]BIOQUANT, Ruprecht-Karls-Universität Heidelberg, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany, [3]ESBS, 1, Bld Sébastien Brandt, BP10413, 67412-ILLKIRCH, France

TABLE OF CONTENTS

1. ABSTRACT

It is now clear that a detailed picture of cell regulation requires a comprehensive understanding of the abundant short protein motifs through which signaling is channeled. The current body of knowledge has slowly accumulated through piecemeal experimental investigation of individual motifs in signaling. Computational methods contributed little to this process. A new generation of bioinformatics tools will aid the future investigation of motifs in regulatory proteins, and the disordered polypeptide regions in which they frequently reside. Allied to high throughput methods such as phosphoproteomics, signaling networks are becoming amenable to experimental deconstruction. In this review, we summarise the current state of linear motif biology, which uses low affinity interactions to create cooperative, combinatorial and highly dynamic regulatory protein complexes. The discrete deterministic properties implicit to these assemblies suggest that models for cell regulatory networks in systems biology should neither be overly dependent on stochastic nor on smooth deterministic approximations.

2. INTRODUCTION

The paradigm of "Structure Determines Function" has offered useful guidance in protein research, at least since it has been known that these molecules can form stable folded structures. Challenges to this dogma are likely to prove futile. Nevertheless, the idealised reader that we have in mind for this review is someone who will look upon the emerging trends in cellular signaling and may like to ask themselves what these terms - structure and function - actually embody in the context of cell regulatory proteins and whether the traditional biochemical text book view is no longer adequate.

Proteins are the primary effectors of the cell - the cellular components responsible for mediating the vast majority of different functions/activities that are needed for the cell to function. An inspection of the Gene Ontology (GO) (1) gives a sense of the huge range of different functions that proteins are involved in.

Many proteins have a modular architecture in the sense that they contain a variety of globular domains, which contribute to different molecular functions such as catalysis and ligand binding. Over the last 20 years, it has gradually been established that sequences of many regulatory proteins also contain abundant short, conserved motifs constituting a second class of module that contribute to molecular functions of the proteins. This is illustrated by the well-known example of c-Src, a protein composed of three globular domains interspersed with short segments of flexible peptide, where it is shown that the roles of the domains and linear motifs are neatly linked: the kinase domain acts on linear motifs containing a phosphorylatable tyrosine, the SH3 domain binds to proline-rich peptides, and the SH2 domain binds to phosphotyrosines. The SH3 and SH2 domains switch between *cis* and *trans* peptide interactions to regulate the activity of the kinase domain in the open and closed conformations (2).

Src demonstrates common aspects of signaling activity since many protein molecules are regulated by post-translational modifications (PTM) that may mediate allosteric effects but, more often, create binding sites important for protein-protein interactions where ligand domains can bind to phosphorylated, methylated or ubiquitylated sites. Some of the earliest peptide motifs to be defined function in cell cycle regulation, hence the original definition of linear motif was provided by Tim Hunt (3).

> *"These motifs are linear, in the sense that three-dimensional organization is not required to bring distant segments of the molecule together to make the recognizable unit. The conservation of these motifs varies: some are highly conserved while others allow substitutions that retain only a certain pattern of charge across the motif."*

Then - there were few; now - there are many. And yet, we may still only know a tiny fraction of the true number and how they promote cross-talk between the signaling networks.

The sequence of amino acids determines the structure and the folding of a protein. Until recently, structural biology focused its attention mainly on the well-ordered regions (globular domains) of proteins that are typically less difficult to crystallize. The functions of globular domains have also been studied using a range of molecular and biochemical techniques (4, 5). Many bioinformatics tools are also available for studies of globular domains, including SMART, Pfam, CDD, SCOP and CATH (6-10). As a consequence there has been a tendency to avoid the unstructured proteins and regions (e.g. by removing them in expression constructs). Since linear motifs are predominantly found in regions of protein sequence that are obviously natively disordered (11), the lack of resources to study this type of protein module has hampered our understanding of their function. An important recent development has been the establishment of DisProt, a database of protein disorder (12).

While it is fairly easy to develop good detection models for globular domains based on high-quality multiple alignments, this is far more difficult for linear functional sites. The primary reason is their short length (typically 3 to 10 amino acids long) and often poorly conserved sequences. The short length of the motifs also makes them much more likely to arise/disappear spontaneously via mutations, making them evolutionarily more labile. Natural selection can thus operate, by point mutation, to fine-tune cell regulatory processes including those affecting development. The experimental verification of short functional sites is also often difficult for several reasons. Some sites are only transiently used, and thus difficult to capture by molecular methods. Even when captured, it can be difficult to determine which protein has used the site (e.g. it is often difficult to determine which kinase has phosphorylated a particular residue). It is also important to underline that the roles that linear motifs and domains play in the cellular interaction networks are quite different, not least because of the affinity with which they bind to their interacting partners. While domains can bind to each other with relatively strong affinities (in the order of nanomolar (11)), linear motifs bind with lower affinity, usually between 1.0 and 150 micromolar e.g. (13, 14).

This review will focus on the role of the linear motifs in cell regulation, protein complex formation, their classification and the new bioinformatics methods available to predict hitherto undiscovered motifs. A glossary is provided for key terms that readers may be unfamiliar with (Table 1).

3. COMPLEX AND COOPERATIVE – SIGNAL INTEGRATION AND DETERMINISTIC SIGNALING

3.1. Interactions as descriptors of protein function

The characteristics of linear motif interactions make them well suited for use in the integration of cellular signals (15). To perform their functions proteins need to interact with each other and with other components of the cell and they do this in many different ways. For example, enzymes interact with small molecules and macromolecules that are the reactants in the reactions they catalyse. Structural proteins, such as tubulin polymerise to form large structural assemblies (16). Targeting of ER resident proteins, that are retained in the ER compartment or are recycled via retrograde transport, is regulated through the interaction of their C-terminal KDEL motif with the KDEL receptor (17).

The protein-protein interactions and hence functions that arise via interactions between proteins can be categorised into several classes as summarized in Figure 1.

Domain-domain interactions typically involve large interfaces between protein domains. Many examples of this binding mode can be found in the literature and in on-line resources e.g. the DIMA database (18).

Mutual fit interactions also tend to involve relatively large interfaces. The defining property of these interactions is that the individual components can only form a stable structure in the presence of the other components. One example is the p53 tetramerisation domain (19), another is the synaptic SNARE complex, which assembles into a parallel four-stranded helical bundle (20).

Table 1. Glossary

Term	Definition
Complex epitope	Most antibodies recognize a three-dimensional surface feature that is determined by the folded structure of a protein antigen.
Deterministic engine	A deterministic engine will always produce the same results on the same input.
Discrete deterministic process	A predictable system, in which chance does not play a role, but with discontinuous elements that cannot be described by the smooth models.
Globular Domain	A region of a protein sequence that folds autonomously and possesses its own function. Sometimes used synonymously with "structural domains" or "folded domains".
Hub Protein	A term coming from interaction networks for proteins that make large numbers of interactions. These usually have substantial native disorder (known or predicted).
Induced Fit	A binding mode in which one molecule adopts a shape that fits to the template provided by the other molecule during the binding step. The term originally comes from enzyme theory.
Interaction networks	Protein-protein interaction maps have become useful aids for understanding the biological complexity of the proteome. Interaction network resources like STRING are ensembles of widely varying experimental evidence, ranging from low throughput biochemistry to large-scale yeast 2-hybrid. Not to be confused with models of signaling networks.
Intrinsically Unstructured Polypeptide	Alternative to Native Disorder. Intrinsically disordered has also been used. Common abbreviations are IUP and IUR.
Linear Epitope	Antibody recognition site determined by the linear sequence of amino acids independent of tertiary structure. Required for Western blots and other immunodetection methods.
Linear Motif	Short regions of proteins (typically peptides between 3 and 8 amino acid residues long) that embody a distinct molecular function independent of the larger sequence/structure context. Nearly always involved in regulation. The function is almost always mediated by interactions with one or more globular domain classes. Acronyms include LM, ELM and SLiM.
Modular Protein	A protein with multiple structural components, each providing separate functionality. The overall function is the aggregate of the subfunctions.
Mutual Induced Fit (or Mutual Fit)	A binding interaction mode in which neither partner provides a rigid template. Instead, the flexible partners adapt to each other during binding in the creation of a stable folded structure. When the interacting partners are natively disordered proteins, they are frequently homologous. Examples are the tetramerisation domain of p53, collagen triple-helices and coiled coils. Mutual fit is also reported for e.g. protein-RNA interactions.
Natively Disordered Polypeptide	Used interchangeably with Intrinsically Unstructured Polypeptide (IUP). Natively unfolded has also been commonly used. Terminology and definitions lack standardisation. In the widest sense it can be taken to mean all peptide regions that do not form an autonomous, stable 3D structure in the native solvent, thus including regions such as single trans-membrane helices, coiled coils, and collagen helices.
Scaffold Protein	A term coming from biochemical studies for a protein that is thought to have a major role in assembling a signaling protein complex. For example, scaffold proteins are reported for a number of kinases and their regulators. Sometimes used synonymously with "Docking protein".
Smooth deterministic process	A predictable system, in which chance does not play a role, which can be described by differential equations, e.g. the rate laws of chemistry.
Stochastic Process	A process in which chance plays a role. Probability distributions are required to simulate future evolution of the system.

Induced fit interactions can also occur between structural domains and relatively large natively unstructured regions of other proteins. The natively unstructured region is then induced to form a stable structure, but only in the presence of the interacting structural domain. An example of a large segment of induced fit is seen in the binding of a segment of SARA to SMAD transcription factors (21).

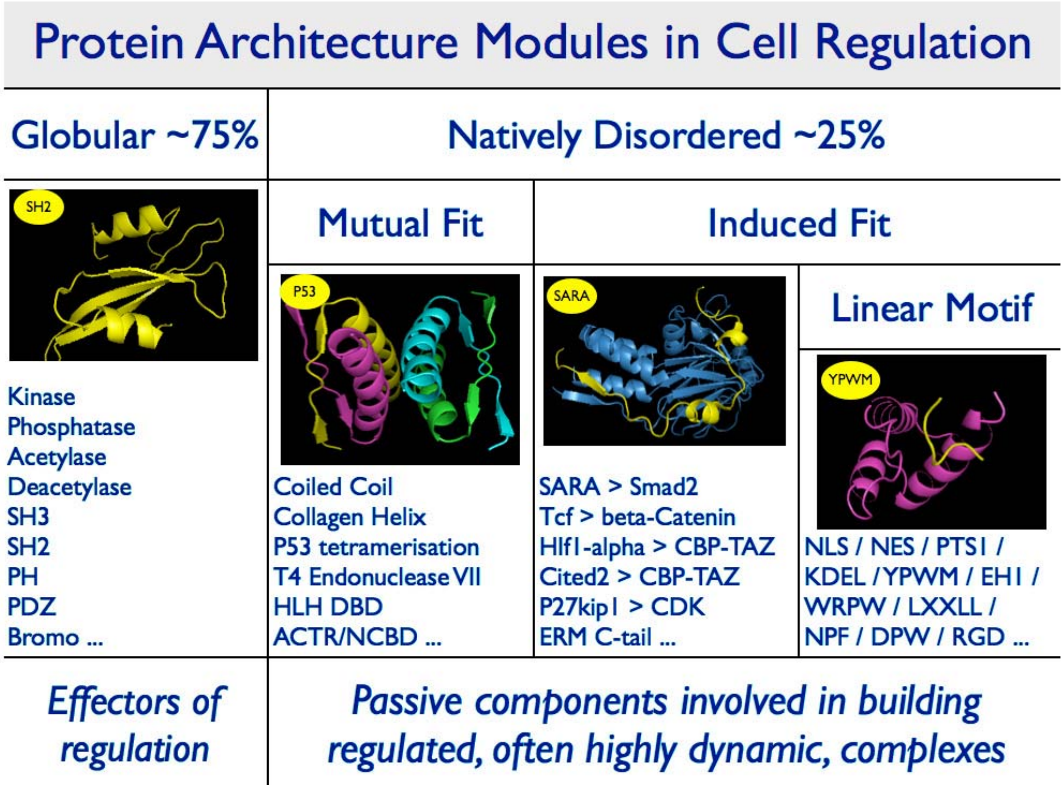
Linear motif-domain interactions are a subset of induced fit interactions, where the templated structure is induced in a short peptide of only a few residues. As a consequence of the small number of residues involved, such interactions tend to be transient and have low binding affinities. Therefore, they are well suited for mediating functions that require a fast response to changing stimuli. A classical example is the dynamic interaction between a kinase and its phosphorylation site: the kinase domain needs to bind to the peptide, attach the phosphate, and then dissociate again, potentially to then go on to interact in a similar way with other substrates. An additional feature is the richness, within a given length of sequence, of potential motif-domain interactions, which is higher than the domain-domain potential for a given length of sequence. Remarkable numbers of interactions have been ascribed to motif-rich regulatory proteins such as p53 and CBP/P300 (e.g. see their interaction sets presented in the STRING server at http://string.embl.de/). The term “hub protein” is increasingly being used to describe such proteins that are located in central positions in regulatory networks.

3.2. Cooperative binding of linear motifs

While in some cases a single linear motif interaction seems sufficient to mediate a given function, e.g. peroxisomal targeting via PTS1 (22), more often cooperativity among several motifs is required. For example, the interaction between T-cell antigen receptor (TCR) and the ZAP-70 protein-tyrosine kinase requires the cooperative interaction of immunoreceptor tyrosine-based activation motif (ITAM) in the TCR with two SH2 domains in ZAP-70 (23, 24). Cooperativity does not require direct contact between the linear motifs themselves: two (or more) independent, moderate affinity interactions combine to give a higher effective affinity in the aggregate.

Multiple linear-motif interactions can also be used to increase the specificity with which two peptides bind to each other. An example of this is in the binding of the SH3 domain of C-terminal Src kinase (CSK) to proline-enriched tyrosine phosphatase (PEP). The two proteins interact via both a conventional proline-rich SH3 recognition linear motif and an additional short hydrophobic linear motif in PEP. This additional interaction is mediated by a second binding pocket of the CSK SH3 domain different to the one used to interact with the proline-rich motif (25).

There are other ways linear motifs can enable interaction between different cellular signals. For example, a protein may contain different modification linear motifs that target the same amino acid residue for different PTMs (considering the modifier enzymes associated with the



Protein Architecture Modules in Cell Regulation

Globular ~75%	Natively Disordered ~25%		
	Mutual Fit	Induced Fit	
			Linear Motif
Kinase Phosphatase Acetylase Deacetylase SH3 SH2 PH PDZ Bromo ...	Coiled Coil Collagen Helix P53 tetramerisation T4 Endonuclease VII HLH DBD ACTR/NCBD ...	SARA > Smad2 Tcf > beta-Catenin HIf1-alpha > CBP-TAZ Cited2 > CBP-TAZ P27kip1 > CDK ERM C-tail ...	NLS / NES / PTS1 / KDEL / YPWM / EH1 / WRPW / LXXLL / NPF / DPW / RGD ...
Effectors of regulation	*Passive components involved in building regulated, often highly dynamic, complexes*		

Figure 1. Structural components of regulatory proteins. Globular domains play many roles in signaling. Interactions between globular domains require large complementary surfaces. The autonomously folded domains are expected to be outnumbered by the number of functional peptide elements in the natively disordered regions. These have two main ways of binding to other modules - induced fit when a flexible peptide binds to a well structured template and mutual fit when two flexible peptide elements together fold into a stable structure. While the distinctions are clear, in practice a given interaction might employ varying amounts of the three interaction types: For example, an interface between two globular domains might have several local regions where mutual and induced fits occur. The estimate of ~25% native disorder in the human proteome was obtained using IUPred prediction (provided by Chad Davis, EMBL).

different PTMs as inputs from different cellular signals). This could lead to competition (i.e. an interaction) between the two signals, with different enzymes competing to modify the same residue. The different PTM states of the motif resulting from this interaction would then presumably bind to different interacting domains (via different ligand linear motifs) - thus leading to different output signals from the interaction. Additionally, once a particular PTM is attached to the protein, the globular domain that recognises this PTM can form an interaction with the modified motif. This interaction may help sterically hinder interactions between modifier enzymes responsible for different PTMs that can target the same residue.

An example of such overlapping motifs is lysine 9 of histone H3 in mammalian cells. This residue may be either acetylated or methylated (26): methylation is linked to transcriptional silencing while acetylation is involved in transcriptional regulation. Overlapping linear motifs of this kind are abundant in histone tails; the catalogue of these has been called the "histone code" (27). Figure 2 shows a peptide from the N-terminus of Histone H3, tri-methylated at lysine-4, in complex with the PHD finger of BPTF (28). Methylation of this residue has been shown to be significantly more abundant in actively expressed chromatin regions, compared to other regions (29). A further example is found in the transcriptional co-activator p300. Two lysine residues within the cell-cycle regulatory region 1 of p300 (1020 and 1024) are both acetylated and sumoylated. Recent data suggests that competition between acetylation and sumoylation at these sites may be involved in regulating the activation of p300 (30). (Additional examples involving sumoylation sites can be found in reference (31)).

Another example of this kind of signal combination is the sumoylation of lysine 21 of IkappaBalpha. If this residue is not sumoylated, it is available for ubiquitylation by E3 ubiquitin ligases, targeting the protein for degradation via the proteasome pathway (32). Degradation of IkappaBalpha leads to activation of NF-kappaB signaling, an important step in the immune response to infection, along with many other external stimuli.

3.3. Protein complexes and linear motifs

Given the low affinity of linear motif interactions, they tend not to be involved in forming large, stable protein complexes (such as an RNA polymerase holoenzyme), although they often regulate the function of such complexes (as with the motif-rich CTD element of RNA Pol II). However, in a few cases, regulatory multiprotein complexes, mediated solely using linear motif interactions, have been described. A good example of this is the signaling complex assembled during TCR-induced signaling that involves the transmembrane receptor LAT, the SH2/SH3 adapter GRB2 and several other components (33). A combination of phosphotyrosine peptide ligands in LAT interacts with the SH2 domain in GRB2, while SH3 domains in GRB2 interact with proline-rich peptide ligands in SOS1 and CBL2, along with several other linear motif-domain interactions. The SH3 domain-peptide interactions cannot form a stable complex alone but must cooperate with the phosphotyrosine interactions.

In some cases, quite long-lived protein complexes contain all components of a PTM cycle; e.g. an amino acid residue is phosphorylated, the modification creates a binding motif for a target protein, then finally the phosphate is removed to stop the signaling: kinase, substrate, ligand, phosphatase can all be present in the complex for the duration of the cycle. A complex of this kind is formed between NimA-related kinase (Nek2), protein phosphatase 1 alpha, and Nek2-associated protein (C-Nap1), a substrate for both Nek2 and PP1 (34).

Another example, is the complex formed in human neurons between protein kinase A (PKA), protein phosphatase 2B (PP2B), and the A kinase anchoring protein, (AKAP79) (35). The two enzymes are known to cooperate to regulate the phosphorylation state of substrates such as AMPA receptor glutamate receptor 1 (GluR1) (36, 37).

In a third example, both mammalian and yeast MAP kinase cascades employ scaffold proteins to ensure signaling specificity and efficiency. Studies on the yeast mating pheromone pathway showed that the Ste5p scaffold protein selectively binds the MAP3K Ste11p, MAPKK Ste7p, and MAPK Fus3p complex and connects it to upstream activators (38). Furthermore, when activated, Fus3p and other yeast Map kinases have been shown by ChIP-Chip analyses to physically occupy the genes they regulate (39).

Do these kinases represent the exception or the rule? Will most protein kinases prove to have analogous scaffolded substrate interactions? Will other non-phosphorylation-based motifs also promote such organization?

3.4. Stochastic versus deterministic nature of linear motifs

Once a persistent regulatory complex has been assembled, the phosphorylation state of the bound substrate protein is not strongly influenced by the affinity of kinase and phosphatase for the substrate peptide. Instead other factors are much more important regulators, e.g. other protein-protein interactions, such as allosteric regulation of the enzymatic components, interactions with competitive inhibitors of the reaction, cellular localization of the complex or concentration of the small-molecule reactants (in this case ATP and ADP). The collective affinity of interactions that hold the PTM complex together is much higher than those of the individual interactions between each linear motif and its interacting partner. Thus, the low affinity stochastic behavior of the individual linear motif interactions could be considered overwhelmed by the high affinity of the collective interactions holding the complex together. If the effect of the collective interactions is sufficiently robust, then the system has become deterministic: the regulatory protein complex is now a deterministic engine.

A deterministic system is one that behaves predictably, i.e. given a particular input, it will always produce the same output. The rate laws of chemistry are deterministic because of the large numbers of molecules involved, so that chemical reactions become predictable to high accuracy. In a cell biological context it might be unclear whether deterministic or stochastic models should be favoured, since regulatory proteins may be present in only a few thousand molecules per cell, or sometimes just a few hundred. Biochemical pathway modeling tools therefore often offer both possibilities (40, 41).

Because of a reliance on chance events, a stochastic signaling system might be less efficient than a deterministic one. A deterministic cell signaling system requires that regulatory complexes be assembled; otherwise it will be difficult to avoid stochastic features such as Brownian motion from predominating. Then, under the normal cellular operating conditions, the set of regulatory signals input into the complex are guaranteed to produce the same result every time. If this is indeed the case, determinism is simply an emergent property of a system based on stochastic but highly cooperative signaling elements dynamically creating large protein complexes that function as deterministic signaling engines. The creation, disassembly and fusion of these complexes implies that cell signaling models should be discretely deterministic and that the smooth deterministic differential equations used to model metabolic fluxes (40, 42) may be inappropriate.

It should be obvious why cell regulation is likely to have evolved to include deterministic signaling structures: failsafe behaviors will help to guarantee cell survival. Tyrosine kinases (TKs) illustrate that random diffusion of substrate molecules can be irrelevant to regulatory function. Well studied TKs such as Src, EGFR and Abl appear to have very weak sequence specificity (43), being capable of phosphorylating any accessible tyrosine (though rates may vary for different peptide substrates *in vitro*). Yet Src and the insulin receptor do not phosphorylate the same substrate proteins: diffusion is ruled out and substrate specificity can only be achieved by complex-mediated substrate delivery.

That diffusion may play a subordinate role in cell signaling is highlighted by a recent FRET-based intracellular study of a protein tyrosine phosphatase. The PTP1B tyrosine phosphatase activity revealed remarkably

Table 2. Classification of linear motifs according to the ELM database.

Functional type	Description	Regular expression	ELM link
PTM	Sumoylation	[VILMAFP]K.E	MOD_SUMO
	N-Myristoylation	(^MG|^G)[^EDRKHPFYW]..[STAGCN][^P]	MOD_NMyristoyl
	N-glycosylation.	.(N)[^P][ST]..	MOD_N-GLC_1
Localization/Targeting	KDEL/ER retrieving	[KRHQSAP][DENQT]EL$	TRG_ER_KDEL_1
	Nuclear export signal	[DEQ].{0,1}[LIM].{2,3}[LIVMF].{2,3}[LMVF].[LMIV].{0,3}[DE]	TRG_NES_CRM1_1
	ER retention/retrieving	^M[DAL][VNI]R[RK]|^M[HL]RR	TRG_ER_diArg_1
Binding/ligand	Mapk docking site	[KR]{0,2}[KR].{0,2}[KR].{2,4}[ILVM].[ILVF]	LIG_MAPK_1
	PDZ binding motif	.[ST].[VIL]$	LIG_PDZ_1
	SH3 binding motif	[RKY]..P..P	LIG_SH3_1
Cleavage	Furin	R.[RK]R.	CLV_PCSK_FUR_1
	Proprotein convertase 7	[R]…[KR]R.	CLV_PCSK_PC7_1
	Taspase 1	Q[MLVI]DG..[DE]	CLV_TASPASE1

discrete spatial regulation for this soluble cytosolic protein (44). It was found to be most active at the endoplasmic reticulum but moderately active at the plasma membrane. The authors were not able to fit the observed PTP1B activity gradient to a smooth reaction diffusion model because the discrete spatial activity was robust to a wide range of parameterisation (enzyme concentration, substrate concentration, cell shape, etc.).

Current models of cell regulatory systems often don't include information about the non-catalytic interactions that maintain complexes between components of the system. For example, a recent report modeling the phosphorylation states of the dopamine and cAMP-regulated phosphoprotein of 32 kDa (DARPP-32) (45), which is a substrate of both PKA and PP2B, does not include models of the interaction of these enzymes with the scaffold protein AKAP79 mentioned above. System models of this kind might be improved by incorporating the non-catalytic interactions that establish complexes of the system components.

4. CLASSIFICATION AND EXAMPLES OF LINEAR MOTIFS

As described for the ELM server (46) it is convenient to classify linear motifs into four types of functional sites: ligand sites (LIG), PTM sites (MOD), proteolytic cleavage and processing sites (CLV), and sites for subcellular targeting (TRG) (for examples see Table 2). These functional assignments are useful, in that they encompass the range of peptide motif activity, but are somewhat arbitrary. For example, modification sites usually also act as ligands, while cell compartment targeting motifs are a full subset of the ligand motifs.

Other classifications are no doubt possible, for example a structural classification. About one third of liganded motifs in the ELM resource adopt a helical conformation in the bound state (even if lacking stable structure when unbound). Many of these seem to use mainly hydrophobic interactions in the binding interface. Beta augmentation of motifs accounts for another third. This is the formation of an extra strand to an already existing beta sheet in the ligand domain, as shown in Figure 2 depicting the PHD finger/H3K4 interaction. Beta augmentation is much more common than had been anticipated and sometimes involves fascinating *cis-trans* rearrangements. Since there is an excellent recent review of this interaction mode (47) it is not treated more here, but researchers interested in linear motif biology should be well informed on beta augmentation. The remaining third of linear motifs have irregular interaction topologies.

4.1. Phosphopeptide motifs and their phosphopeptide-binding domains

Phosphorylation is the most abundant PTM of eukaryotic proteins. It has gradually become clear that only a small subset of phosphosites lead to allosteric regulation of globular domain function, as exemplified by the Tyr416 of the Src kinase (48). The recent system-wide attempt to mine the proteome for phosphotyrosine motifs has revealed that two-third of known phosphotyrosine sites are involved in mediating the interaction with phosphotyrosine binding domains, e.g. SH2, whereas the remaining sites regulate processes like enzyme activity or nucleic acid binding (49). Thirteen phosphopeptide-binding domains are listed in Table 3 and there is every reason to suppose that the list will become much larger. Figure 3 shows the interaction of a phosphopeptide with the 14-3-3 domain. Of the large domain families, SH2 and PTB are currently best understood (50). WD40 is a huge domain family of central importance in linear motif biology throughout the cell. Some bind to phosphopeptides, while others bind unmodified motifs e.g. the clathrin box (51). However, there are other domain families such as BRCT and FHA that are clearly generic phosphopeptide binders. Yet, so far, only a few of their members have been investigated.

The BRCT domain of the nuclear protein BRCA1 has an important role in tumor suppression (52) and DNA damage response (53). In the human proteome, there are about 40 BRCT domains in ~20 proteins with varied domain architectures (http://smart.embl-heidelberg.de entry: SM00292); a few of these proteins have a single copy of BRCT, but mostly the domains occur as pairs or higher copy number repeats. BRCT domain pairs from BRCA1 and MDC1 preferentially recognize phosphoserine peptides (54). Many of the BRCT ligands are likely to be at pSQ motifs phosphorylated by the checkpoint kinases, ATM, ATR, DNA-PK (55). The known BRCA1-binding motifs are pS..F.K (high affinity) or pS..F (lower affinity) (54). The structure of the latter complex is shown in Figure 4. Paired BRCT domains are bound together at a hydrophobic interface and present a phosphopeptide binding surface spanning both domains.

Table 3. Protein domain families with one or more members that bind phosphopeptides.

Protein Domain	Modified Amino Acid	PDB Entry	Human proteins with domain[1]	Reference
SH2	pY	1I3Z	111/100	(50)
PTB	pY	1IRS	31/25	(50)
C2 (PKCdelta)	pY	1YRK	129/134	(145)
BRCT	pS	1TI5	21/22	(146)
FF	pS	1H40	6/4	(147)
MH2	pS	1U7F	n.a.[2]/8	(148)
SRI	pS	2C5Z	n.a./n.a.	(149)
14-3-3	pS/pT	1QJA	12/9	(150)
WW	pS/pT	1I8H	45/37	(151)
WD40	pS/pT	1NEX	264/207	(152)
Polo-box	pS/pT	1UMW	n.a./5	(153)
FHA	pT	1GXC	24/23	(154)
Cks	pT	2ASS	n.a./2	(155)

[1] The data numbers obtained from two sources: SMART_genomic_mode http://smart.embl-heidelberg.de/browse.shtml and from a domain search in the DR field of the human Uniprot/SwissProt entries (18,054 entries when surveyed). [2] n.a. = not available

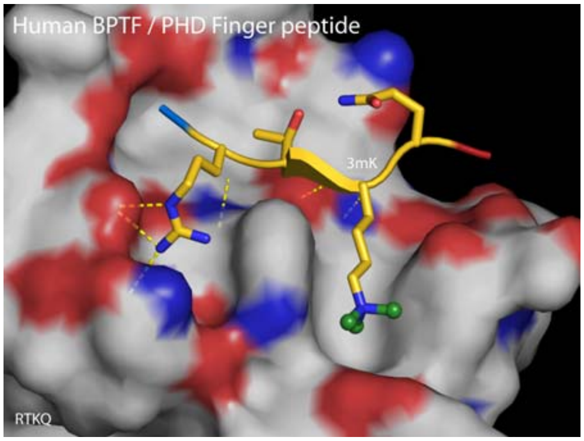


Figure 2. Reading the “histone code”. The PHD Finger of human BPTF (solid surface) bound with the histone H3 (1-15)K4me3 peptide (sticks). This motif includes a di- or tri-methylated lysine, which is a marker for active promoters. Sidechains are shown for the pattern "RTKQ" where the four sidechains make very specific interactions: in particular the trimethylated K resides in a long hydrophobic pocket that cannot accept the unmethylated lysine. Dotted lines indicate hydrogen bonds. The histone tail is known to be disordered when unbound but adopts a short local beta conformation in the binding site of human BPTF. Binding as an extra strand to an existing beta sheet is termed “beta augmentation”. See ref (28) and entry pdb:2F6J for the crystal structure. Key to structural figures. Binding domain surface (white) with oxygen (red) and nitrogen (blue); ELM peptide (gold sticks) with N- (blue) and C- (red) termini; sidechains with oxygen (red), nitrogen (blue), phosphorous (orange), and methyl groups (dark-green). Putative hydrogen bonds indicated by yellow dotted bands. Images produced using PyMOL (http://pymol.sourceforge.net/).

Perhaps the single copy BRCT domains must also dimerise: it is worth considering that their dimerisation might be aided by scaffolding proteins, which are important for many aspects of phosphorylation, and which might then add an additional regulatory interaction to BRCT phosphopeptide binding. It remains possible that some BRCT domains bind phosphopeptide as monomers. Since consensus phosphopeptide motifs have so far been defined for just a few BRCT domains, the range of different binding motif patterns could be quite large in the BRCT family, comparable to SH2 in variation.

The BRCT domain is sometimes found in association with a second phosphopeptide-binding domain, FHA. The “Forkhead associated” domain was first identified in Forkhead transcription factors but is not limited to the nucleus, being found in a wide range of signaling proteins and in several kinesins. FHA occurs in more than 20 proteins in the human proteome (http://smart.embl-heidelberg.de entry: SM00240). In contrast to BRCT, FHA is usually found as a single copy.

Most of the work on FHA domain function involves proteins with major roles in cell cycle and DNA damage response. FHA has been shown to be a phosphothreonine-binding module (56, 57). The yeast Rad53 checkpoint protein is unusual in containing two FHA domains. These have differential but interconnecting function in Rad9 activation (58) and have different sequence specificities. Rad53 FHA1 shows a preference for a large aliphatic amino acid at the pT+3 position while the FHA2 has a preference for an acidic amino acid (D or E) at this position (see Figure 5). Although weak *in vitro* binding of phosphoserine and phosphotyrosine peptides has been observed, all higher affinity FHA interactions utilize phosphothreonine, which may be an essential requirement for biological FHA ligands. The Ki67 FHA binds a larger triply-phosphorylated peptide, where one of the phosphorylations is in the sequence specific location, a second makes non-specific backbone contacts and a third makes no direct contact – yet is able to increase binding affinity by stabilizing the bound conformation of the phosphopeptide (59). This complex illustrates the point that clusters of phosphorylation sites can have a cooperative regulatory function.

4.2. Destruction motifs

The "destruction motifs" are a prime example of sequence motifs known to participate in particular biological pathways. Many proteins acting at crucial steps of the cell cycle must be rapidly degraded as soon as their task has been performed. Such proteins often contain destruction motifs that act as signals facilitating their rapid degradation at the required moment. The best characterised destruction motifs are the KEN-box (consensus KEN) (60) and the D-box (reported consensus R..L but the motif conservation may actually be R..L..[ILV]) (61). Both motifs allow the timely recruitment of the anaphase-promoting ubiquitin ligase complex APC/C, which in turn initiates ubiquitination and subsequent proteasome-mediated degradation of the protein (62, 63). Recognition of destruction boxes is performed by two proteins, Cdh1 and Cdc20, which act as co-activators of the APC/C at distinct steps of the cycle. Cdc20 joins the APC/C in early mitosis. During anaphase it is replaced by Cdh1. The D-box is recognised by both Cdc20 and Cdh1, whereas the

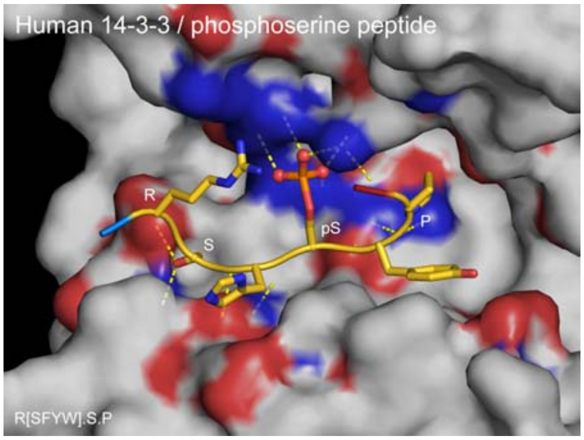


Figure 3. Phosphopeptide (pS/pT) recognition by 14-3-3 protein. 14-3-3 proteins interact with phosphoserine or phosphothreonine containing motifs. Here human 14-3-3 is shown bound to a phosphoserine containing peptide matching the consensus phosphopeptide RS.pS.P lying within, and only partially occupying, a deep groove in the protein. See ref (162) and pdb:1QJB for the crystal structure.

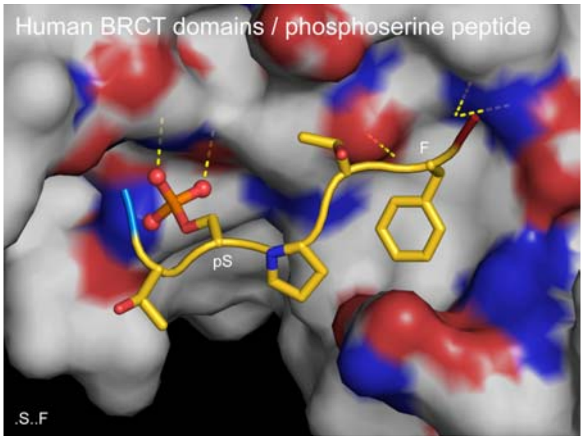


Figure 4. Phosphoserine recognition by a BRCT domain pair. The BRCA1 BRCT domains (solid surface) are bound to a phosphopeptide motif (stick representation) from the Bach1 helicase. The phosphate group of phosphoserine makes 3 strong charged interactions while the Phe nests in a hydrophobic depression. Since none of the adjacent residues make high affinity interactions, the motif is summarized as pS..F. This phospho-protein mediated interaction of the BRCT domain has a central role in cell-cycle check point and DNA repair functions. See ref (54) and entry pdb:1T15 for the complex structure.

KEN-box is preferentially recognized by Cdh1. Cdc20 itself contains a KEN box, which is recognized by Cdh1, ensuring the temporal degradation of Cdc20 and its replacement by Cdh1 as a cofactor of the APC/C (60). Some APC/C-target proteins contain only the D-box, others contain only the KEN-box, a few may contain both (62). Therefore, the D-box and KEN can act, both independently and as co-ordinated signals, for protein degradation. Instances of proteins containing active D-box and/or KEN-box can be found in the corresponding entries for these motifs, recently annotated in the ELM database (http://elm.eu.org/browse.html).

Destruction motifs are exemplary of problematic issues related to short motifs in general. Their statistical occurrence in proteins is very high and only a fraction of them will prove to be biologically active. The discrimination of "true" and "false" destruction motifs is not obvious, even on the basis of apparently straightforward experimental procedures. For instance, a protein in which a potential destruction motif has been mutated may become resistant to proteasome-mediated degradation due to serious misfolding and aggregation, and not because of losing a specific APC/C-targeting site. This stresses the requirement for discriminative criteria, which can assist or supplement the experimental approach for identification of truly active motifs. First, one can use criteria related to the particular pathway involved. Putative destruction motifs are more likely to be active if they are found in proteins related to cell cycle processes. Second, one can use criteria related to functional motifs in general. For instance, putative destruction motifs are more likely to be active when they are found in annotated cell cycle proteins, natively disordered polypeptide, and if they are conserved in the orthologous proteins. These criteria have been applied in a survey of KEN box candidates (64).

4.3. Nuclear export signal

The dynamics of protein localization between cytosol and nucleus is mainly controlled by the activity of soluble transport receptors, importins and exportins that regulate proteins shuttling into and out of the nucleus, respectively.

In the nucleus, cargoes for export form a complex with their exportin receptors and RanGTP. This trimeric complex is subsequently translocated through the nuclear pore complex (NPC). After translocation, the complex dissociates on the cytoplasmic side of the NPC through the conversion of RanGTP to the GDP-bound form. The best studied exportin is CRM1 (also designated exportin 1), which binds to proteins that have the nuclear export signal (NES), a short hydrophobic linear motif. The NES motif was first identified in the viral HIV-1 Rev protein (65) and in the cellular protein A phosphorylation inhibitor (PKI) (66). The drug leptomycin B inhibits this pathway by covalently binding to CRM1 (67). Therefore, this drug has been very useful for the identification of proteins that are exported in a CRM1-dependent manner.

A wide variety of functional NES sequences have been identified. The NES consensus sequence is reported to be #.{2,3}#.{2,3}#.# where the hydrophobic amino acids (#) are usually leucine or isoleucine (68). Upstream or downstream of the hydrophobic motif, negatively charged amino acids are common and may be required (see the ELM entry TRG_NES_CRM1_1). Paraskeva *et al.* (69) have shown that isolated HIV-1 Rev NES binds more weakly to CRM1 than the full length Rev protein. This implies that NES might require an appropriate sequence context to adopt the conformation needed to bind

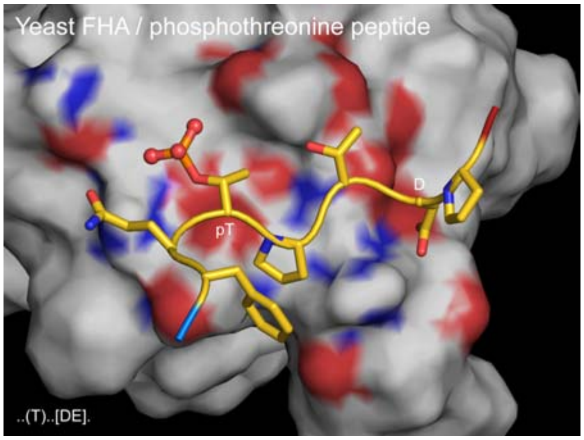


Figure 5. Phosphothreonine recognition by an FHA domain. FHA domain is a signal transduction module, which recognizes phosphothreonine containing peptides on the ligand proteins. Here the yeast Rad53 FHA1 domain (surface) is bound to a Rad9 peptide (stick) corresponding to a pT..[DE]. The surface pocket accepting the pThr samples not just the phosphate but also the gamma-methyl group, hence pSer is not a good fit. The only other strong interaction is charge complementarity with the Asp residue at position +3. See ref (163) and entry pdb:1K3N for the complex structure.

CRM1, and that the flanking regions might contribute to the binding affinity. Currently, there are no available NES/CRM1 complex structures that might explain what determines the strength of this interaction.

Another interesting feature of the NES motif is that, whereas other exportins such as CAS, Exp-t and Exp4 bind to their substrates with affinities in the nanomolar range, CRM1 binds to most NES cargoes with 100-500 fold lower affinity (69). A possible explanation of this difference is suggested to be that the NESes have evolved to maintain low affinity binding in order to avoid defects in export-complex disassembly (70). However, a complicating factor is undoubtedly the fact that the hydrophobic NES is frequently misidentified deep inside globular domains (71, 72). Of course, if just this peptide motif is taken out of the domain context and cloned into a reporter construct, it is hardly surprising that it gives a positive result. The requirement for linear motifs to be accessible is absolute and domain context must be carefully evaluated.

The unfeasible NES-like motifs deeply buried in globular domains can be contrasted with the well understood NES in MapKapK2 which conditionally folds back onto the adjacent kinase domain, thereby becoming unavailable for export signaling, depending on the nearby phosphorylation state (73). Another well defined NES lies within the p53 tetramerization domain and is only accessible in the p53 monomer (74). Thus, it is not the case that an already folded NES motif is automatically false but its structural state must be conditionally regulated. More generally, it might be very common for linear motifs adjacent in sequence to a well defined globular domain to have open and closed states that are regulated by phosphorylation and we would encourage researchers to think in these terms.

4.4. Sumoylation

The Small Ubiquitin-related MOdifiers (SUMO) are proteins that become attached to numerous substrate proteins within the nuclear compartment (75, 76). They are synthesised as inactive precursors which, after maturation and activation, are covalently linked onto the lysine residue of the linear motif #K.E found in the target proteins. This process, evolutionarily conserved in all eukaryotes, is termed sumoylation and is a reversible regulatory event (77).

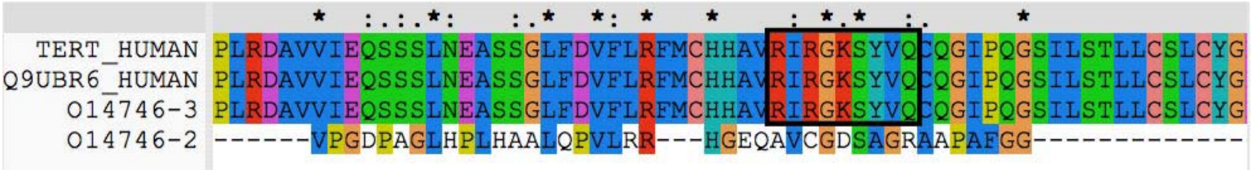
Sumoylation has been shown to have a role in transcriptional regulation (78). Unlike ubiquitination, SUMO regulation does not result in protein degradation but rather localization of proteins to specific subnuclear complexes, e.g. the promyelocytic leukemia protein PML nuclear bodies where intense transcriptional activity occurs (79). In these macromolecular complexes, proteins that are covalently linked to SUMO co-exist with proteins that recognise SUMO through a second class of linear motif, called SIM (SUMO-Interacting Motif) defined by a short hydrophobic stretch such as [IV].[IV][IV]. It has become clear that SIMs are central to the understanding of sumoylation, since they can dynamically interact with the covalently linked SUMOs (80). Protein recruitment into nuclear bodies is not the only way through which sumoylation affects transcription. In fact, it was recently reported that SUMO can directly influence higher order chromatin structure (81). Furthermore, SUMO also appears to be involved in maintaining genomic integrity, e.g. preventing the accumulation of recombinogenic structures at damaged replication forks (82).

Invertebrates have a single SUMO gene, whereas vertebrates have three: SUMO-1 and two paralogues, SUMO-2 and SUMO-3, the latter pair showing 96% similarity between themselves but only 45% with SUMO-1 (83). These three genes share some common properties, but also show different localisation and functions. RanGAP1 is preferentially modified with SUMO-1 (84), the topoisomerase II with SUMO-2 and -3 (85). Moreover, SUMO-2 and -3 themselves contain a #K.E site and can therefore form polymeric chains (86).

Nevertheless, the way paralogous-specific interactions arise is still not understood. The #K.E motif is common to all, as is the usage of only one E2 conjugating enzyme (Ubc9) for all three SUMOs. It has been suggested that variant SIMs play a role in determining the binding specificity of the SUMO paralogues (87). SUMO-specific proteases (SENPs) have also been related to the differential localisation of the three SUMOs (88). Nevertheless, the regulatory picture is far from being complete as indeed is the overall understanding of Sumoylation: so similar and yet so different from ubiquitination.

5. BEHAVIOUR OF LINEAR MOTIF SEQUENCES

Short protein motifs share the problem of overprediction with transcription factor (TF) binding sites,



```
                   *  :.:.*:   :.*  *: *   *    : *.*  :.     *
  TERT_HUMAN PLRDAVVIEQSSSLNEASSGLFDVFLRFMCHHAVRIRGKSYVQCQGIPQGSILSTLLCSLCYG
Q9UBR6_HUMAN PLRDAVVIEQSSSLNEASSGLFDVFLRFMCHHAVRIRGKSYVQCQGIPQGSILSTLLCSLCYG
    O14746-3 PLRDAVVIEQSSSLNEASSGLFDVFLRFMCHHAVRIRGKSYVQCQGIPQGSILSTLLCSLCYG
    O14746-2 ------VPGDPAGLHPLHAALQPVLRR---HGEQAVCGDSAGRAAPAFGG-------------
```

Figure 6. A missing phosphorylation site in one of the alternatively spliced isoforms of the human telomerase. Observe that the boxed site phosphorylated by the protein kinase PKB (regular expression R.R..([ST])…) is not present in the fourth sequence. The multiple sequence alignment of the known alternative transcripts was produced with ClustalW.

partly because, like linear protein motifs, TF binding sites are also short subsequences and are degenerate in the sense that the same TF can bind a set of similar but not identical sites (89). In that field, the best results have been achieved by predictive schemes that combine statistical models to represent the site (90) with phylogenetic footprinting (91).

The prediction of linear motifs has also benefited from considering evolutionary information. This has been implemented in two ways (i) focusing on the conservation of certain instances during the process of sequence divergence (92), and (ii) identifying convergently evolved patterns in a set of sequences that share a functional trait, like a common interacting partner (93, 94).

Multiple sequence alignments of related protein sequences can provide insights into the evolutionary dynamics and characteristics of linear motifs. Examining regions of these alignments known to contain functional instances of linear motifs reveals that individual instances are rarely conserved in all related sequences. Focusing on amino acid positions within conserved motifs, it is often the case that some positions within the motif show no variation in amino acid residue. Other positions vary, but only between a limited set of amino acids (typically those with similar physico-chemical properties). A final class of positions appears to accept any possible amino acid. This pattern of conservation, is a result of the fact that only some positions are important for linear motif function. Some of the side chains of residues are important for the interaction, these are the positions where no, or only limited, variation of amino acid residues is observed. Other residues function as spacers, or by providing main-chain interactions, and they are found to accept a wide range of amino acid residues. The molecular role of linear motifs, in a regulatory context, is determined by the inherent flexibility of the surrounding residues and the disordered and evolutionarily variable context in which the motif is situated actually enables recognition with low affinity (11).

Gain or loss of a linear motif is likely to happen by single point mutation. This evolutionary plasticity makes linear motifs into small modular units that can tune the evolution of protein function once the globular domain architectures have been established. This would explain the convergent appearance of the same linear motif in unrelated proteins with similar function. There is a limit to this plasticity since linear motifs contain a functional value that cannot be gained or lost purely by chance as linear motifs are subject to selection.

There are some situations that make proteins less sensitive to linear motif loss. Paralogous proteins tend to lose motifs quite easily relative to orthologous sequences (95). Recurrent motifs can present different numbers of copies in different species (e.g. the DPW and NPF motifs in Epsins). Additionally, the splicing process adds even more complexity to linear motif evolution, since isoforms of a protein can differ in whether or not they contain a motif. There are a few known examples of motifs that are present in some, but not all the alternative transcripts. This is the case of the phosphorylation site MOD_PKB_1 in the catalytic subunit of the human telomerase shown in Figure 6.

The ELM database contains experimentally validated motifs, the majority of annotations being drawn from human, mouse and other vertebrates. The majority of these (66%) are conserved only within vertebrates. However, a significant proportion (22%) is conserved across vertebrates, plants and yeast (Chica *et al.*, in press). In a signaling system multiple signals may act in concert to regulate the behavior/output of the system. In a cell, this could be the regulation by modification of multiple linear motifs simultaneously. The strong conservation of some motifs across such a wide range of phyla suggests that they may be involved in core pathways. In contrast, where motifs are conserved only in a clade such as the vertebrates this is likely to be a result of the combinatorial composition of the network, and how regulation has been tuned during evolution. This can act both to release evolutionary pressure on existing motifs and select for new ones. Overall, the evolutionary stability of a linear motif is a function of its importance in the cellular network, and how that network has been evolving.

The evolutionary behavior of linear motifs becomes an issue when designing computational tools to predict them. This is illustrated when using multiple sequence alignments (MSAs) to assess the extent of conservation of motifs within equivalent positions in a set of related proteins. In general, MSA software is not trained with alignment of linear motif like regions in mind – focusing instead on accurate alignment of globular regions. This results in sub-optimal performance of MSA algorithms in disordered protein regions (Thompson, personal communication). Multiple sequence alignment algorithms give different results in some extreme cases, like repetitive and C-terminal motifs. Figure 7 compares the alignment of these kinds of motifs, calculated with MAFFT and ClustalW. For these cases, the former manages to use the motifs as small anchors inside the disordered regions more often than the latter.

6. METHODS TO PREDICT LINEAR MOTIFS

Until very recently, few tools were available for bioinformatical analysis of short functional sites; only in the last few years has there been some progress in the development of computational approaches to identify linear motifs. The main problem is the difficulty of discriminating between true and false positives, as a consequence of the complex pattern of evolutionary conservation and the lack of statistical significance due to the short length of the linear motif. The linear motif predictors can be divided in two categories: firstly, methods aimed at identifying new instances of already known linear motifs on protein sequences; secondly, methods focused on discovering new / *de novo* linear motifs. Table 4 lists a range of different methods of both categories of predictor.

Early attempts to predict short motifs in protein sequences involved describing a 'consensus' sequence that capture the signature of highly conserved residues in the motif. PROSITE (http://www.expasy.ch/prosite/) (96) made the first systematic attempt to catalogue known motifs. The PROSITE database has collected a number of linear protein motifs, representing them as regular expression patterns. PROSITE patterns have been very useful, but also suffer from severe over prediction problems and more recently the database has focused on protein signature and domain annotation.

Nevill-Manning *et al.* (97) implemented a method based on automatically constructing consensus sequences in order to identify motifs from families of aligned protein sequences. Software based on their methods, eMOTIF (98), has been implemented although it is currently unavailable.

SCANSITE (http://scansite.mit.edu/) (99) is a web-accessible tool that predicts motifs important in cellular signaling such as phosphorylation motifs or peptides binding to SH2 domains, 14-3-3 domains or PDZ domains. Each sequence motif is represented as a position-specific scoring matrix (PSSM) based on results from oriented peptide library and phage display experiments.

The Eukaryotic Linear Motif (ELM) resource (http://elm.eu.org/) (46) stores manually curated information about known linear motifs: it combines the use of regular expressions with logical filters (or rules), based on contextual information, to discriminate between likely true and false positives in order to improve the predictive value of ELM. The currently implemented context filters are a) taxonomic range filter, b) cell compartment filter, c) globular domain filter. In addition known ELM instances and predictions in sequences similar to ELM instance sequences, where the motif is positionally conserved, are identified and displayed.

A similar resource has been subsequently developed by Balla *et al.* (100). The Minimotif (MnM) database (http://sms.engr.uconn.edu) contains 312 minimotifs extracted from the literature and other online resources and a web-based simple motif search (SMS) system for identifying linear motifs in proteins. Homology analysis, surface prediction and frequency scores in complete proteomes are used to estimate the probability that the identified minimotifs are biologically functional.

The identification of signature-like putative functional sites has been recently implemented by Ben-Tal and his group (92). This novel method implemented in the QuasiMotiFinder program, uses motifs and signatures as defined by PROSITE. Each putative motif is assigned scores that depend a) on its physico-chemical similarity with respect to the original motif; b) on the degree of evolutionary conservation within homologous sequences. The total score is used to calculate the statistical significance of the putative motif.

The AutoMotif Server (AMS) (http://ams2.bioinfo.pl/) (101) predicts PTM sites in proteins based only on sequence information. The PTMs are taken from the Swiss-Prot and ELM databases and sequence models for all types of PTMs are trained by support vector machine.

In addition there are increasing numbers of specific predictors for individual PTMs and protein sorting motifs. In Table 5 we have collected some representative sites. The ExPasy proteomics tools page (http://www.expasy.ch/tools/) is a good place to start looking for these resources.

As sequence database annotation becomes more extensive it has become clear that motifs can be enriched with certain keywords with good statistical significance. Copley used transcriptional keywords to detect new examples of the EH1 transcriptional repressor motif (102). Researchers can undertake equivalent motif/keyword explorations interactively using SIRW (http://sirw.embl.de/index.html) (103). This resource was used to demonstrate that KEN motifs are significantly enriched with cell cycle keywords and GO terms. The candidate list was further refined using native disorder prediction and phylogenetic conservation scoring (64). We anticipate that this pipeline - keyword enrichment, disorder prediction and conservation scoring - will become widely applicable in motif discovery.

A more difficult task is *de novo* linear motif discovery: their short length, high flexibility and low-binding affinities make them awkward, both for experimental and *in silico* analysis. In addition, the available training data sets are far from comprehensive and make it hard to develop fully automated discovery algorithms. Our current knowledge in the linear motif field is still poor: somewhat more than a hundred classes of linear motif are known in eukaryotes but it has been estimated (14) that hundreds of motif classes mediating protein interaction have still to be discovered. However, in the last few years many proteome-scale interaction data sets have became available and this has allowed the implementation of a new generation of bioinformatics tools to discover *de novo* linear motifs involved in protein interactions.

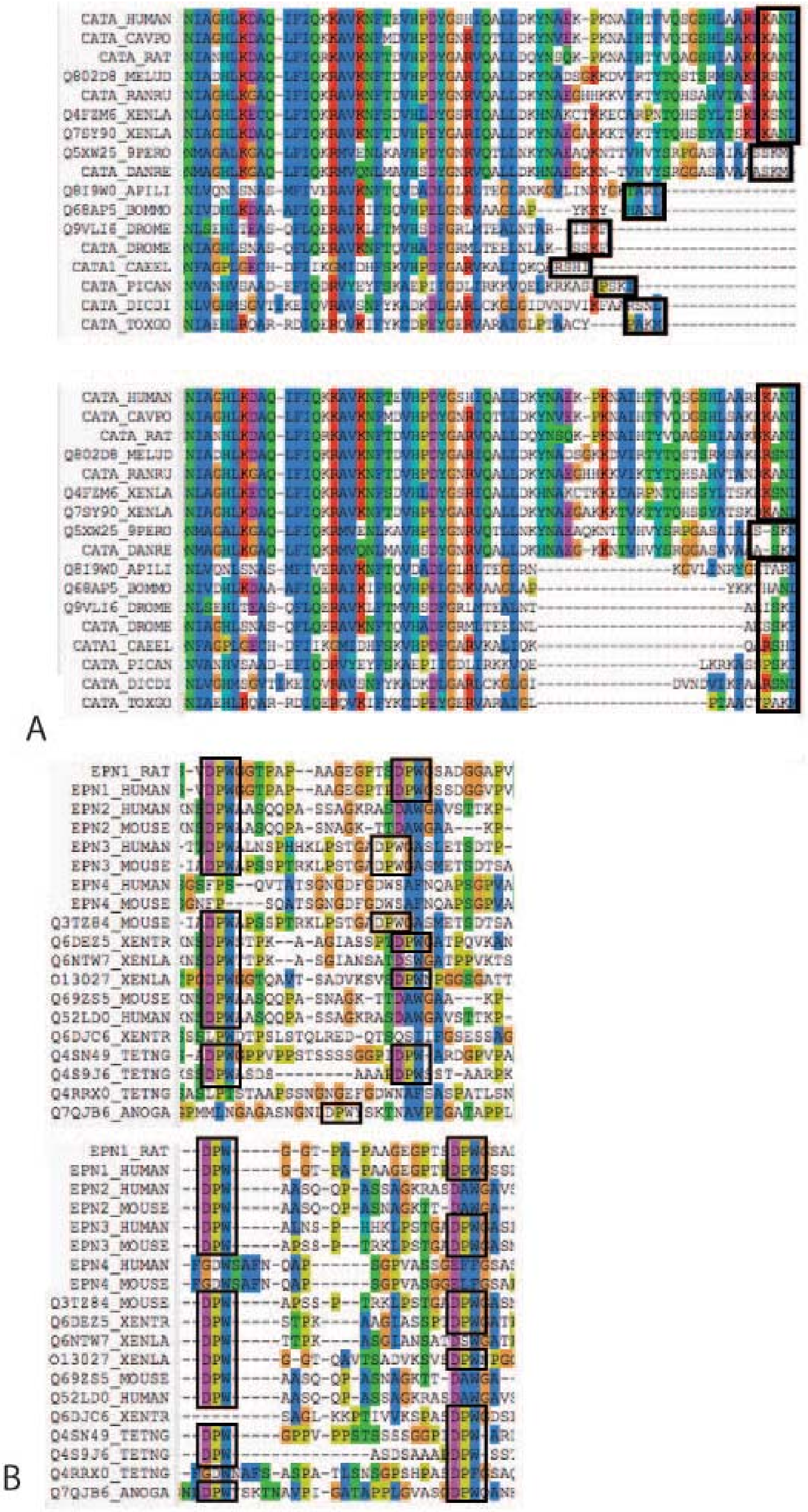


Figure 7. (A) Multiple sequence alignment of Catalase C-termini. The peroxisomal targeting motif PTS-1 is often misaligned by ClustalW (top) but not by MAFFT (below). (B) Misalignment of repetitive DPW motifs in Epsin. On the top, the multiple sequence alignment was calculated with ClustalW (164), where several motifs are unaligned. Below, the same sequences as aligned by MAFFT (165), where more instances are correctly aligned.

Table 4. List of methods to predict new instances and *de novo* motifs.

	Server	Method	Advantage	Disadvantage	url or reference
METHODS TO IDENTIFY NEW INSTANCES OF KNOWN MOTIFS	**PROSITE**	Regular expression matching	First attempt to catalogue LM	Stopped adding motifs due to high number of false positive matches. Currently the main focus is on globular domains	http://www.expasy.ch/prosite
	SCANSITE	Profile-based methods that uses data coming from oriented peptide library technique	>60 motifs. Quantitative representation of patterns is suitable for measuring features like motif specificity	Restricted to phosphorylation sites and motifs involved in signaling	http://scansite.mit.edu/
	ELM	Regular expression matching plus contextual filtering	Context-based rules and logical filters reduce the amount of false positives. >130 motifs are manually curated	Incomplete coverage of known motifs	http://elm.eu.org/
	Minimotif Miner	Regular expression matching plus contextual filtering	Large number of regular expressions	Motifs have little extra annotation	http://sms.engr.uconn.edu
	QuasiMotiFinder	Matching of patterns similar to PROSITE signatures plus evolutionary filtering	Evolutionary filtering reduces number of false predictions	Restricted to the set of motifs in PROSITE	(92)
	AutoMotifServer	Prediction of motifs based on trained support vector machine (SVM). Each type of PTM trained separately	The server predicts a good number of PTMs not present in other resources	The score assigned to the predicted instances is not biologically significant	http://ams2.bioinfo.pl/
	SIRW	Combine Regular expression with keyword search	Very intuitive method for prediction of new instances. Enrichment with GO terms can provide significant support	Low throughput interactive method	http://sirw.embl.de/
DE NOVO MOTIF DISCOVERY	**DILIMOT**	Identification of over-represented motifs in a set of proteins interacting with a target protein	First attempt at de novo motif prediction. Authors themselves found and tested new motifs	Only applicable to proteins present in interaction datasets. Only returns identities at motif conserved positions	http://dilimot.embl.de/
	SLiMFinder	Identification of over represented motifs in set of proteins, typically the set is an interaction dataset	Is able to retrieve motif matches with semi-conserved positions	Mainly applicable to proteins present in interaction datasets	http://bioinformatics.ucd.ie/shields/software/slimfinder/
	D-MOTIF/D-STAR algorithm	Detection of correlated (co-occurring) short sequence motifs	Improve detection from sparse and noisy interaction data	Rather stringent	(105)

DILIMOT (DIscovery of LInear MOTif) (http://dilimot.embl.de/) (93) is a method for *de novo* discovery of linear motifs within a set of proteins. The method detects over-represented short sequences in a set of sequences that share a common functional characteristic (e.g. interaction partners). Since most linear motif sequences are found in unstructured regions, parts of the sequences of the set of protein (e.g. coiled coil regions, globular domain) are discarded before starting the motif search. Finally the conservation of the motif in the orthologous protein is calculated and the statistical significance is assessed.

SLiMFinder (http://bioinformatics.ucd.ie/shields/software/slimfinder/) (104) is a method for finding potential shared motifs in unrelated proteins using a model of convergent evolution and, for the first time, assign a significance value to each motif. SlimFinder is comprised of two algorithms: a) SLiMBuild that identifies convergently evolved LM in a dataset of unrelated proteins b) SLiMChance that calculates a significance value associated with such motif predictions.

Both programs can find statistically over-represented motifs in non-homologous sequences; the main difference between the two methods is that DILIMOT masks all but one arbitrarily selected homologous protein prior to motif discovery, whereas SLiMFinder searches for motifs in all proteins and then weights results according to the evolutionary relationship of the proteins containing the motif. Moreover SLiMFinder is able to predict patterns with semiconserved positions.

Tan *et al.* (105) proposed a novel approach of mining correlated de-novo motifs from interaction data based on a MTM (many to many) comparison. This approach aims overcome some of the problems of the sparse and noisy nature of the interaction data sets within which only a limited number of interactions are observed for many proteins. Their model implies that interactions are mediated by pairs of motifs co-occuring in separate

Table 5. Example of some specialized predictors

Predictor	PTM	Link	Reference
NetPhos	Generic phosphorylation sites in eukaryotic proteins	http://www.cbs.dtu.dk/services/NetPhos/	(156)
NetCGlyc	C-mannosylation sites in mammalian proteins	http://www.cbs.dtu.dk/services/NetCGlyc/	(157)
ChloroP	Chloroplast transit peptides and their cleavage sites in plant proteins	http://www.cbs.dtu.dk/services/ChloroP/	(158)
SulfoSite	Predicts tyrosine sulfation sites in protein sequences	http://sulfosite.mbc.nctu.edu.tw/	(159)
Sulfinator	Predicts tyrosine sulfation sites in protein sequences	http://expasy.org/tools/sulfinator/	(160)
Myristoylator	Predicts N-terminal myristoylation of proteins	http://expasy.org/tools/myristoylator/	(161)

interacting proteins. The use of a correlated (co-occurrence) motif pairs approach has the advantage of increasing the number of motifs that can be detected of being more stringent compared to the other approaches and in addition not requiring any prior knowledge of protein grouping.

7. NATIVELY DISORDERED PROTEIN AND PREDICTION METHODS

Until recently the dominant paradigm in protein function prediction has been relating the 3-dimensional structure to the protein's function (106). The realisation that intrinsically unstructured polypeptide (IUP) regions could have a functional role has lead to the development of disorder prediction methods. In contrast to the well established field of protein structure prediction, prediction methods for intrinsic disorder are less well developed. However, many of the strategies applied to protein structure prediction have been retooled for disorder prediction. Echoing the first secondary structure prediction methods, the simplest disorder prediction methods use statistical prediction methods like propensity (107) or some measure of the physico-chemical properties of the amino acids (108). More sophisticated disorder prediction methods use artificial intelligence methods like neural networks in the same way as the most successful of the protein structure prediction algorithms (109). In general, the most recent programs outperform the earlier methods: a key factor has been the establishment of the DisProt IUP database (12), which was not available to the earlier generation of predictors.

7.1. Definition-based predictions

GlobPlot defines a propensity for a particular amino acid to be in either an ordered or a disordered region of a protein (107). This is plotted as a running sum (with a smoothing window) such that the slope of the graph indicates order or disorder (Figure 8A). The propensity to be in an ordered or disordered state is based on data from SCOP. In SCOP, amino acids are characterised as being part of structured regions or random coils. Therefore, the default GlobPlot propensity scale measures the difference in the propensity for an amino acid to be in the ordered set (i.e. defined secondary structure part of SCOP) versus being in the disordered set (i.e. in the random coil part of SCOP). In addition to the default definition (known as the Russell/Linding definition) GlobPlot provides other propensity scales based on other data sets. These scales are all available on the GlobPlot site for the user to explore. All of them are attempting to capture the propensity for a given amino acid to be in an ordered or a disordered region. In general, they correlate well with physico-chemical properties of the amino acids, for example, in Figure 2 of reference (107) the propensities are shown against hydropathy.

7.2. Physico-chemical predictions

One way or another, most prediction methods attempt to encapsulate the physico-chemical properties of the amino acids and use this to predict the likelihood of a particular residue to be in an ordered or a disordered region. One of the first, by Uversky *et al.* is simply a measure of the hydrophobicity of the protein (108). This method successfully manages to distinguish between ordered and disordered sequences by averaging the hydrophobicity over the length of the protein and plotting this against the net charge of the protein. The result is a function that often distinguishes between primarily ordered and disordered proteins. For a clear example see Figure 3 in reference (108) where the plot of the net charge of the protein as a function of the hydrophobicity of the protein produces a clear separation between ordered and disordered proteins. More recently, this method has been updated and implemented as FoldIndex (110). The use of a sliding window then allows the prediction of disorder for segments of protein sequence. FoldIndex takes the original equation of Uversky *et al.* and adds constants to the linear equation used to distinguish between ordered and disordered proteins in the net charge versus hydrophobicity plot. The addition of these constants to the equation transforms the scale from 1 to 0 to a score where positive indicates ordered and negative indicates disordered.

IUPred intrinsically captures physico-chemical information in its algorithm (111). The assumption underlying the method in IUPred is that amino acids in disordered regions tend to be those that do not have the capacity to form many interactions with other residues (112). Inter-residue interactions are responsible for forming stable 3-dimensional structures. The sliding window approach is used in IUPred to generate a potential for the residues in the window to form stabilising interactions. The potential is based on a statistical method for calculating the potential for the interactions. The sum of the interaction energies in the window is a function of the amino acid composition. This function embodies the chemical type of the amino acids and crucially, the potential for them to form interactions (112). The order-disorder tendency of the p53 sequence plotted by IUPred is shown in Figure 8B.

7.3. Artificial Learning prediction

The data from the DisProt resource has been used to analyse the frequency of different amino acids in the disordered polypeptide segments and used to make predictions about the likelihood that a sequence will be found in ordered or disordered conformations. A wide range of properties of amino acids has been applied to

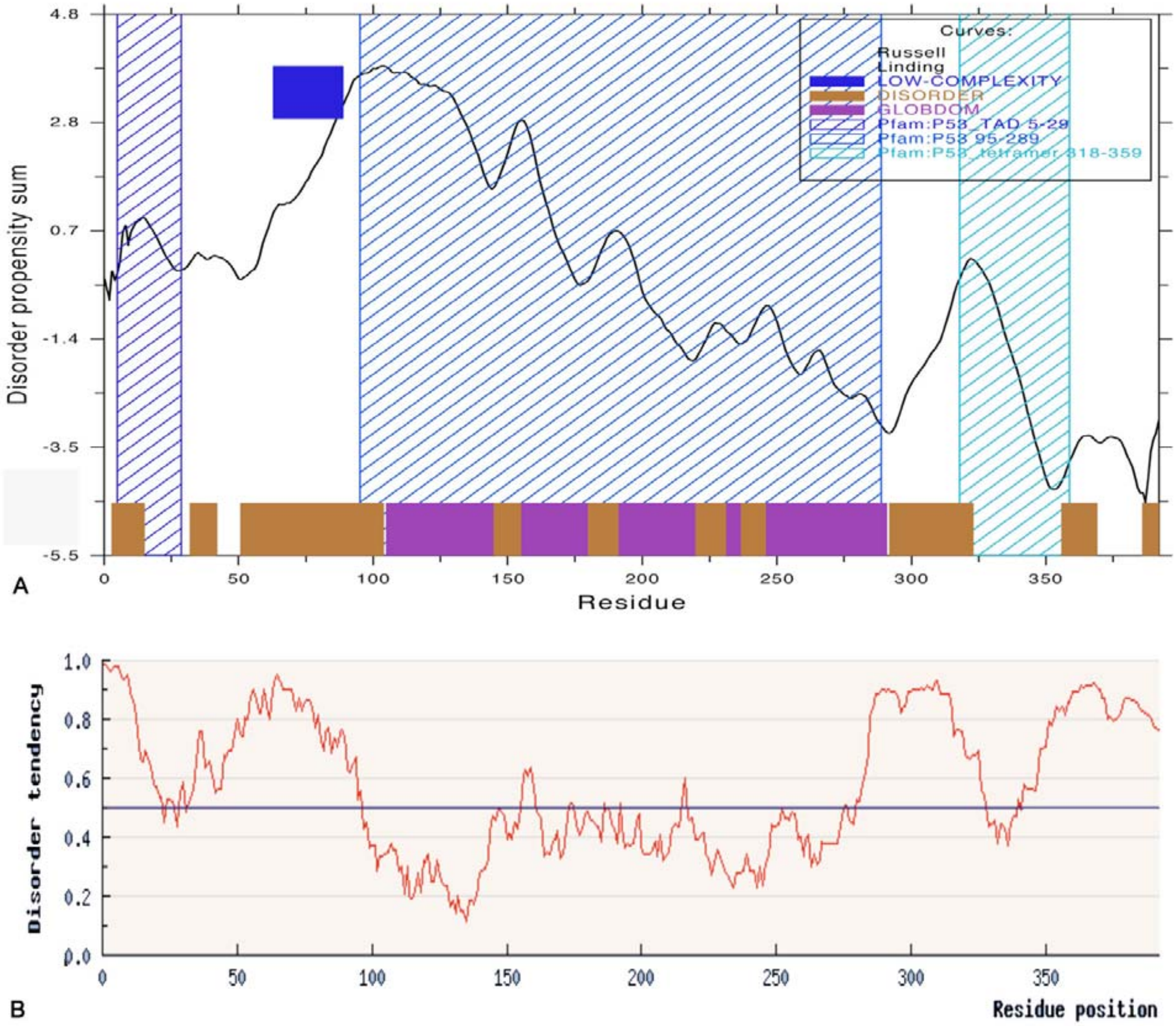


Figure 8. Output of two popular disorder predictors for the sequence of protein p53. Both plots were taken using the default parameters of GlobPlot (A) and IUpred (B). For (A) an upward slope indicates disorder. For (B) a positive score indicates disordered regions. The hashed regions in (A) indicate Pfam entry annotations of which only the central DNA binding domain is a true globular domain. Numerous studies have revealed that the N- and C-termini are natively disordered, though short segments are known to make induced fit interactions while the tretramerisation module self-associates by mutual fit (166-169).

classify the amino acids as order promoting or disorder promoting. As an example, one of the best predictors was a residue contact scale (113). Other high performers included hydrophobicity scales. DISOPRED also uses artificial learning techniques to classify protein sequences into ordered and disordered regions (114). The data set at the heart of this method is a set of non-redundant protein sequences with high resolution X-ray structures. Disordered residues are indicated by the lack of coordinates in the structure - similar to some of the scoring schemes in GlobPlot. However, rather than relying on statistical differences in the propensity of different amino acids to be in ordered or non-ordered regions, the DISOPRED server defines a Support Vector Machine using a window of 15 residues to predict the likelihood of order.

There are many other native disorder predictors that perform well, such as RONN, DisProt and FoldUnfold (115-118). A useful list is maintained at the DisProt site (http://www.disprot.org/predictors.php). The general recommendation is to build up an idea of a protein's structure by using several of the disorder predictors and cross-compare to domain databases.

There are now more than 500 publications on natively disordered proteins (119). Protein disorder prediction has come of age and should now be used routinely in any bioinformatics analysis of protein sequence, whether on the small or large scale. It is an essential adjunct to linear motif biology and researchers in transcription, cell signaling and other aspects of cell regulation need to be aware of the role of protein disorder in cell regulation. As well as the primary literature, there are a number of recent reviews that go into the topic in more depth than is covered here and are excellent points of entry e.g. Romero *et al.* (120; 121). These reviews can also help to bring the concept of native disorder into undergraduate teaching courses since the topic is neglected in text books (with the honorable exception of Garrett (122)).

8. ANTIBODIES AS TOOLS FOR THE EXPERIMENTAL INVESTIGATION OF MOTIFS

The success of the computational methods for the discovery of linear motifs depends on good experimental data used to derive the data sets that these methods use as their basis. The experimental determination of linear motifs has been greatly aided by the use of antibodies. In particular, antibodies have successfully been applied to the detection of phosphorylated residues and phosphorylation motifs in proteins (95). They are used primarily because they can detect the surface exposed motifs. Typically this involves raising an antibody to a peptide that expresses the motif and then using this antibody to test the surface exposure or accessibility of the motif (for an example of this technique see (123)). The precise structure, or lack thereof, of bound peptides and motifs to domains or antibodies is not known in the majority of cases. Whilst linear motifs are found in Intrinsically Disordered Regions (IDR) it has been suggested by many authors, that motifs may form induced fit structures (for a good review see (124)). The P..P motif structure, for example, is found in an IDR however upon binding it forms a left handed helix (125). It is clear that there exists a range of states, from stable structures through to stable unstructured proteins, encompassing induced fit and transient structures. Linear motifs tend to occur towards the unstructured end of this scale.

In one example of such a strategy, Kikuchi *et al.* developed an antibody that recognises an N-glycosylation site (126). The antibody is capable of recognising the native motif but not the denatured protein. Phage display was used to isolate the trimer motif specifically recognised by the antibody (126). Alternatively, degenerate peptide libraries can be used to isolate the motif. For example, raising an antibody to recognise the phosphorylated form of the PKA substrate consensus sequence RR.T* requires a library containing the sequence CxxxxxRRxT*xxxx; where x is any amino acid (127). The antibody is exposed to the peptide library, and specifically binding antibodies are purified. These purified antibodies are expected to selectively bind to the phosphorylated T with R at positions -2 and -3. They do not bind to the non-phosphorylated form or to peptides without the R at positions -2 and -3. The motif can be validated by constructing conjugates of the motif and another protein – typically Bovine Serum Albumin. The antibody is exposed to both the conjugate and the protein without the motif. The antibody should specifically bind to the conjugate but not to the protein without the motif. The antibody can then be used in assays to see if the motif is present in other proteins, perhaps when the sequence of the other proteins in a sample is not known. Critically in this example, the motif recognition is context dependent - i.e. the antibody only recognises the modified motif. Once these antibodies have been raised they can be used, for example, to detect substrates of particular families of kinases.

The study of signaling pathways, and in particular the role of phosphorylation can be carried out using antibodies that recognise particular motifs, for example the motif recognized by PKB and PKC kinases in reference (128). Antibodies specific to different phosphorylation motifs are used to trace signaling pathways by identifying substrates of specific kinases (128). Kinases are a major drug target and therefore the development of any tools that are likely to increase the range of targets for kinase inhibitors is likely to be of interest to the drug development industry. Antibodies can be raised against known substrates of a particular enzyme and then used to identify unknown substrates (127), assuming that the substrates all share the same recognition motif. Peptide arrays can also be used to determine the motif that the antibody recognises and the specificity of the recognition. Once an antibody has been raised for a particular motif, this antibody can be exposed to the substrate protein to see if they react. Different antibodies can be raised to the phosphorylated and unphosphorylated substrate. For example, antibodies against phosphotyrosine that can discriminate between phosphorylated and non-phosphorylated tyrosine and are broadly reactive against any phosphotyrosine containing proteins have been very successful in studying intracellular signaling mechanisms (127). Phospho-specific antibodies have dramatically increased the rate at which phosphorylation events can be studied in the cell (129). Antibodies can be raised using degenerate peptide libraries of the substrate motif. The motif can then be worked out using the kinase substrate antibody matrix which is based on the relative frequency of amino acids at each position (128), in a method similar to oriented peptide libraries (130). A range of motif discovery tools such as ScanSite, DILIMOT and SlimFinder, as well as motif databases such as Phospho.ELM (131) and ELM (46) can then be used to help reveal the antigenic or phosphorylated sites in the protein. Linear motif detection experiments can be parallelised by the use of Flow Assisted Cell Sorters (FACS) and fluorescent labeling of the antibodies (129). To investigate cell signaling in cancer, Irish *et al.* used FACS and a set of labeled antibodies raised against phosphorylated motifs using a library of synthetic peptides (132). This process forms the basis of a number of patents that describe the process of raising antibodies to motifs important for the regulation of kinase substrates (127).

Much work has been done on the experimental validation of linear motifs. The most common approach is to use deletional analysis to attempt to alter the phenotype. The Cryptochrome (CRY) protein has been studied in a number of species, notably by Hemsley *et al.* in *Drosophila Melanogaster* (133). The CRY protein contains a large domain similar to other photolyases. In Drosophila there is a second C-terminal putatively disordered domain unique to Drosophila. This second region has previously been

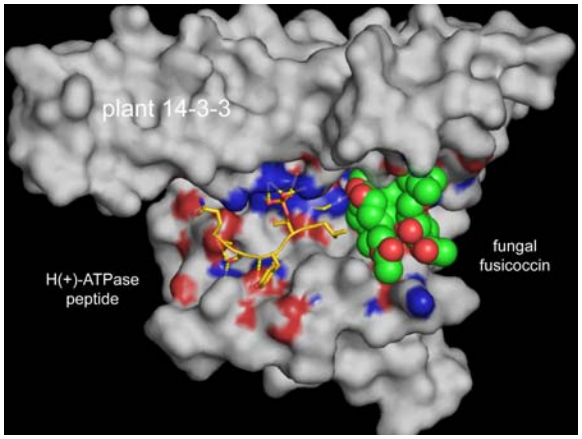


Figure 9. Toxin and drug interactions. A 14-3-3 protein from tobacco (*Nicotiana tabacum*) is shown with a phosphopeptide from H(+)-ATPase and a fungal phytotoxin (fusicoccin) bound in the deep groove of the protein. The toxin stabilizes the interaction between the ATPase and the 14-3-3 protein. This causes permanent activation of the proton pump in the guard cells of the leaf, so that the stomata open and the plant wilts. The phosphopeptide (sticks) and the toxin (space-filling, green carbons/red oxygens) are both accommodated in the deep groove of 14-3-3 and make hydrophobic interactions with each other. Compare with Figure 3, which shows just a phosphopeptide interaction with a human 14-3-3 protein. See (139) and pdb:1O9F for the crystal structure.

shown to be important for signal transduction and regulation. The same region in the human protein, although not similar on the sequence level, is known to be disordered. Hemsley *et al.* test the hypothesis that the disordered C-terminal region, despite no conservation at the sequence level, could have conserved functional activity, mediated by short linear motifs. Motif prediction algorithms and searches of the ELM database indicate the presence of a number of linear motifs, some of which are shared with mammalian versions of CRY (133). Deletional/mutational analysis of dCRY was able to delineate the position of the linear motif, by noting loss of function (i.e. ability to interact with partner proteins) when sections of the protein were removed. Site directed mutagenesis together with the yeast two-hybrid experiments provided further evidence for the motif being responsible for the interaction of dCRY and its partners. Immunoprecipitation experiments using CRY conjugated to Hemagglutinin were also used to provide further evidence for the importance of the motif and to identify residues that determine specificity. Work by Losi in humans was able to identify key residues for determining interactions between CRY and other partner proteins in the pathway (134). In both cases, the experimental validation was unable to unambiguously delineate the motifs responsible. In the case of the work by Partch *et al.* (135), this is due to limiting the mutational analysis to proteolysis. The bioinformatics approach described here is new and further refinement will no doubt lead to more successful results.

Antibodies and the linear epitopes that they recognise have been used to define targets for antibody-based therapeutics against HIV. In this case, the antibodies are raised using the peptide arrays to linear epitopes on surface accessible regions of key HIV proteins. The aim is the application of antibodies to disrupt the action of HIV, for an example see Huang *et al.* (136). The power of antibodies to recognise and specifically bind protein motifs has had an enormous impact on the study of protein-protein interactions. This is likely to continue as the demand for more and more information about the type and specificity of protein interactions increases. Particular motifs of interest, perhaps identified using tools such as those shown in Table 4, can be investigated using antibodies. An example of such work is Friedman (137). They present an investigation of MAPK substrates using phospho-specific antibodies. Antibodies have successfully been used to study other systems such as the 14-3-3 kinases in Arabidopsis (138).

To conclude this section, antibodies are very important tools for linear motif biology. With such simple targets, there is obviously a risk of cross-reaction and we end with a word of warning: design experiments such that a second independent method is used to ensure that the protein itself can be verified.

9. LINEAR MOTIF INTERACTIONS AND DRUG DISCOVERY

In the search for new drugs, pharmaceutical companies have been notoriously leery of targeting almost anything except cellular enzymes. In the context of cell signaling through linear motifs, classical targets are therefore kinases, phosphatases, methylases, acetylases and other enzymes of PTM. The efficacy of the tyrosine kinase inhibitor Gleevec (imatinib) in treating cancers, e.g. CML and GIST, had such an impact that some 30% of pharmaceutical research became devoted to kinases. While there will no doubt be some useful new treatments arising from this focus, the redundancy in kinase site specificities is one of the important reasons why many of them will never be precisely targeted.

We consider that the conservatism of pharmaceutical companies in avoiding protein-protein interaction targets is overblown and that at least some linear motif interactions will themselves prove to be excellent targets, namely those motifs that bind in grooves of ligand domains. It is often said that drugs are merely poisons administered at lower doses. The fungal phytotoxin fusicoccin shows the way: 14-3-3 proteins have a large and deep groove in which to accept the phosphopeptide ligand (Figure 9) and this is a potent target for fusicoccin (139). One of the first domain-peptide interactions to be selected for inhibition studies was SH2 (140, 141). There has been good progress, at least in the case of the Grb2 SH2, where a macrocyclic phosphopeptide mimetic has been obtained with nanomolar affinity and for which antiproliferative effects have been demonstrated (142).

The most striking examples of the potential for linear motif disruptors are the nutlins that block ubiquitin

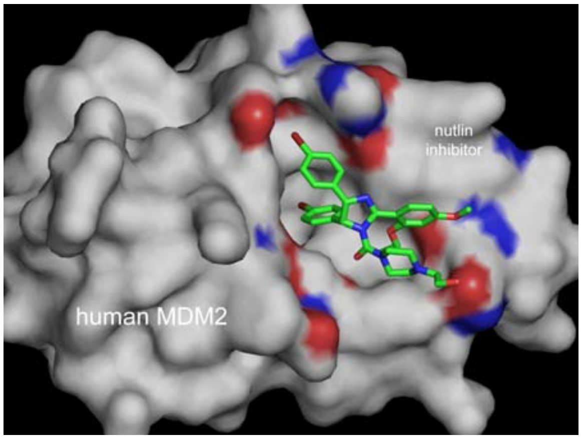


Figure 10. Candidate drug molecule binding a human protein target. The synthetic inhibitor nutlin (sticks) is shown bound to human MDM2. Two hydrophobic pockets in the protein, one of them very deep, accommodate a pair of bromophenyl rings from the inhibitor. Disruption of MDM2/p53 interaction can lead to apoptosis and tumour suppression. See (170) and pdb:1RV1 for the crystal structure.

mediated destruction of p53. Nutlins achieve their effect by binding in the P53 peptide-binding cavity of the MDM2 group of E3 ubiquitin ligases, (Figure 10), stabilizing cellular P53 protein. The potential for combination therapy is illustrated by the effect of DNA Topoisomerase inhibitor in combination with nutlin on retinoblastoma cells: i.e. dramatic restoration of apoptosis in apparently death-resistant cells (143).

10. CONCLUSION

The last few years have at last begun to see development of the computational tools that are needed to aid and complement the experimental investigation of motif-rich regulatory proteins. The prediction of natively disordered protein segments has progressed very rapidly and is good enough that it can already be considered a mature field (which should not be taken to mean that future improvements are in any way undesired). Useful computational resources focused on linear motifs, such as the ELM resource, can now be used as aids to regulatory motif biology, with an important educational role. The first wave of tools to predict novel motifs have been developed, but need to be improved to reach their full potential as aids to experimental research. A vast amount of data is being unleashed by the high throughput proteomics techniques and will be avidly mined by bioinformaticians. However, our present knowledge of cellular signaling processes is still poor and the present understanding of cell regulation represents only a tiny glimpse at the true quantity of protein interactions and regulatory processes in eukaryotic cells. It is essential to introduce context dependence into motif analysis. In this respect, Linding *et al.* have laid down a marker by developing NetworKIN, an integrative computational approach that combines sequence motifs and protein association networks, to predict which protein kinases target experimentally identified phosphorylation sites *in vivo* (144). Looking forward, we can expect a 3-pronged attack: One prong is represented by the improved computational methods for identifying regulatory motifs. These will be allied to high throughput methods such as phosphoproteomics and systematic cell complex identification for revealing regulatory components in bulk. But, it would be a fallacy to suppose that these massive data generating schemes will provide intimate understanding of cell regulation. The third prong will be, as ever, targeted experimental investigation into the nooks and crannies (the nodes and the feedback loops) of the signaling networks.

11. ACKNOWLEDGEMENTS

We thank all contributors to the ELM resource for making our work possible. This work was supported in part by the EU ProteomeBinders and EU EMBRACE (LHSG-CT-2004-512092) grants. GT was supported by EMBO and DAAD short-term fellowships.

12. REFERENCES

1. M.A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G.M. Rubin, J.A. Blake, C. Bult, M. Dolan, H. Drabkin, J.T. Eppig, D.P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J.M. Cherry, K.R. Christie, M.C. Costanzo, S.S. Dwight, S. Engel, D.G. Fisk, J.E. Hirschman, E.L. Hong, R.S. Nash, A. Sethuraman, C.L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S.Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E.M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried & R. White: The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32, D258-61, (2004)
2. T.J. Boggon & M.J. Eck: Structure and regulation of Src family kinases. *Oncogene* 23, 7918-7927, (2004)
3. T. Hunt: Protein sequence motifs involved in recognition and targeting: a new series. *TIBS* 15, 305, (1990)
4. H.L. Liu & J.P. Hsu: Recent developments in structural proteomics for protein structure determination. *Proteomics* 5, 2056-2068, (2005)
5. R.R. Copley, T. Doerks, I. Letunic & P. Bork: Protein domain analysis in the era of complete genomes. *FEBS Lett* 513, 129-134, (2002)
6. R.D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L. Sonnhammer & A. Bateman: Pfam: clans, web tools and services. *Nucleic Acids Res* 34, D247-51, (2006)
7. A. Andreeva, D. Howorth, S.E. Brenner, T.J. Hubbard, C. Chothia & A.G. Murzin: SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 32, D226-9, (2004)
8. I. Letunic, R.R. Copley, B. Pils, S. Pinkert, J. Schultz & P. Bork: SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* 34, D257-60, (2006)
9. A. Marchler-Bauer, J.B. Anderson, M.K. Derbyshire, C. DeWeese-Scott, N.R. Gonzales, M. Gwadz, L. Hao, S. He, D.I. Hurwitz, J.D. Jackson, Z. Ke, D. Krylov, C.J.

Lanczycki, C.A. Liebert, C. Liu, F. Lu, S. Lu, G.H. Marchler, M. Mullokandov, J.S. Song, N. Thanki, R.A. Yamashita, J.J. Yin, D. Zhang & S.H. Bryant: CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* 35, D237-40, (2007)

10. L.H. Greene, T.E. Lewis, S. Addou, A. Cuff, T. Dallman, M. Dibley, O. Redfern, F. Pearl, R. Nambudiry, A. Reid, I. Sillitoe, C. Yeats, J.M. Thornton & C.A. Orengo: The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res* 35, D291-7, (2007)

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

12. M. Sickmeier, J.A. Hamilton, T. LeGall, V. Vacic, M.S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V.N. Uversky, Z. Obradovic & A.K. Dunker: DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* 35, D786-93, (2007)

13. B.T. Seet & T. Pawson: MAPK signaling: Sho business. *Curr Biol* 14, R708-10, (2004)

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

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

16. E. Nogales, M. Whittaker, R.A. Milligan & K.H. Downing: High-resolution model of the microtubule. *Cell* 96, 79-88, (1999)

17. T. Nilsson & G. Warren: Retention and retrieval in the endoplasmic reticulum and the Golgi apparatus. *Curr Opin Cell Biol* 6, 517-521, (1994)

18. P. Pagel, M. Oesterheld, V. Stumpflen & D. Frishman: The DIMA web resource--exploring the protein domain network. *Bioinformatics* 22, 997-998, (2006)

19. P.D. Jeffrey, S. Gorina & N.P. Pavletich: Crystal structure of the tetramerization domain of the p53 tumor suppressor at 1.7. angstroms. *Science* 267, 1498-1502, (1995)

20. M.A. Poirier, W. Xiao, J.C. Macosko, C. Chan, Y.K. Shin & M.K. Bennett: The synaptic SNARE complex is a parallel four-stranded helical bundle. *Nat Struct Biol* 5, 765-769, (1998)

21. G. Wu, Y.G. Chen, B. Ozdamar, C.A. Gyuricza, P.A. Chong, J.L. Wrana, J. Massague & Y. Shi: Structural basis of Smad2 recognition by the Smad anchor for receptor activation. *Science* 287, 92-97, (2000)

22. G.J.J. Gatto, B.V. Geisbrecht, S.J. Gould & J.M. Berg: Peroxisomal targeting signal-1 recognition by the TPR domains of human PEX5. *Nat Struct Biol* 7, 1091-1095, (2000)

23. G.H. Kong, J.Y. Bu, T. Kurosaki, A.S. Shaw & A.C. Chan: Reconstitution of Syk function by the ZAP-70 protein tyrosine kinase. *Immunity* 2, 485-492, (1995)

24. S. Koyasu, A.G. Tse, P. Moingeon, R.E. Hussey, A. Mildonian, J. Hannisian, L.K. Clayton & E.L. Reinherz: Delineation of a T-cell activation motif required for binding of protein tyrosine kinases containing tandem SH2 domains. *Proc Natl Acad Sci U S A* 91, 6693-6697, (1994)

25. R. Ghose, A. Shekhtman, M.J. Goger, H. Ji & D. Cowburn: A novel, specific interaction involving the Csk SH3 domain and its natural ligand. *Nat Struct Biol* 8, 998-1004, (2001)

26. E. Nicolas, C. Roumillac & D. Trouche: Balance between acetylation and methylation of histone H3 lysine 9 on the E2F-responsive dihydrofolate reductase promoter. *Mol Cell Biol* 23, 1614-1622, (2003)

27. T. Jenuwein & C.D. Allis: Translating the histone code. *Science* 293, 1074-1080, (2001)

28. H. Li, S. Ilin, W. Wang, E.M. Duncan, J. Wysocka, C.D. Allis & D.J. Patel: Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* 442, 91-95, (2006)

29. B.D. Strahl, R. Ohba, R.G. Cook & C.D. Allis: Methylation of histone H3 at lysine 4 is highly conserved and correlates with transcriptionally active nuclei in Tetrahymena. *Proc Natl Acad Sci U S A* 96, 14967-14972, (1999)

30. T. Bouras, M. Fu, A.A. Sauve, F. Wang, A.A. Quong, N.D. Perkins, R.T. Hay, W. Gu & R.G. Pestell: SIRT1 deacetylation and repression of p300 involves lysine residues 1020/1024 within the cell cycle regulatory domain 1. *J Biol Chem* 280, 10264-10276, (2005)

31. G. Bossis & F. Melchior: SUMO: regulating the regulator. *Cell Div* 1, 13, (2006)

32. J.M. Desterro, M.S. Rodriguez & R.T. Hay: SUMO-1 modification of IkappaBalpha inhibits NF-kappaB activation. *Mol Cell* 2, 233-239, (1998)

33. J.C. Houtman, H. Yamaguchi, M. Barda-Saad, A. Braiman, B. Bowden, E. Appella, P. Schuck & L.E. Samelson: Oligomerization of signaling complexes by the multipoint binding of GRB2 to both LAT and SOS1. *Nat Struct Mol Biol* 13, 798-805, (2006)

34. N.R. Helps, X. Luo, H.M. Barker & P.T. Cohen: NIMA-related kinase 2 (Nek2), a cell-cycle-regulated protein kinase localized to centrosomes, is complexed to protein phosphatase 1. *Biochem J* 349, 509-518, (2000)

35. V.M. Coghlan, B.A. Perrino, M. Howard, L.K. Langeberg, J.B. Hicks, W.M. Gallatin & J.D. Scott: Association of protein kinase A and protein phosphatase 2B with a common anchoring protein. *Science* 267, 108-111, (1995)

36. S.J. Tavalin, M. Colledge, J.W. Hell, L.K. Langeberg, R.L. Huganir & J.D. Scott: Regulation of GluR1 by the A-kinase anchoring protein 79 (AKAP79) signaling complex shares properties with long-term depression. *J Neurosci* 22, 3044-3051, (2002)

37. M. Colledge, R.A. Dean, G.K. Scott, L.K. Langeberg, R.L. Huganir & J.D. Scott: Targeting of PKA to glutamate receptors through a MAGUK-AKAP complex. *Neuron* 27, 107-119, (2000)

38. K. Yoshioka: Scaffold proteins in mammalian MAP kinase cascades. *J Biochem (Tokyo)* 135, 657-661, (2004)

39. D.K. Pokholok, J. Zeitlinger, N.M. Hannett, D.B. Reynolds & R.A. Young: Activated signal transduction kinases frequently occupy target genes. *Science* 313, 533-536, (2006)

40. M. Ullah, H. Schmidt, K.H. Cho & O. Wolkenhauer: Deterministic modelling and stochastic simulation of

biochemical pathways using MATLAB. *Syst Biol (Stevenage)* 153, 53-60, (2006)

41. S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes & U. Kummer: COPASI--a COmplex PAthway SImulator. *Bioinformatics* 22, 3067-3074, (2006)

42. D. Gilbert, H. Fuss, X. Gu, R. Orton, S. Robinson, V. Vyshemirsky, M.J. Kurth, C.S. Downes & W. Dubitzky: Computational methodologies for modelling, analysis and simulation of signalling networks. *Brief Bioinform* 7, 339-353, (2006)

43. R. Bose, M.A. Holbert, K.A. Pickin & P.A. Cole: Protein tyrosine kinase-substrate interactions. *Curr Opin Struct Biol* 16, 668-675, (2006)

44. I.A. Yudushkin, A. Schleifenbaum, A. Kinkhabwala, B.G. Neel, C. Schultz & P.I. Bastiaens: Live-cell imaging of enzyme-substrate interaction reveals spatial regulation of PTP1B. *Science* 315, 115-119, (2007)

45. E. Fernandez, R. Schiappa, J.A. Girault & N. Le Novere: DARPP-32 is a robust integrator of dopamine and glutamate signals. *PLoS Comput Biol* 2, e176, (2006)

46. P. Puntervoll, R. Linding, C. Gemund, S. Chabanis-Davidson, M. Mattingsdal, S. Cameron, D.M. Martin, G. Ausiello, B. Brannetti, A. Costantini, F. Ferre, V. Maselli, A. Via, G. Cesareni, F. Diella, G. Superti-Furga, L. Wyrwicz, C. Ramu, C. McGuigan, R. Gudavalli, I. Letunic, P. Bork, L. Rychlewski, B. Kuster, M. Helmer-Citterich, W.N. Hunter, R. Aasland & T.J. Gibson: ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* 31, 3625-3630, (2003)

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

48. M.A. Young, S. Gonfloni, G. Superti-Furga, B. Roux & J. Kuriyan: Dynamic coupling between the SH2 and SH3 domains of c-Src and Hck underlies their inactivation by C-terminal tyrosine phosphorylation. *Cell* 105, 115-126, (2001)

49. M.L. Miller, S. Hanke, A.M. Hinsby, C. Friis, S. Brunak, M. Mann & N. Blom: Motif decomposition of the phosphotyrosine proteome reveals a new N-terminal binding motif for SHIP2. *Mol Cell Proteomics* 7, 181-192, (2008)

50. M.B. Yaffe: Phosphotyrosine-binding domains in signal transduction. *Nat Rev Mol Cell Biol* 3, 177-186, (2002)

51. M.B. Yaffe & A.E. Elia: Phosphoserine/threonine-binding domains. *Curr Opin Cell Biol* 13, 131-138, (2001)

52. T. Ludwig, P. Fisher, S. Ganesan & A. Efstratiadis: Tumorigenesis in mice carrying a truncating Brca1 mutation. *Genes Dev* 15, 1188-1193, (2001)

53. C.X. Deng & F. Scott: Role of the tumor suppressor gene Brca1 in genetic stability and mammary gland tumor formation. *Oncogene* 19, 1059-1064, (2000)

54. J.A. Clapperton, I.A. Manke, D.M. Lowery, T. Ho, L.F. Haire, M.B. Yaffe & S.J. Smerdon: Structure and mechanism of BRCA1 BRCT domain recognition of phosphorylated BACH1 with implications for cancer. *Nat Struct Mol Biol* 11, 512-518, (2004)

55. J.N. Glover, R.S. Williams & M.S. Lee: Interactions between BRCT repeats and phosphoproteins: tangled up in two. *Trends Biochem Sci* 29, 579-585, (2004)

56. K. Hofmann & P. Bucher: The FHA domain: a putative nuclear signalling domain found in protein kinases and transcription factors. *Trends Biochem Sci* 20, 347-349, (1995)

57. D. Durocher & S.P. Jackson: The FHA domain. *FEBS Lett* 513, 58-66, (2002)

58. B.L. Pike, S. Yongkiettrakul, M.D. Tsai & J. Heierhorst: Diverse but overlapping functions of the two forkhead-associated (FHA) domains in Rad53 checkpoint kinase activation. *J Biol Chem* 278, 30421-30424, (2003)

59. I.J. Byeon, H. Li, H. Song, A.M. Gronenborn & M.D. Tsai: Sequential phosphorylation and multisite interactions characterize specific target recognition by the FHA domain of Ki67. *Nat Struct Mol Biol* 12, 987-993, (2005)

60. C.M. Pfleger & M.W. Kirschner: The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1. *Genes Dev* 14, 655-665, (2000)

61. M. Glotzer, A.W. Murray & M.W. Kirschner: Cyclin is degraded by the ubiquitin pathway. *Nature* 349, 132-138, (1991)

62. A. Castro, C. Bernis, S. Vigneron, J.C. Labbe & T. Lorca: The anaphase-promoting complex: a key factor in the regulation of cell cycle. *Oncogene* 24, 314-325, (2005)

63. J.M. Peters: The anaphase promoting complex/cyclosome: a machine designed to destroy. *Nat Rev Mol Cell Biol* 7, 644-656, (2006)

64. S. Michael, G. Trave, C. Ramu, C. Chica & T.J. Gibson: Discovery of candidate KEN-box motifs using cell cycle keyword enrichment combined with native disorder prediction and motif conservation. *Bioinformatics* 24, 453-457, (2008)

65. U. Fischer, J. Huber, W.C. Boelens, I.W. Mattaj & R. Luhrmann: The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs. *Cell* 82, 475-483, (1995)

66. W. Wen, J.L. Meinkoth, R.Y. Tsien & S.S. Taylor: Identification of a signal for rapid export of proteins from the nucleus. *Cell* 82, 463-473, (1995)

67. N. Kudo, B. Wolff, T. Sekimoto, E.P. Schreiner, Y. Yoneda, M. Yanagida, S. Horinouchi & M. Yoshida: Leptomycin B inhibition of signal-mediated nuclear export by direct binding to CRM1. *Exp Cell Res* 242, 540-547, (1998)

68. M.J. Zhang & A.I. Dayton: Tolerance of diverse amino acid substitutions at conserved positions in the nuclear export signal (NES) of HIV-1 Rev. *Biochem Biophys Res Commun* 243, 113-116, (1998)

69. E. Paraskeva, E. Izaurralde, F.R. Bischoff, J. Huber, U. Kutay, E. Hartmann, R. Luhrmann & D. Gorlich: CRM1-mediated recycling of snurportin 1 to the cytoplasm. *J Cell Biol* 145, 255-264, (1999)

70. D. Engelsma, R. Bernad, J. Calafat & M. Fornerod: Supraphysiological nuclear export signals bind CRM1 independently of RanGTP and arrest at Nup358. *EMBO J* 23, 3643-3652, (2004)

71. O. Hantschel, B. Nagar, S. Guettler, J. Kretzschmar, K. Dorey, J. Kuriyan & G. Superti-Furga: A myristoyl/phosphotyrosine switch regulates c-Abl. *Cell* 112, 845-857, (2003)

72. J. Kadlec, E. Izaurralde & S. Cusack: The structural basis for the interaction between nonsense-mediated mRNA decay factors UPF2 and UPF3. *Nat Struct Mol Biol*

11, 330-337, (2004)

73. W. Meng, L.L. Swenson, M.J. Fitzgibbon, K. Hayakawa, E. Ter Haar, A.E. Behrens, J.R. Fulghum & J.A. Lippke: Structure of mitogen-activated protein kinase-activated protein (MAPKAP) kinase 2 suggests a bifunctional switch that couples kinase activation with nuclear export. *J Biol Chem* 277, 37401-37405, (2002)

74. J.M. Stommel, N.D. Marchenko, G.S. Jimenez, U.M. Moll, T.J. Hope & G.M. Wahl: A leucine-rich nuclear export signal in the p53 tetramerization domain: regulation of subcellular localization and p53 activity by NES masking. *EMBO J* 18, 1660-1672, (1999)

75. T. Sternsdorf, K. Jensen, B. Reich & H. Will: The nuclear dot protein sp100, characterization of domains necessary for dimerization, subcellular localization, and modification by small ubiquitin-like modifiers. *J Biol Chem* 274, 12555-12566, (1999)

76. C. Endter, J. Kzhyshkowska, R. Stauber & T. Dobner: SUMO-1 modification required for transformation by adenovirus type 5 early region 1B 55-kDa oncoprotein. *Proc Natl Acad Sci U S A* 98, 11312-11317, (2001)

77. H. Poukka, U. Karvonen, O.A. Janne & J.J. Palvimo: Covalent modification of the androgen receptor by small ubiquitin-like modifier 1 (SUMO-1). *Proc Natl Acad Sci U S A* 97, 14145-14150, (2000)

78. R.T. Hay: Role of ubiquitin-like proteins in transcriptional regulation. *Ernst Schering Res Found Workshop* 173-192, (2006)

79. S. Muller, M.J. Matunis & A. Dejean: Conjugation with the ubiquitin-related modifier SUMO-1 regulates the partitioning of PML within the nucleus. *EMBO J* 17, 61-70, (1998)

80. O. Kerscher: SUMO junction-what's your function? New insights through SUMO-interacting motifs. *EMBO Rep* 8, 550-555, (2007)

81. S. Cai, C.C. Lee & T. Kohwi-Shigematsu: SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nat Genet* 38, 1278-1288, (2006)

82. H.L. Klein: A SUMOry of DNA replication: synthesis, damage, and repair. *Cell* 127, 455-457, (2006)

83. R.T. Hay: SUMO: a history of modification. *Mol Cell* 18, 1-12, (2005)

84. H. Saitoh & J. Hinchey: Functional heterogeneity of small ubiquitin-related protein modifiers SUMO-1 versus SUMO-2/3. *J Biol Chem* 275, 6252-6258, (2000)

85. Y. Azuma, A. Arnaoutov & M. Dasso: SUMO-2/3 regulates topoisomerase II in mitosis. *J Cell Biol* 163, 477-487, (2003)

86. M.H. Tatham, E. Jaffray, O.A. Vaughan, J.M. Desterro, C.H. Botting, J.H. Naismith & R.T. Hay: Polymeric chains of SUMO-2 and SUMO-3 are conjugated to protein substrates by SAE1/SAE2 and Ubc9. *J Biol Chem* 276, 35368-35374, (2001)

87. C.M. Hecker, M. Rabiller, K. Haglund, P. Bayer & I. Dikic: Specification of SUMO1- and SUMO2-interacting motifs. *J Biol Chem* 281, 16117-16127, (2006)

88. P. Heun: SUMOrganization of the nucleus. *Curr Opin Cell Biol* 19, 350-355, (2007)

89. W.W. Wasserman & A. Sandelin: Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* 5, 276-287, (2004)

90. G.D. Stormo: DNA binding sites: representation and discovery. *Bioinformatics* 16, 16-23, (2000)

91. B. Lenhard, A. Sandelin, L. Mendoza, P. Engstrom, N. Jareborg & W.W. Wasserman: Identification of conserved regulatory elements by comparative genome analysis. *J Biol* 2, 13, (2003)

92. R. Gutman, C. Berezin, R. Wollman, Y. Rosenberg & N. Ben-Tal: QuasiMotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns. *Nucleic Acids Res* 33, W255-61, (2005)

93. V. Neduva & R.B. Russell: DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Res* 34, W350-5, (2006)

94. N.E. Davey, R.J. Edwards & D.C. Shields: The SLiMDisc server: short, linear motif discovery in proteins. *Nucleic Acids Res* 35, W455-9, (2007)

95. V. Neduva & R.B. Russell: Linear motifs: evolutionary interaction switches. *FEBS Lett* 579, 3342-3345, (2005)

96. A. Bairoch: PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res* 20 Suppl, 2013-2018, (1992)

97. C.G. Nevill-Manning, T.D. Wu & D.L. Brutlag: Highly specific protein sequence motifs for genome analysis. *Proc Natl Acad Sci U S A* 95, 5865-5871, (1998)

98. J.Y. Huang & D.L. Brutlag: The EMOTIF database. *Nucleic Acids Res* 29, 202-204, (2001)

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

100. S. Balla, V. Thapar, S. Verma, T. Luong, T. Faghri, C.H. Huang, S. Rajasekaran, J.J. del Campo, J.H. Shinn, W.A. Mohler, M.W. Maciejewski, M.R. Gryk, B. Piccirillo, S.R. Schiller & M.R. Schiller: Minimotif Miner: a tool for investigating protein function. *Nat Methods* 3, 175-177, (2006)

101. D. Plewczynski, A. Tkacz, L.S. Wyrwicz & L. Rychlewski: AutoMotif server: prediction of single residue post-translational modifications in proteins. *Bioinformatics* 21, 2525-2527, (2005)

102. R.R. Copley: The EH1 motif in metazoan transcription factors. *BMC Genomics* 6, 169, (2005)

103. C. Ramu: SIRW: A web server for the Simple Indexing and Retrieval System that combines sequence motif searches with keyword searches. *Nucleic Acids Res* 31, 3771-3774, (2003)

104. R.J. Edwards, N.E. Davey & D.C. Shields: SLiMFinder: A Probabilistic Method for Identifying Over-Represented, Convergently Evolved, Short Linear Motifs in Proteins. *PLoS ONE* 2, e967, (2007)

105. S.H. Tan, W. Hugo, W.K. Sung & S.K. Ng: A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC Bioinformatics* 7, 502, (2006)

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

107. R. Linding, R.B. Russell, V. Neduva & T.J. Gibson: GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* 31, 3701-3708, (2003)

108. V.N. Uversky, J.R. Gillespie & A.L. Fink: Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* 41, 415-427, (2000)

109. A.K. Dunker, C.J. Brown, J.D. Lawson, L.M. Iakoucheva & Z. Obradovic: Intrinsic disorder and protein function. *Biochemistry* 41, 6573-6582, (2002)

110. J. Prilusky, C.E. Felder, T. Zeev-Ben-Mordehai, E.H. Rydberg, O. Man, J.S. Beckmann, I. Silman & J.L. Sussman: FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* 21, 3435-3438, (2005)

111. Z. Dosztanyi, V. Csizmok, P. Tompa & I. Simon: IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21, 3433-3434, (2005)

112. Z. Dosztanyi, V. Csizmok, P. Tompa & I. Simon: The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* 347, 827-839, (2005)

113. R.M. Williams, Z. Obradovi, V. Mathura, W. Braun, E.C. Garner, J. Young, S. Takayama, C.J. Brown & A.K. Dunker: The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac Symp Biocomput* 89-100, (2001)

114. J.J. Ward, L.J. McGuffin, K. Bryson, B.F. Buxton & D.T. Jones: The DISOPRED server for the prediction of protein disorder. *Bioinformatics* 20, 2138-2139, (2004)

115. Z.R. Yang, R. Thomson, P. McNeil & R.M. Esnouf: RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 21, 3369-3376, (2005)

116. K. Peng, P. Radivojac, S. Vucetic, A.K. Dunker & Z. Obradovic: Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 7, 208, (2006)

117. O.V. Galzitskaya, S.O. Garbuzynskiy & M.Y. Lobanov: FoldUnfold: web server for the prediction of disordered regions in protein chain. *Bioinformatics* 22, 2948-2949, (2006)

118. F. Ferron, S. Longhi, B. Canard & D. Karlin: A practical overview of protein disorder prediction methods. *Proteins* 65, 1-14, (2006)

119. R.B. Russell & T.J. Gibson: A careful disorderliness in the proteome: Sites for interaction and targets for future therapies. *FEBS Lett* (2008)

120. P. Romero, Z. Obradovic & A.K. Dunker: Natively disordered proteins: functions and predictions. *Appl Bioinformatics* 3, 105-113, (2004)

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

122. Garret,R.H. and Grisham,G.M. (2006) Brooks/Cole Publishing Co.

123. H. Dumortier, J. Klein Gunnewiek, J.P. Roussel, Y. van Aarssen, J.P. Briand, W.J. van Venrooij & S. Muller: At least three linear regions but not the zinc-finger domain of U1C protein are exposed at the surface of the protein in solution and on the human spliceosomal U1 snRNP particle. *Nucleic Acids Res* 26, 5486-5491, (1998)

124. P. Tompa & M. Fuxreiter: Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci* 33, 2-8, (2008)

125. C.D. Chen & B. Kemper: Different structural requirements at specific proline residue positions in the conserved proline-rich region of cytochrome P450 2C2. *J Biol Chem* 271, 28607-28611, (1996)

126. M. Kikuchi, M. Kataoka, T. Kojima, T. Horibe, K. Fujieda, T. Kimura & T. Tanaka: Single chain antibodies that recognize the N-glycosylation site. *Arch Biochem Biophys* 422, 221-229, (2004)

127. M.J. Combs & Y. Tan: Production of motif specific and context independent antibodies using peptide libraries as antigens. *U.S. Patent No. 6,982,318* (2000)

128. H. Zhang, X. Zha, Y. Tan, P.V. Hornbeck, A.J. Mastrangelo, D.R. Alessi, R.D. Polakiewicz & M.J. Comb: Phosphoprotein analysis using antibodies broadly reactive against phosphorylated motifs. *J Biol Chem* 277, 39379-39387, (2002)

129. S.H. Diks & M.P. Peppelenbosch: Single cell proteomics for personalised medicine. *Trends Mol Med* 10, 574-577, (2004)

130. M. Rodriguez, S.S. Li, J.W. Harper & Z. Songyang: An oriented peptide array library (OPAL) strategy to study protein-protein interactions. *J Biol Chem* 279, 8802-8807, (2004)

131. F. Diella, C.M. Gould, C. Chica, A. Via & T.J. Gibson: Phospho.ELM: a database of phosphorylation sites--update 2008. *Nucleic Acids Res* 36, D240-4, (2008)

132. J.M. Irish, R. Hovland, P.O. Krutzik, O.D. Perez, O. Bruserud, B.T. Gjertsen & G.P. Nolan: Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell* 118, 217-228, (2004)

133. M.J. Hemsley, G.M. Mazzotta, M. Mason, S. Dissel, S. Toppo, M.A. Pagano, F. Sandrelli, F. Meggio, E. Rosato, R. Costa & S.C. Tosatto: Linear motifs in the C-terminus of D. melanogaster cryptochrome. *Biochem Biophys Res Commun* 355, 531-537, (2007)

134. A. Losi: Flavin-based Blue-Light photosensors: a photobiophysics update. *Photochem Photobiol* 83, 1283-1300, (2007)

135. C.L. Partch, M.W. Clarkson, S. Ozgur, A.L. Lee & A. Sancar: Role of structural plasticity in signal transduction by the cryptochrome blue-light photoreceptor. *Biochemistry* 44, 3795-3805, (2005)

136. J.H. Huang, Z.Q. Liu, S. Liu, S. Jiang & Y.H. Chen: Identification of the HIV-1 gp41 core-binding motif--HXXNPF. *FEBS Lett* 580, 4807-4814, (2006)

137. A. Friedman & N. Perrimon: High-throughput approaches to dissecting MAPK signaling pathways. *Methods* 40, 262-271, (2006)

138. B. Fuller, S.M.J. Stevens, P.C. Sehnke & R.J. Ferl: Proteomic analysis of the 14-3-3 family in Arabidopsis. *Proteomics* 6, 3050-3059, (2006)

139. M. Wurtele, C. Jelich-Ottmann, A. Wittinghofer & C. Oecking: Structural view of a fungal toxin acting on a 14-3-3 regulatory complex. *EMBO J* 22, 987-994, (2003)

140. T.R.J. Burke, Z.J. Yao, D.G. Liu, J. Voigt & Y. Gao: Phosphoryltyrosyl mimetics in the design of peptide-based signal transduction inhibitors. *Biopolymers* 60, 32-44, (2001)

141. T.K. Sawyer, R.S. Bohacek, D.C. Dalgarno, C.J. Eyermann, N. Kawahata, C.A.r. Metcalf, W.C. Shakespeare, R. Sundaramoorthi, Y. Wang & M.G. Yang: SRC homology-2 inhibitors: peptidomimetic and nonpeptide. *Mini Rev Med Chem* 2, 475-488, (2002)

142. J. Phan, Z.D. Shi, T.R.J. Burke & D.S. Waugh: Crystal structures of a high-affinity macrocyclic peptide

mimetic in complex with the Grb2 SH2 domain. *J Mol Biol* 353, 104-115, (2005)
143. N.A. Laurie, S.L. Donovan, C.S. Shih, J. Zhang, N. Mills, C. Fuller, A. Teunisse, S. Lam, Y. Ramos, A. Mohan, D. Johnson, M. Wilson, C. Rodriguez-Galindo, M. Quarto, S. Francoz, S.M. Mendrysa, R.K. Guy, J.C. Marine, A.G. Jochemsen & M.A. Dyer: Inactivation of the p53 pathway in retinoblastoma. *Nature* 444, 61-66, (2006)
144. R. Linding, L.J. Jensen, G.J. Ostheimer, M.A. van Vugt, C. Jorgensen, I.M. Miron, F. Diella, K. Colwill, L. Taylor, K. Elder, P. Metalnikov, V. Nguyen, A. Pasculescu, J. Jin, J.G. Park, L.D. Samson, J.R. Woodgett, R.B. Russell, P. Bork, M.B. Yaffe & T. Pawson: Systematic discovery of *in vivo* phosphorylation networks. *Cell* 129, 1415-1426, (2007)
145. C.H. Benes, N. Wu, A.E. Elia, T. Dharia, L.C. Cantley & S.P. Soltoff: The C2 domain of PKCdelta is a phosphotyrosine binding domain. *Cell* 121, 271-280, (2005)
146. R.S. Williams, M.S. Lee, D.D. Hau & J.N. Glover: Structural basis of phosphopeptide recognition by the BRCT domain of BRCA1. *Nat Struct Mol Biol* 11, 519-525, (2004)
147. M. Allen, A. Friedler, O. Schon & M. Bycroft: The structure of an FF domain from human HYPA/FBP11. *J Mol Biol* 323, 411-416, (2002)
148. B.M. Chacko, B.Y. Qin, A. Tiwari, G. Shi, S. Lam, L.J. Hayward, M. De Caestecker & K. Lin: Structural basis of heteromeric smad protein assembly in TGF-beta signaling. *Mol Cell* 15, 813-823, (2004)
149. E. Vojnic, B. Simon, B.D. Strahl, M. Sattler & P. Cramer: Structure and carboxyl-terminal domain (CTD) binding of the Set2 SRI domain that couples histone H3 Lys36 methylation to transcription. *J Biol Chem* 281, 13-15, (2006)
150. M.K. Dougherty & D.K. Morrison: Unlocking the code of 14-3-3. *J Cell Sci* 117, 1875-1884, (2004)
151. M.A. Verdecia, M.E. Bowman, K.P. Lu, T. Hunter & J.P. Noel: Structural basis for phosphoserine-proline recognition by group IV WW domains. *Nat Struct Biol* 7, 639-643, (2000)
152. S. Orlicky, X. Tang, A. Willems, M. Tyers & F. Sicheri: Structural basis for phosphodependent substrate selection and orientation by the SCFCdc4 ubiquitin ligase. *Cell* 112, 243-256, (2003)
153. A.E. Elia, L.C. Cantley & M.B. Yaffe: Proteomic screen finds pSer/pThr-binding domain localizing Plk1 to mitotic substrates. *Science* 299, 1228-1231, (2003)
154. D. Durocher, J. Henckel, A.R. Fersht & S.P. Jackson: The FHA domain is a modular phosphopeptide recognition motif. *Mol Cell* 4, 387-394, (1999)
155. B. Hao, N. Zheng, B.A. Schulman, G. Wu, J.J. Miller, M. Pagano & N.P. Pavletich: Structural basis of the Cks1-dependent recognition of p27 (Kip1) by the SCF (Skp2) ubiquitin ligase. *Mol Cell* 20, 9-19, (2005)
156. N. Blom, S. Gammeltoft & S. Brunak: Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* 294, 1351-1362, (1999)
157. K. Julenius: NetCGlyc 1.0.: prediction of mammalian C-mannosylation sites. *Glycobiology* 17, 868-876, (2007)
158. O. Emanuelsson, H. Nielsen & G. von Heijne: ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci* 8, 978-984, (1999)
159. T.Y. Lee, H.D. Huang, J.H. Hung, H.Y. Huang, Y.S. Yang & T.H. Wang: dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res* 34, D622-7, (2006)
160. F. Monigatti, E. Gasteiger, A. Bairoch & E. Jung: The Sulfinator: predicting tyrosine sulfation sites in protein sequences. *Bioinformatics* 18, 769-770, (2002)
161. G. Bologna, C. Yvon, S. Duvaud & A.L. Veuthey: N-Terminal myristoylation predictions by ensembles of neural networks. *Proteomics* 4, 1626-1632, (2004)
162. K. Rittinger, J. Budman, J. Xu, S. Volinia, L.C. Cantley, S.J. Smerdon, S.J. Gamblin & M.B. Yaffe: Structural analysis of 14-3-3 phosphopeptide complexes identifies a dual role for the nuclear export signal of 14-3-3 in ligand binding. *Mol Cell* 4, 153-166, (1999)
163. C. Yuan, S. Yongkiettrakul, I.J. Byeon, S. Zhou & M.D. Tsai: Solution structures of two FHA1-phosphothreonine peptide complexes provide insight into the structural basis of the ligand specificity of FHA1 from yeast Rad53. *J Mol Biol* 314, 563-575, (2001)
164. Chenna, R. Sugawara, H. Koike, T. & Lopez, R: Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* 31, 3497-500, (2003)
165. Katoh, K. Kuma, K. Toh, H & Miyata, T: MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33, 511-8 (2005)
166. A. Ayed, F.A. Mulder, G.S. Yi, Y. Lu, L.E. Kay & C.H. Arrowsmith: Latent and active p53 are identical in conformation. *Nat Struct Biol* 8, 756-760, (2001)
167. S. Bell, C. Klein, L. Muller, S. Hansen & J. Buchner: p53 contains large unstructured regions in its native state. *J Mol Biol* 322, 917-927, (2002)
168. R. Dawson, L. Muller, A. Dehner, C. Klein, H. Kessler & J. Buchner: The N-terminal domain of p53 is natively unfolded. *J Mol Biol* 332, 1131-1141, (2003)
169. H.J. Dyson & P.E. Wright: Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6, 197-208, (2005)
170. L.T. Vassilev, B.T. Vu, B. Graves, D. Carvajal, F. Podlaski, Z. Filipovic, N. Kong, U. Kammlott, C. Lukacs, C. Klein, N. Fotouhi & E.A. Liu: *In vivo* activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* 303, 844-848, (2004)

Abbreviations: PTM; post-translational modification; #: hydrophobic amino acid; NES: nuclear export signal; NPC: nuclear pore complex; SUMO: Small Ubiquitin-related Modifiers; PEP: proline-enriched tyrosine phosphatase; TCR: T-cell antigen receptor; MSA: multiple sequence alignment; TF: transcription factor; IUP: intrinsically unstructured polypeptide; IDR: Intrinsically Disordered Regions; TM: FoldIndex.

Key Words: Linear motif, Globular domain, Cell regulation, Cell signaling, Protein interaction, Phosphopeptide, Natively disordered protein, Prediction methods, Network topology, Protein complex, Review

Send correspondence to: Toby Gibson, Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany, Tel: 49 6221 3878398, Fax: 49 6221 3878517, E-mail: toby.gibson@embl.de

http://www.bioscience.org/current/vol13.htm