A Million Peptide Motifs for the Molecular Biologist

Peter Tompa,1,2,* Norman E. Davey,3 Toby J. Gibson,4 and M. Madan Babu5,*
1VIB Structural Biology Research Center (SBRC), Vrije Universiteit Brussel, Brussels 1050, Belgium
2Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest 1519, Hungary
3Department of Physiology, University of California, San Francisco, CA 94158, USA
4Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany
5MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK
*Correspondence: ptompa@vub.ac.be (P.T.), madanm@mrc-lmb.cam.ac.uk (M.M.B.)
http://dx.doi.org/10.1016/j.molcel.2014.05.032

A molecular description of functional modules in the cell is the focus of many high-throughput studies in the postgenomic era. A large portion of biomolecular interactions in virtually all cellular processes is mediated by compact interaction modules, referred to as peptide motifs. Such motifs are typically less than ten residues in length, occur within intrinsically disordered regions, and are recognized and/or posttranslationally modified by structured domains of the interacting partner. In this review, we suggest that there might be over a million instances of peptide motifs in the human proteome. While this staggering number suggests that peptide motifs are numerous and the most understudied functional module in the cell, it also holds great opportunities for new discoveries.

Introduction

A molecular description of the full complement of functional modules of proteins in the cell, ranging from protein complexes to posttranslational modification sites (Figure 1), is the focus of many current high-throughput studies. Due to the relentless pace of the generation of novel information, a census of distinct molecular entities at various levels of description is needed to understand and appreciate how organisms use these basic tools to increase their functional versatility (Hartwell et al., 1999). In an early and seminal paper addressing this problem, Chothia (1992) suggested that there are about a thousand distinct protein folds in nature. For instance, the ~22,000 or so protein coding genes in the human genome are estimated to contain 35,000 instances of folded domains belonging to about 10,000 families (Sammut et al., 2008). Their functions are often realized through interactions with other proteins, mediated by an estimated 10,000 distinct interaction types (Aloy and Russell, 2004), resulting in several hundred thousands of binary interactions (Hart et al., 2006; Stumpf et al., 2008; Venkatesan et al., 2009) and 600–1,000 distinct protein complexes (Havugimana et al., 2012; Levy et al., 2009; Perica et al., 2012). While the exact number of instances of these functional modules will likely be influenced by the nuances of the definition used and by biological processes such as alternative splicing, these estimates nevertheless represent a cornerstone of our concepts of modularity of protein structure, function, and evolution.

In recent years it has become increasingly clear that the functional and molecular toolkit of eukaryotes extends beyond structured domains (Babu et al., 2012; Dyson and Wright, 2005; Tompa, 2012; Uversky and Dunker, 2010). A large portion of functional biomolecular interactions is mediated by compact interaction interfaces within intrinsically disordered regions (IDRs; i.e., segments of the polypeptide chain that do not autonomously fold into a defined tertiary structure) that are recognized and/or posttranslationally modified by structured domains of the interacting partner (Bhattacharyya et al., 2006; Davey et al., 2012a; Diella et al., 2008; Dinkel et al., 2014; Hornbeck et al., 2012). These short interaction interfaces (which we herein refer to as “peptide motifs”) to include both binding motifs and posttranslational modification (PTM) sites (see next section) are typically less than ten residues in length and enable both high functional diversity and functional density to polypeptide segments containing them. On evolutionary timescales, they can evolve very rapidly by appearing de novo or conversely disappearing with equal rapidity, conferring exceptional evolutionary plasticity on the interactome (Davey et al., 2012b; Mosca et al., 2012).

In this review, we discuss that an inventory of peptide motifs in proteins lags far behind that of domains and protein complexes. We then describe the current understanding of peptide motifs and provide estimates that there might be more than a hundred thousand—and possibly up to a million—instances in the human proteome. This overwhelming number implies that the quest for their identification and functional characterization is a formidable challenge. However, acquiring a deeper understanding of peptide motifs is imperative for gaining a more complete molecular description of the physiological and pathological processes of the cell.

Attributes of Peptide Motifs

The known instances of peptide motifs have been shown to mediate a wide range of important cellular tasks (Table 1). They typically fall into two major classes: binding motifs (Figures 2A–2C), which mediate interactions with globular domains, and posttranslational modification sites (Figure 2D), which are recognized and altered by modifying enzymes. These distinct terms (which might also be referred to as linear motifs [LMs] or short/eukaryotic linear motifs [ELMs/SLIMs]) denote functional sites that often occur within disordered regions of proteins. The two terms are frequently used synonymously, although they describe somewhat distinct, but overlapping, sets of functional modules.
For binding motifs, amino acid sequence specificity determinants can be described. These determinants are often discovered by identifying recurring patterns (or motifs) of residues from functionally equivalent peptide sequences. Posttranslational modification sites are less homogeneous, and many appear to lack specificity determinants outside of the modified residue (e.g., protein phosphatase sites), although certain enzymes have recognizable sequence specificities (e.g., proline-directed kinases) for the region they modify.

In the ELM database (http://elm.eu.org), the most comprehensive repository of experimentally characterized peptide motifs, there are about 2,475 documented instances of functional binding motifs and posttranslational modification sites in 1,547 proteins. These motifs are recognized and/or modified by 86 globular domain families, and the majority of them were discovered by low-throughput binding assays and structure-function studies of the domain-peptide motif complex (Dinkel et al., 2014). The number of posttranslational modification sites discovered by high-throughput studies (HTSs), however, is significantly higher, in a range of well above 100,000 sites (Beltrao et al., 2012; Choudhary and Mann, 2010). The number of structural domains known to participate in peptide-mediated interactions has also rapidly increased in recent years (~200 domain families; Stein et al., 2011). The documented instances of binding motifs, structural domains, and PTM sites form much of the basis of our understanding of peptide motifs. Nonetheless, as we discuss below, the number of actual instances of peptide motifs and peptide-domain interactions in the entire human proteome is likely to be significantly higher than what has been documented in current databases.

Based on the known instances of peptide motifs, the general features and evolutionary principles of binding motifs and PTM sites are becoming increasingly clear (Davey et al., 2012b). Binding motifs are on average 6–7 amino acids in length, with only 3–4 core positions conferring the majority of the interaction specificity (see consensus in Figure 2C). The limited binding surface area results in low-affinity, transient, and modulatable interactions, usually in the low micromolar range. PTM sites in general have weaker intrinsic specificity determinants (Perkins et al., 2010; Van Roey et al., 2012). Often, experimental approaches allow only the detection of the modified residue. Subsequent computational analysis is required to identify residues complementary with the enzyme’s active site and infer the motif directing the modification (Mahrus et al., 2008; Zielinska et al., 2010).

As a result of the involvement of a small number of residues in peptide motifs, these functional modules exhibit high evolutionary plasticity, and they can arise de novo and evolve convergently by one or a few mutations within disordered regions (Davey et al., 2012b; Holt et al., 2009; Mosca et al., 2012). In accordance, the peptide motifs are highly evolutionarily variable and, compared to globular domains within the same protein sequence, less conserved across large evolutionary distances. Most domain-binding motifs in human proteins are not typically conserved outside vertebrates, and PTM sites in particular are rapidly gained and lost during evolution (Beltrao et al., 2012; Hoit et al., 2009; Kim et al., 2012), though a subset of motifs involved in critical regulatory processes (e.g., in cell cycle; Galea et al., 2008) can be conserved over large evolutionary distances (Davey et al., 2012a, 2012b).

The amino acid composition of the polypeptide segment in which peptide motifs are embedded shows strong resemblance to that of IDRs (Fuxreiter et al., 2007). The majority of known binding motifs and tens of thousands of PTM sites have been discovered in predicted disordered regions (Beltrao et al., 2012; Davey et al., 2012b). This preference likely reflects (a) the accessibility of such motifs within disordered regions, which allows recognition by their binding partner, and (b) the flexibility of such motifs, which enables them to adopt a specific conformation upon binding (Davey et al., 2012b; Galea et al., 2008).

**Figure 1. Functional Modules of Proteins**

Modularity of protein function is manifest at distinct levels, from the recurrent use of whole-protein elements (complexes and structural domains [folds]) to protein parts (binding motifs and modification sites). The figure highlights their typical length scale and estimated number of instances in the human proteome, as outlined in detail in this manuscript.
evolutionary mechanism for generating proteins that can be subjected to extensive regulation by cellular pathways involved in protein homeostasis. Consequently, peptide motifs influence the life of almost all proteins from the cradle to the grave, regulating and coordinating their processing, localization, and degradation, as illustrated through the example of the cell cycle phosphatase CDC25A (Boutros et al., 2007). The biochemical activity of this enzyme is modulated at multiple levels by its long disordered region harboring numerous peptide motifs (see Figure 2A), which mediate regulatory interactions with cognate domains (Figure 2B). The newly synthesized protein is targeted within the cell via localization signals (e.g., nuclear localization signal [NLS] motif to target to the nucleus). The protein is posttranslationally modified both constitutively and conditionally on peptide motifs that act as modification sites (e.g., Chk1 kinase phosphorylation sites) or as docking sites (e.g., cyclin box).

---

Table 1. Motif Instances and Their Regulatory Roles

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Examples of Binding Partner</th>
<th>Motif</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binding Motifs (Mediate Interactions with Globular Domains)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classical binding motifs</td>
<td>act as simple binding modules, often functioning in the assembly of protein complexes as scaffolding modules</td>
<td>SH2 domain family</td>
<td>pYXXφ</td>
<td>phosphodependent SH2 domain binding motif</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UEV domain of TSG101</td>
<td>P[T][S]AP</td>
<td>Tsg101 UEV domain binding motif</td>
</tr>
<tr>
<td><strong>Functional Subsets</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trafficking motifs</td>
<td>mediate binding to protein necessary for localization to a particular subcellular location</td>
<td>exportin-1</td>
<td>φX(2,3)φX(2,3)φXφ</td>
<td>nuclear export signal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AP2</td>
<td>YXXφ</td>
<td>endocytic signal</td>
</tr>
<tr>
<td>Targeting motifs</td>
<td>retain protein at a subcellular location by binding to a previously localized protein</td>
<td>PCNA</td>
<td>QXXφXX[FHT][FHY]</td>
<td>replication fork localization motif</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>microtubule-associated protein EB family</td>
</tr>
<tr>
<td>Docking motifs</td>
<td>recruit enzyme to a protein and promote modification at a site distinct from the docking site</td>
<td>cyclin family</td>
<td>[RK]XLX(0,1)φ</td>
<td>CDK/cyclin kinase complex docking motif</td>
</tr>
<tr>
<td></td>
<td></td>
<td>calcineurin</td>
<td>PXIX[IV]</td>
<td>calcineurin phosphatase docking motif</td>
</tr>
<tr>
<td>Degron motifs</td>
<td>a functional subset of docking motifs that controls the stability of a protein by recruiting ubiquitin ligases, resulting in polyubiquitination and subsequent proteasomal degradation</td>
<td>Cdc20 and Cdh1 APC/C subunits</td>
<td>RXXLXXφ</td>
<td>APC/C-specific degradation motif</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fbw7 SCF ubiquitin ligase complex subunit</td>
<td>p[ST]PXxp[ST]</td>
<td>phosphodependent SCF degradation motif</td>
</tr>
<tr>
<td><strong>Posttranslational Modification Sites (Sites Are Recognized and Altered by Modifying Enzymes)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moiety addition sites</td>
<td>recognized by modifying enzymes for the attachment or removal of biochemical functional groups</td>
<td>SUMO-conjugating enzyme UBC9</td>
<td>φ(K)X(DE)</td>
<td>site SUMOylated by SUMO ligase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Polo-like kinase family</td>
<td>[DE]X([ST])φ</td>
<td>site phosphorylated by Polo-like kinases</td>
</tr>
<tr>
<td>Cleavage sites</td>
<td>recognized by proteases as sites for proteolytic cleavage</td>
<td>caspase-3</td>
<td>[DE]XD/[GSAN]</td>
<td>site of scission by caspase protease</td>
</tr>
<tr>
<td>Structural modification sites</td>
<td>recognized by enzymes that catalyze the structural conversion of a region</td>
<td>peptidyl-prolyl cis/trans-isomerase Pin1</td>
<td>p<a href="P">ST</a></td>
<td>site isomerised by PIN prol cis/trans-isomerase</td>
</tr>
</tbody>
</table>

φ represents bulky hydrophobic residues; X represents any amino acid; p signifies a phosphorylation; [] signifies a site of modification; | signifies a site of cleavage; {} signifies how many times the preceding residue can be used; and [] indicates the use one of the residues within the brackets to create a peptide motif.
The enzyme’s activity is controlled by peptide motifs, which mediate recruitment of binding partners to form functionally distinct complexes (e.g., 14-3-3 binding motif). In addition, its stability and half-life is controlled by degradation motifs, resulting in its ubiquitination and degradation by the proteasome (e.g., TrCP degron). The example of CDC25A highlighted here illustrates how peptide motifs may form an essential part of the homeostasis of many—if not all—proteins and are involved in nearly all cellular processes. This premise makes a large number of motif instances in the human proteome very likely.

High-Throughput Studies Suggest a Very High Number of Binding Motif Instances

Transient interactions mediated by short peptide motifs are often overlooked in classical proteomics methods of protein-protein interaction discovery that are based on tandem affinity purification tag (TAP-tag) or yeast two-hybrid (Y2H) experiments, as such methods are typically biased toward stable interactions (Diella et al., 2008; Landry et al., 2013; Liu et al., 2012; Neduva and Russell, 2006). Motif-domain interactions are inherently weaker, generally promiscuous, and more transient than domain-domain interactions (Schreiber et al., 2009; Zhou, 2012). The protein-protein interactions they mediate often depend on other factors, such as the PTM status of the motif and the abundance of the interacting protein, both of which are not generally accounted for in experiments identifying interactions (Landry et al., 2013). This may not be an issue for certain protein domains, such as SH3 (Tonikian et al., 2009) and PDZ (Belotti et al., 2013), as the interactions they mediate are rather stable and are much better represented in HTSs. However, it

![Diagram](image-url)

**Figure 2. Binding Motifs and Posttranslational Modification Sites in Proteins**

(A) Architecture of functional modules in the cell cycle phosphatase CDC25A. The protein is largely intrinsically disordered, with the exception of the phosphatase catalytic domain (residues 376–482, PDB ID: 1C25) in the C terminus. The protein contains multiple peptide motifs that regulate the protein from translation to degradation, such as intracellular localization (a nuclear localization signal [NLS], yellow), regulated degradation by polyubiquitination (degrons KEN box, red, and a phophodependent noncanonical DSG degron, green), binding to cyclin (cyclin box, blue), and direct regulation of enzymatic activity (phosphodependent 14-3-3 binding motifs, orange). The phosphorylation state of CDC25A is controlled by multiple sites with specificities for various kinases including Chk1 and Cdk1.

(B) Structures of instances of the peptide motif classes found in CDC25A, solved in other proteins that have convergently evolved the motifs, bound to their corresponding protein domains. Peptide motifs are color coded according to the representation of the modular domain architecture (i.e., the order of appearance of structured domains from N to C terminus in the protein) in (A). The structures from left to right are as follows: a CDK/cyclin docking motif from PS3 bound to cyclin A (PDB ID: 1H26), a SCF-TrCP degron from β-catenin bound to the SCF substrate recognition subunit (PDB ID: 1P22), a KEN box from BUBR1 bound to the APC substrate recognition subunit (PDB ID: 4GGD), a NLS from pRb bound to Importin alpha (PDB ID: 1PJM), and two 14-3-3 motifs from PKC bound to a 14-3-3 dimer (PDB ID: 2WH0).

(C) Convergently evolved instances of the regulatory motifs found in CDC25 that have been experimentally validated in other nonhomologous eukaryotic proteins as documented in the Eukaryotic Linear Motif resource (ELM). Sequences from CDC25 are color coded according to the representation of the domain architecture in (A). The major specificity-determining residues in the motifs are in bold and highlighted in light blue. Peptide motifs were validated in human unless stated: *, peptide motifs validated in Xenopus laevis; †, sequence of protein used in the structural panel.

(D) p300 is a transcription coactivator that regulates hundreds of transcription factors. The protein is 2,414 amino acids in length and contains nine domains (colored) and long connecting disordered regions (white) that constitute more than half of its length. UniProt (UniProt Consortium, 2013) and PhosphoSite (Hornbeck et al., 2012) list 123 PTMs in p300, such as phosphorylation (P, yellow), acetylation (A, blue), methylation, or double methylation (M, purple). Certain residues undergo two types of modification, such as acetylation and sumoylation (S, pink) or acetylation and ubiquitination (U, green). As it appears, PTMs can occur within disordered protein regions and also in folded domains, often in their exposed loop regions.
becomes an important consideration for identifying PTM site-mediated interactions using Y2H, as the relevant modifying enzyme that modifies and activates the motif may not be expressed under standard experimental conditions (Liu et al., 2012). The bias of human genome-scale Y2H interaction studies in determining motif-mediated interaction is also highlighted by the observation that only 1% of the detected interactions are mediated by peptide motifs, whereas in certain instances, such as the TGF-β pathway, more than half of the interactions are actually mediated by peptide motifs (Neduva and Russell, 2006).

Still, recent HTSs and computational studies (e.g., direct peptide library- and protein chip-based approaches of motif discovery) indicate very high numbers of individual motif instances (Liu et al., 2012; Neduva and Russell, 2005) in the human proteome. Several distinct structural domain families in the human proteome (e.g., SH2, SH3, PDZ, WW, and PTB domains) that function through peptide motif recognition are represented multiple times; for example, there are approximately 270 PDZ, 280 SH3, and 110 SH2 domains in the human genome (te Velthuis et al., 2011; Wang et al., 2010). Typically, each member recognizes multiple, often similar, peptide motifs in different proteins. For example, a single SH3 domain can bind peptide motifs in up to 80 different partners (Landgraf et al., 2004), and a recent analysis has shown that ~600 different human proteins have a C-terminal PDZ domain-binding motif (Kim et al., 2012). While the peptide motifs occur frequently in multiple proteins, it will be important to infer what proportion of these are actually functionally relevant (see below). Given (i) the above-mentioned observations on the typical number of motif-mediated interactions that protein domains participate in and (ii) that no interacting peptide partner has yet been identified for about 75% of the ~10,000 structural domain families (Stein et al., 2011), we suggest that many novel instances and classes of peptide-domain interactions remain to be discovered.

Unbiased Estimates on the Number of Binding Motifs in the Human Proteome

A simple census based on the foregoing fragmentary estimates might project the number of binding motif instances in the range of tens or even hundreds of thousands. Can we provide independent evidence for this formidable number of binding motifs? To overcome the statistical problem of sequence comparisons involving peptide motifs, we devised two straightforward ways of estimating the number of binding motif instances in the human proteome, based on independent sequence/structure attributes. Both estimates put the number of binding motifs above a hundred thousand.

Binding motifs can be predicted by virtue of their relationship with local structural disorder (Fuxreiter et al., 2007) and their tendency to undergo induced folding upon binding to their partner (Fuxreiter et al., 2004; Wright and Dyson, 2009). A predictor based on these premises, α-MoRF-Pred II, identifies short recognition features that adopt α-helical structure in the bound state (α-MoRFs); the predictor estimates about 3 such motifs in every 1,000 residues in eukaryotes (Cheng et al., 2007). Assuming ~22,000 proteins of 500 amino acids on average, this estimation suggests ~33,000 α-MoRFs in the human proteome. Given that only about 27% of the binding motifs are in an α-helical conformation when bound (Mohan et al., 2006) (the rest can be in β strand or another conformation), this allows us to estimate that the human proteome might contain as many as 122,000 MoRFs.

A different calculation can be made using the ANCHOR program that predicts likely binding motifs. ANCHOR works by analyzing the sequence of a disordered region and estimating the energy of interaction of such a region with a general interaction partner (Mészáros et al., 2009). Proteome-scale ANCHOR estimations suggested that higher eukaryotes have about 30% of their residues in disordered regions, 60% of which are likely to be involved in binding (Mészáros et al., 2009). By conservative estimates on the human proteome as above, and with an upper limit of 15 residues per motif (Tompa et al., 2009), a simple calculation (22,000 × 500 × 0.3 × 0.6 / 15) estimates a number of about 132,000 putative binding motifs in the human proteome.

What about the Number of Posttranslational Modification Sites?

These estimates all put the number of binding motifs above the hundred thousand mark. Here we discuss that the number of PTM sites is several times higher, and together there may easily be a million instances of peptide motifs in the human proteome.

Eukaryotic proteins are subject to more than 300 distinct types of enzyme-catalyzed posttranslational modifications (Jensen, 2004), most of which have yet to be investigated in detail. PTMs regulate many of the dynamic interactions of the interactome by creating a new binding interface, modulating the affinity of existing binding surfaces either positively or negatively, or altering the activity of a folded domain (Beltrao et al., 2012; Beltrao et al., 2013; Nishi et al., 2011; Van Roey et al., 2012). Consequently, a large portion of higher eukaryotic proteomes encodes modifying enzymes. For example, hundreds of kinases, phosphatases, and ubiquitin ligases are encoded in the human genome (Bhore et al., 2013; Manning et al., 2002). Any of these enzymes can modify multiple sites on a large number of substrate proteins. For instance, 6,000 classical NX(T/S) glycosylation motifs are modified by a single oligosaccharyltransferase complex (Zielinska et al., 2010), as many as 1,001 sites may be phosphorylated by the acidophilic casein kinase II (Meggio and Pinna, 2003), and 333 aspartate-directed cleavage sites, described in 292 substrates, have been attributed to caspase-like proteases alone (Mahrus et al., 2008). If a single enzyme can modify such a large number of sites, the presence of such large numbers of modifying enzymes in the human genome therefore suggests the presence of very high numbers of PTM sites. In fact, mass spectrometry (MS)-based HTS studies suggest that we have only seen the tip of the iceberg as characterized by low-throughput studies: a recent analysis of human modification sites cataloged 31,165 phosphorylation, 22,057 ubiquitination, 8,042 acetylation, and 6,422 proteolytic cleavage sites (Beltrao et al., 2012), and in another study it was suggested that there may be more than 100,000 sites in the phosphoproteome of human cells (Choudhary and Mann, 2010).

Unbiased Estimates of the Number of PTMs in the Human Proteome

An important limitation even of HTS studies is that these experiments are generally carried out in a limited set of conditions; i.e.,
they never sample all possible cell states/conditions, and they invariably introduce a systematic bias against chemically labile modifications. Furthermore, we know very little about the stoichiometry of the modifications, so many may be very rare and rapidly demodified (Beltrao et al., 2013). To assess the real numbers, one might refer to a few actual proteins that have been extensively studied by a combination of high-throughput and low-throughput methods due to their biological and/or biomedical importance. For example, the central transcription coactivator p300 is 2,414 residues in length and contains 123 identified PTMs (Figure 2D), such as phosphorylation, acetylation, methylation, ubiquitination, and sumoylation, as compiled in the PhosphoSite (http://www.phosphosite.org) (Hornbeck et al., 2012) and UniProt (www.uniprot.org) (UniProt Consortium, 2013) databases. Similar studies of p53 (62 PTMs/393 residues in length), tau protein (123/758), BRCA1 (88/1863), α-synuclein (22/140), c-Myc (45/439), Hdm2 (68/491), retinoblastoma protein (72/928), and Rad18 (43/495) suggest that the PTM density of the proteome (523/5,507) may be as high as one modification in every ten residues. This suggests close to a million PTM instances in the human proteome. Given that there are more than 300 different PTM types (Jensen, 2004), of which we usually focus on only a handful as mentioned above, this number does not seem to be exaggerated.

Interplay between Binding Motifs and PTM Sites Increases Regulatory Potential

By all the above-mentioned estimates and considerations, the full complement of peptide motifs (the “motifome”) may well reach into the millions in eukaryotes such as human. Many regions will contain several distinct motifs, including overlapping autonomous recognition sites, different types of modification of the same residue by different enzyme classes, and the same modification on a single site by distinct enzymes under different cellular conditions. Furthermore, due to the promiscuity of peptide motifs, which allows the recognition of the same binding motif by multiple members of the same structural domain family, the number of motif-mediated interactions by a protein region can far exceed the number of instances of peptide motifs.

The issue of the real number of instances of peptide motifs is further complicated by the complex interplay between the binding motifs and PTM sites, in which the two categories cooperate and complement each other. A PTM within the binding motif, or more often, in the immediate flanking regions, is a common mechanism to regulate interactions mediated by a binding motif—referred to as PTM switches (Van Roey et al., 2012, 2013). For example, the canonical nonconstitutive ligand motifs (e.g., 14-3-3 binding sites) require modification of specific residues to be recognized by their recognition domains. Adjacent or overlapping binding motifs can function competitively and allow the manifestation of even more complex switches. Conversely, the cooperative multivalent binding of low-affinity binding motifs can recruit binding partners in a high-avidity interaction. Such conditional multisite interaction interfaces, which are dependent on the state of the cell, result in complex regulatory mechanisms, which are exploited in information processing and cellular signaling (Van Roey et al., 2012). Higher-order signaling phenomena might also arise from phase transitions mediated by high motif valency (Li et al., 2012; Tompa, 2013). Complex patterns of modifications at multiple PTM sites in the interface region can integrate multiple signals to regulate an interaction in either a binary or rheostatic manner (Wu, 2013). Not unexpectedly, peptide motifs are enriched in nonconstitutive, alternatively spliced exons, tuning the regulatory potential of a protein by adding or removing peptide motifs (Buljan et al., 2012; Romero et al., 2006; Weatheritt et al., 2012) and also facilitating the rewiring of the interactome in different tissues (Buljan et al., 2012, 2013; Ellis et al., 2012).

Discussions of modular protein functions are somewhat hampered due to the definition of function, as applied to genes and gene products. For instance, function may be defined differently by different people in the sense that no objective definition is offered (Greenspan, 2011). In this sense, just like the question of what is a gene and thus how many genes are there in the genome (Pearson, 2006), the question of what is a peptide motif and how many are there in the proteome is rendered somewhat philosophical by the above-described complex combinatorial interplay between nearby binding motifs and between adjacent binding motifs and PTM sites, and also due to the issue of defining function at different molecular and cellular levels.

Are All Binding Motifs and PTM Sites Functionally Relevant?

Given the enormous plasticity in the evolution of binding motifs and PTM sites, they may arise convergently and be lost rapidly in a short timescale (Davey et al., 2011, 2012b; Holt et al., 2009). Not surprisingly, many PTM sites and binding motifs are not evolutionarily well conserved (Beltrao et al., 2012; Freschi et al., 2011; Holt et al., 2009). This rapid evolution may give rise to “noisy” interactions that may not have functional consequence (Landry et al., 2013; Levy et al., 2009), thus raising an important question: how many of these binding motifs, PTM sites, and interactions mediated by these modules are functionally relevant? While all peptide motifs recognize their binding partner, and it is unlikely that all these binding events are functionally important and hence evolutionarily conserved across diverse lineages, it has been noted that their rapid evolution may lead to new interactions, upon which selection can operate. This may eventually lead to functional rewiring of molecular interactions and the emergence of species-specific phenotypes during the course of organismal evolution (Buljan et al., 2013; Kim et al., 2012; Mosca et al., 2012). Thus, absence of conservation of some peptide motifs across orthologous proteins does not mean they are nonfunctional. Depending on the complexity of the proteome of an organism, specificity in peptide-mediated interactions can be achieved by continually eliminating promiscuous, nonfunctional binding motifs by negative selection (Zarringar et al., 2003).

In accord, recent studies on individual proteins suggest that nonconserved PTM sites may be functional (Holt et al., 2008). A good example is Securin, an inhibitor of Separase, the protease responsible for the separation of sister chromatids. In yeast, two Cdk1 phosphorylation sites inhibit the APC/C-mediated ubiquitination of Securin. These sites protect Securin from APC/C-mediated degradation until the appropriate point of the
cell cycle when the Cdc14 phosphatase can remove the phosphate group (Holt et al., 2008). These Cdk1 phosphorylation sites are important regulatory modifications in S. cerevisiae. However, the phosphodependent stabilization mechanism is absent outside of species closely related to yeast.

Taken together, in addition to the evolutionary conservation of peptide motifs, the relevance of their occurrence in proteomes needs to be characterized in the appropriate model organism to assess their true functionality. A careful choice of experimental design to test motif functionality is crucial, as their functionality is sensitive to expression level and may lead to incorrect inferences about their functional relevance (Gibson et al., 2013). Therefore, an important challenge for the community is to not only identify binding motifs and PTM sites, but also functionally characterize these peptide motifs by investigating them in the right biological context.

Conclusions and Future Directions

All the points raised here highlight the exceptional evolutionary agility and functional plasticity of peptide motifs. They are evolutionarily, functionally, dynamically, and contextually different from other types of functional protein modules, filling a distinct and self-evidently essential gap in the continuum of functional protein module types (Figure 1).

When we consider the functional repertoire of peptide motifs, the possibility of de novo evolution, and their propensity to be enriched in alternatively spliced regions and IDRs, the evidence strongly suggests that they play an important role in creating functional diversity in the proteome (Buljan et al., 2013; Weatheritt et al., 2012; Weatheritt and Gibson, 2012). On an evolutionary timescale, mutations in IDRs with embedded binding motifs facilitate rewiring of interactomes and tuning of regulation through the gain and loss of binding motifs and modification sites, thereby enabling the emergence of complexity and new phenotypes (Beltrao et al., 2009; Buljan et al., 2012, 2013; Mosca et al., 2012). If mutations disrupt binding peptides and/or PTM sites, they may result in a disease outcome, e.g., Noonan syndrome, Usher syndrome, cherubism, and aberrant signaling in cancer (Guettler et al., 2011; Pajkos et al., 2012; Reimand and Bader, 2013; Vacic et al., 2012). Their evolutionary plasticity is also exploited by invading pathogens, such as viruses, which often mimic peptide motifs of their host proteins to hijack their cellular pathways (Davey et al., 2011; Hagai et al., 2014).

Identifying peptide motifs through sequence similarity searches is generally subject to high levels of statistical uncertainty, and they are elusive to identify experimentally. However, to gain a better and more complete description of the complex physiological and pathological processes of the cell, much more focus should be placed not only on identifying them, but also establishing their functionality through a combination of high- and low-throughput studies. A particular emphasis should be placed on identifying and functionally characterizing species-specific peptide motifs that are not evolutionarily conserved, but may be relevant, by studying them in the right biological context and in the appropriate model system. This endeavor will have rich reward not only in terms of a better understanding of cell physiology, but also in our ability to modulate the regulation of therapeutically important proteins.

ACKNOWLEDGMENTS

The authors thank Dr. Angela Bekesi for preparing Figure 2D and Tilman Flock, Natasha Latysheva, and Robert Weatheritt for their comments on our manuscript. This work was supported by the Odysseus grant G.0029.12 from Research Foundation Flanders (FWO) to P.T. M.M.B. acknowledges the Medical Research Council (MC_U105185859), HFSP (RGY0073/201), and EMBO Yi for funding his research.

REFERENCES


