Review

The application of modular protein domains in proteomics

Joshua A. Jadwin, Mari Ogiue-Ikeda, Kazuya Machida*

Department of Genetics and Developmental Biology, Raymond and Beverly Sackler Laboratory of Genetics and Molecular Medicine, University of Connecticut Health Center, 400 Farmington Avenue, Farmington, CT 06030, USA

ARTICLE INFO

Article history:
Received 24 March 2012
Revised 13 April 2012
Accepted 13 April 2012
Available online 21 April 2012

Edited by Marius Sudol, Gianni Cesareni, Giulio Superti-Furga and Wilhelm Just

Keywords:
Modular protein domain
Proteomics
Domainomics
Motif scanning
Domain scanning
Multiplex scanning

ABSTRACT

The ability of modular protein domains to independently fold and bind short peptide ligands both in vivo and in vitro has allowed a significant number of protein–protein interaction studies to take advantage of them as affinity and detection reagents. Here, we refer to modular domain based proteomics as “domainomics” to draw attention to the potential of using domains and their motifs as tools in proteomics. In this review we describe core concepts of domainomics, established and emerging technologies, and recent studies by functional category. Accumulation of domain–motif binding data should ultimately provide the foundation for domain-specific interactomes, which will likely reveal the underlying substructure of protein networks as well as the selectivity and plasticity of signal transduction.

© 2012 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

1. Introduction

Eukaryotic proteins are modular in nature. Many proteins contain independently folding globular domains capable of binding short peptide motifs even when both domain and motif are removed from the context of their full-length protein [1,2]. Modular protein interacting domains facilitate protein–protein interactions required for a diverse set of cellular processes including signal transduction and subcellular localization. Domains are categorized based on structural and sequence homology, with each domain family recognizing motifs with similar characteristics, such as phosphorylated tyrosine (pTyr) or proline rich sequences (Table 1).

The combination of modular domains within a protein contributes to its biological function by defining its protein interaction network. The post-translational modification (PTM) of amino acid side chains within a peptide motif can modulate domain–motif binding, providing the basis for the elegant and complicated protein signaling networks required for life [3]. Over the past decade, exploitation of a number of high-throughput proteomic technologies including increasingly sensitive mass spectrometry (MS) and protein microarrays has led to the dissection of vast protein interaction networks and the role of PTMs in altering network topology [4]. Accurate quantification of protein–protein interactions and PTM is now possible [5].

Because of the role of modular domains in assembling protein complexes, a significant portion of proteomics studies take advantage of modular domains as a means to assess protein–protein interactions (e.g., as bait in pull-down or probes in microarray). Here we refer to this modular domain-based proteomics as “domainomics.” While this may be a somewhat artificial segmentation, it is meant to draw attention to the potential of domains and their motifs as tools in contemporary proteomics. The emergence of domainomics as a unique sub-genre of proteomics raises a number of important questions: (1) What are the characteristics of modular domains as a research tool? (2) How are assays tailored to address specific scientific questions? (3) What technologies are available for exploiting domains as tools? (4) What lessons can be...
Table 1

<table>
<thead>
<tr>
<th>Recognition</th>
<th>Name</th>
<th>Size (a.a)</th>
<th>Domain</th>
<th>Protein</th>
<th>Affinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phosphotyrosine</td>
<td>PTB</td>
<td>100–150</td>
<td>36</td>
<td>31</td>
<td>nM–μM</td>
</tr>
<tr>
<td></td>
<td>SH2</td>
<td>~100</td>
<td>120</td>
<td>110</td>
<td>nM–μM</td>
</tr>
<tr>
<td></td>
<td>PTP(^b)</td>
<td>250–280</td>
<td>50</td>
<td>38</td>
<td>nM–μM</td>
</tr>
<tr>
<td></td>
<td>14-3-3</td>
<td>~250</td>
<td>8</td>
<td>8</td>
<td>nM–μM</td>
</tr>
<tr>
<td></td>
<td>BRCT</td>
<td>90–100</td>
<td>46</td>
<td>24</td>
<td>nM–μM</td>
</tr>
<tr>
<td></td>
<td>FF</td>
<td>50–60</td>
<td>25</td>
<td>6</td>
<td>nM–μM</td>
</tr>
<tr>
<td></td>
<td>FHA</td>
<td>65–100</td>
<td>31</td>
<td>31</td>
<td>nM–μM</td>
</tr>
<tr>
<td></td>
<td>MH2/DWB</td>
<td>~200</td>
<td>8</td>
<td>8</td>
<td>nM–μM</td>
</tr>
<tr>
<td></td>
<td>polo-box</td>
<td>~200</td>
<td>5</td>
<td>5</td>
<td>nM–μM</td>
</tr>
<tr>
<td>Polynucleotide</td>
<td>EVH1/WH1</td>
<td>~115</td>
<td>8</td>
<td>8</td>
<td>μM</td>
</tr>
<tr>
<td></td>
<td>GYF</td>
<td>~60</td>
<td>3</td>
<td>3</td>
<td>μM</td>
</tr>
<tr>
<td></td>
<td>SH3</td>
<td>~60</td>
<td>291</td>
<td>217</td>
<td>nM–μM</td>
</tr>
<tr>
<td></td>
<td>WW(^a)</td>
<td>38–40</td>
<td>88</td>
<td>48</td>
<td>μM</td>
</tr>
</tbody>
</table>
| 3. Application design

As affinity and detection reagents, modular domains and their short peptide ligands (motifs) can be used in three basic proteomic designs: motif scanning (motifs are scanned), domain scanning (domains are scanned), and multiplex scanning, each differing in their concept and execution (Fig. 1). “Motif scanning” surveys possible interaction partners containing a binding motif for a modular domain of interest. Typically, a modular domain is used as an affinity probe to either a library of synthesized peptides (e.g., SPOT arrays) or a whole proteome (e.g., far-Western and pull-down experiments). Motif scanning has been frequently used to define binding consensus motifs [11–13], to determine binding sites [14,15], and to identify interacting proteins within a particular cellular environment, such as during growth factor treatment [7,16]. In addition, a labeled domain can be used as a quantitative profiling tool for determining the presence or absence of modular domain binding sites in a group of cancer cell lysates or tissues [17].

“Domain scanning” uses a peptide binding motif as bait to screen a library of domains (e.g., domain microarrays) or a proteome (e.g., pull-down experiments). This approach is often used to determine binding partners when a putative domain binding motif is known to play an important role. For example, downstream effector docking sites are often examined using domain scanning [18,19].

In “multiplex scanning,” binary interactions between many domains and motifs are simultaneously analyzed in the same experimental system. By determining interactions between nearly all...
domains belonging to a functional subgroup and their putative peptide motifs in equivalent assay conditions, multiplex scanning allows for construction of domain interaction maps or classification of domain specificity [2,20,21]. Assays can be multiplexed by expanding the experimental scale of either motif scanning or domain scanning. For example, simultaneous analysis of 61 peptides and 159 domains using a microarrays-in-microplates system has been conducted [22]. Alternatively, use of one tube multiplex assays, in which multiple domains compete in binding, have also been reported [23-25]. In addition to these different designs, binding assays can be categorized as “in vivo or in vitro” and “direct or indirect.” Domain–motif interactions detected in vitro should be confirmed by in vivo methods, such as co-immunoprecipitation or in vivo pull-downs. Interactions observed in lysate-free systems, e.g., peptide arrays, and far-Westerns are considered direct, while immunoprecipitation or pull-down may detect direct, and indirect interactions due to bridging across precipitated proteins (e.g., Fig. 2E). Here, the presence of a known binding site for the bait within an identified protein suggests direct binding. In either case, direct binding should be confirmed using orthogonal assay systems. Importantly, indirect binding does not necessarily mean such an interaction is not specific; rather, it may represent an important functional subcellular complex [26].

4. Domain-based proteomics assays

In this section we briefly describe core technologies for domainomics, namely, in silico screening, forward- and reverse-phase arrays, phage display, bead array, and photocrosslinking, as well as other methods (Fig. 2).

4.1. In silico screen

A number of searchable databases are available to obtain sequences, domain structures, PTM sites, and known interactions including NCBI gene, Human Protein Reference Database (HPRD), SMART, Pawson lab website, PROSITE, Domino, Phospho.ELM, PhosphoSitePlus, MnM, and PhosphoNET. Prior to either “domain scanning” or “motif scanning” studies (Fig. 1), it is advisable to use web-based prediction tools to test potential domain–motif interactions, possibly eliminating unnecessary experiments and helping efficient experimental design [27]. Prediction tools include Scansite, NetPhorest, SMALL, and DomPep [28-30]. Prediction models are typically based around position specific scoring matrices (PSSMs), which are trained using a set of experimentally determined interactions. Based on the training set, PSSMs define the likelihood of observing a particular amino acid at a specific position within a peptide that is known to bind a domain [2,21]. Using this information, peptide binding scores can be calculated and used to predict the likelihood of novel peptide binding. However, the predictive ability of PSSM-based models may be limited by interaction training sets that are insufficient to define every position within the matrix, the lack of sufficient negative interaction training data, and their inability to incorporate the effect of residue interdependence on peptide binding [31-33].

4.2. Forward phase array

Analogous to antibody microarrays, modular binding domains can be immobilized on a solid surface and probed with fluorophore/enzyme-labeled peptides or proteins (Fig. 2A) [2]. The labeling can be direct, or indirect via a biotin-streptavidin or epitope-tag system [34,35]. Fluorescence-based detection systems are more amenable to quantitative analysis, while enzymatic detection may provide signal amplification [36]. Labels should be sufficiently spaced from motifs to minimize steric hindrance in ligand binding [37,38]. However, a major concern for the forward phase format is whether immobilization of recombinant domain proteins might disrupt the conformation and orientation required for ligand binding [39-41]. For instance, covalent binding provides strong association to solid surfaces, but the modification of chemical groups which occurs could interfere with domain binding to specific ligands. Thus, the immobilization method for a given domain must be carefully optimized with an appropriate evaluation method that incorporates modular domain-specific positive and negative controls, e.g., modified and unmodified peptides, and wild type and mutated domains. Currently domain arrays are mainly used to survey synthetic peptides or purified proteins. A prototypic domain array containing SH2, SH3, WW, PDZ, 14-3-3, FHA, PH, and EF domains on a nitrocellulose coated glass slide was introduced by Espejo et al. in 2002 [34]. In this assay, activity of immobilized SH3 domains was confirmed with peptides containing known SH3 motifs. In addition, they showed that endogenous proteins in cell lysates were bound by SH3 domains and could be detected using specific antibodies. Polyacrylamide hydrogel pads have also been used for protein domain immobilization [42]. Macbeath’s group introduced a multiplex forward phase microarray by spotting the full complement of human SH2/PTB domains onto aldehyde-modified glass [22]. This system enabled
affinity calculations for each domain–peptide pair based on the dependence of binding on ligand concentration, providing a more reliable measure than results obtained from a single ligand concentration.

4.3. Reverse phase array

In proteomics, reverse phase arrays are represented by lysate microarrays in which a large number of samples immobilized on a chip are probed with a specific antibody. Similarly, epitope-tagged or labeled domains can be used to probe a microarray or macroarray on which peptides, proteins, or lysates are immobilized (Fig. 2B). As opposed to forward phase domain arrays, there is little concern associated with immobilization process as domains remain in solution. However, binding affinity may be affected by the type of labeling or epitope-tag used. Reverse phase arrays are particularly suitable to screen large-scale libraries of peptide motifs to determine consensus sequences for domains [43,44]. Liu et al. performed a SPOT based array analysis, probing a set of 192 phosphorysine peptides with 50 SH2 domains. They found that the selectivity of SH2 domains is driven not only by permissive (favorable) amino acid motifs but also non-permissive neighboring residues [45]. Our lab has generated a multiplex reverse phase array system in which a group of sample spots are probed with many SH2 domains in register within a 96-well plate apparatus, allowing for the assessment of ~2000 domain–motif interactions in a single plate [8].

4.4. Phage display

Phage display provides a renewable and scalable source of peptides or domains for proteomics screens using conventional tools available to most labs [46]. In this assay, foreign DNA fragments are expressed as fusions with coat proteins on the surface of the bacteria phage (Fig. 2C). Typically, 5–45 amino acid peptides are expressed as N-terminal fusions, however C-terminal fusions are also technologically possible [47]. A process known as biopanning isolates high affinity clones through repetitive binding, washing, eluting and amplification [48]. High affinity clones are identified by sequencing of phage DNA and used to define the specificity of domain binding and natural domain ligands. Using this technology, large-scale multiplex scanning has been performed [49,50]. For example, 10^10 randomized peptides motifs were screened against 145 domains [21]. While phage display is typically used to express peptide motifs, Saksela’s group expressed a full complement of SH3 domains on the phage surface. Using this phage library, they revealed that some SH3 domains can bind ligands with nanomolar affinity, a much higher affinity than previously reported [51]. This library is now publicly available and widely used for domain scanning [52,53].

4.5. Bead array

Bead-based assays have been used successfully for motif scanning of multiple domain families including PHD, SH2 and PTB do-
mains [13,54,55]. These arrays provide an easy, cost-effective complement to microarrays. In bead-based assays, a large number of pooled peptides can be screened simultaneously without the need for prior spatial separation of peptides [56]. Typically, combinatorial peptide libraries are synthesized on beads and probed with epitope-tagged modular domains (Fig. 2D). Interactions are detected using various methods such as “on bead western”, FACS, or magnetic nanoparticle pull-down; peptides are then identified by MS [13,56,57]. Generally, bead-based binding assays do not excel at quantification; as this requires a secondary assay in an orthogonal format using resynthesized peptides. To circumvent this, Astle et al. developed a bead-to-microarray screen. Peptides with positive interactions are cleaved from beads, divided into two fractions. One fraction is spotted on microarray for quantification by domain titration, and the other is used for sequence identification. By virtue of throughput and quantification capability, this system is thought to be suitable for screening of domain–peptide inhibitors [56].

4.6. In vivo photocrosslinking

In vitro substitution of a specific amino acid residue within a protein with an unnatural amino acid by manipulation of the ribosomal machinery was first described in late 80s [58,59]. Decades later, using bacterial amber suppressor tRNA and aminoacyl-tRNA synthetase specific for an unnatural amino acid, it is now possible to express mutated proteins containing unnatural residues harboring specific functional side chains in mammalian cells [60–62].

Incorporation of photo-reactive cross-linking amino acids such as p-benzoyl-l-phenylalanine (pBpa) into a modular domain of interest is particularly promising when studying transient domain–motif interactions (Fig. 2E) [63]. It allows for the capture of domain binding proteins whose interactions may not survive extensive washing during immunoprecipitation. Using this scheme, a photo-reactive domain and its ligands are cross-linked using UV-illumination, immunoprecipitated using an affinity tag, and subjected to LC–MS/MS. Recently, photocrosslinking methods were used to study in vitro SH2 or SH3 domain interactions [64,65]. These studies led to the discovery of previously unidentified substrates. This approach is suitable for motif scanning by trapping proteome-wide interactors in vivo. However, a potential weakness of this technology is that mutation sites and cross-linker side-chain moieties may interfere with physiological ligand interactions [64]. The mutation site should be carefully determined so that the cross-linking site is close enough to the binding surface without causing steric interruption of ligand binding [65]. Because of the availability of a range of unnatural amino acids with various chemical functionalities and their potential capabilities in studying PTM-dependent protein–protein interactions in real time in vivo, studies using these technologies will clearly benefit domainomics [66]. For example, incorporation of a metabolically stable analog of phosphotyrosine could serve as a new tool in pTyr-interacting domain screening [67].

4.7. Other technologies

The yeast two-hybrid system (Y2H) has been an useful method for identifying novel domain interactions [68–70]. Modified Y2H assays provide a method for screening PTM-dependent interactions [71,72]. However, some Y2H screens have high false positive and negative rates [73]. ELISA-based systems have also been used in domainomics [74,75]. Fluorescent polarization has proven to be an useful method, capable of real-time in solution kinetic measurements, for drug screens or validation of microarray data [2,76]. AlphaScreen, a commercially available bead-based proximity assay, also shows promise as a sensitive tool for assessment of domain–motif interactions and drug discovery [77]. In this assay, a pair of proteins are indirectly labeled with either donor or acceptor microbeads. If the proteins are interacting, irradiation of a donor bead triggers activation of a nearby acceptor bead, amplifying the chemiluminescence signal, which provides greater sensitivity and dynamic range [78,79]. Proximity ligation assay is becoming a popular tool due to its high sensitivity and specificity in analyzing protein–protein interactions [80]. Commonly this approach is used in situ where cells or tissues are fixed and incubated with a pair of oligonucleotide-labeled antibodies for two proteins that are hypothesized to interact. Interactions are detected via rolling circle amplification [81]. An in-solution proximity ligation assay has also been reported as a multiplex protein detection tool requiring only 1 μl per sample [82].

5. Survey of domainomics

In this section we focus on representative domains in functional categories, how they are used in domainomics studies and what lessons can be learned from these applications.

5.1. Phosphotyrosine recognizing domains

At this time, there are 3370 experimentally verified tyrosine phosphorylation sites in eukaryotic proteins according to Phos.ELM [83]. The majority of pTyr-binding proteins contain SH2 domains, PTB domains, or both (Table 1) [84,85]. SH2 domains are approximately 100 amino acids in length and bind in a pTyr specific manner via a conserved binding pocket [12,86,87]. PTB domains are 100–150 residues long, are found in scaffold proteins such as Shc and IRS-1, and vary in their dependence on phosphorylation [88]. Several other pTyr binding modules, such as the PKC delta C2 domain and pyruvate kinase M2, have also been reported [15,89].

pTyr-dependent reversible interactions are crucial in propagating signals from activated receptor tyrosine kinases to downstream effectors. For example, upon stimulation, ErbB/EGF receptors auto-phosphorylate multiple Tyr sites on their cytoplasmic tails that serve as docking sites for SH2/PTB containing downstream effectors [90,91]. To focus our methodological discussion, we outline independent domainomics studies on ErbB family receptors. Blagoev et al. used the Grb2-SH2 domain as bait to pull-down the EGFR interactome in HeLa cells [7]. They employed the SILAC method to quantify EGF-dependent interactions. Their study identified 228 proteins, of which 28 proteins, including 6 SH2 proteins and the novel interacter CD59, were selectively enriched upon stimulation. Schulze et al. conducted a SILAC pull-down with pTyr–peptide bait corresponding to all 89 ErbB cytosolic tyrosines, followed by MS [92]. Independent peptide pull-down MS identified 50 proteins, including 10 unique proteins (9 SH2/PTB proteins), most notably STAT5 as an EGFR interactor. EGF and ErbB4 shared diverse interaction partners whereas ErbB2 and ErbB3 preferentially bound Shc and PI3-kinases, respectively. Jones et al. performed a SH2/PTB domain microarray with 33 tyrosine residues of ErbB receptors [22]. This approach identified numerous known and previously undocumented SH2/PTB-pTyr interactions, including 54 involving EGFR, 59 involving ErbB2, 37 involving ErbB3, and 8 involving ErbB4. Binding profiles of ErbB receptor sites were distinct: EGFR and ErbB2 binding became more promiscuous at high ligand concentration, while ErbB3 binding was independent of ligand concentration. A follow up study showed that the ErbB4 receptor displayed a more selective binding profile than the other receptors [93]. This observation appears to be inconsistent with the promiscuous profile of ErbB4 in Schulze's pull-down data [92].

Reasons for this discrepancy are unknown, although the methodologies used in these three studies were quite distinct. First, the
SH2 pull-down performed by Blagoev et al. is capable of detecting direct and indirect binding partners of EGFR in vivo. Second, while Schulze's pTyr-peptide pull-down is a sensitive approach, some tyrosines used in the study may not be phosphorylated in vivo. Thus, observed binding may or may not be physiological. Third, while the sensitivity of the domain microarray approach should very high because domain–peptide interactions occur at fixed concentrations in a buffered solution without any competition or interference from other molecules, the activity of recombinant domain proteins can be lost in vitro due to misfolding, degradation, etc., resulting in poor reporting of physiological binding. These methodological differences should be taken into account in assay design and data interpretation. Limitations of these methodologies suggest that one approach is insufficient to correctly identify all physiological interactions. To illustrate this, we have summarized observed interactions from two studies mentioned above and Scansite predictions for four pYXN motifs of ErbB4 in Table 2. Src binding to pY1188 is shared among three and Grb2 is promi-

Table 2

<table>
<thead>
<tr>
<th>ErbB4 site</th>
<th>Sequence</th>
<th>Pull down</th>
<th>Microarray</th>
<th>Scansite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1162</td>
<td>PMRDKPKQEpYNPVEEN</td>
<td>Grb2</td>
<td>Vav2; RASA1; Arg</td>
<td>PLCg; Grb2; Abl</td>
</tr>
<tr>
<td>Y1188</td>
<td>DLQALDNPeypYNASNGP</td>
<td>Grb2; She</td>
<td>She; Arg</td>
<td>Grb2; She</td>
</tr>
<tr>
<td>Y2020</td>
<td>NGPfKAEDEyypYNPLyL</td>
<td>Grb2</td>
<td>Syk</td>
<td>Grb2; PLCg</td>
</tr>
<tr>
<td>Y2080</td>
<td>EDEYENFLypYNFANT</td>
<td>Grb2</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

5.2. Proline rich sequence recognizing domains

Polyproline sequences are thought to provide a favorable domain–motif interface due to the stability of their helical confirmation, hydrophobicity, and presence on the surface of proteins [94,95]. As a consequence, multiple domain families involved in subcellular localization and the assembly of multiprotein complexes recognize proline rich motifs [96], SH3 domains were the first member of this group, discovered in 1988, followed by WW, EVH, and GYF domains (Table 1) [97–100]. It is worth noting that the first demonstration of polyproline recognition came from an early motif scanning experiment; a cDNA expression library was screened using a GST-tagged Ab1 SH3 domain as bait. Subsequent deletion analysis using the positive clones led to the finding [11].

The specificity of these domains to short proline-rich peptides is generally modest, with affinities in the low micromolar range (Table 1). However biochemical and structural studies suggest greater selectivity may be conferred by a larger binding surface [101,102]. Given the large number of proline-rich sequence recognizing domains and the important role of selectivity in signaling (Table 1), it quickly became apparent that other modular domains such as, FHA, WW and BRCT also share this behavior (Table 1) [103], it quickly became apparent that other modular domains such as, FHA, WW and BRCT also share this behavior (Table 1) [103–106].

While it was surprising when dimeric 14-3-3 proteins were found to be genuine phosphoserine/threonine recognition modules [103], it quickly became apparent that other modular domains such as, FHA, WW and BRCT also share this behavior (Table 1) [104–106].

14-3-3 family proteins have attracted scientists’ attention because of their involvement in important cellular processes, such as signal transduction, and in disease [107]. 14-3-3 domainomics studies have been conducted for specificity mapping by peptide screening [108,109], and protein identification by pull-down MS [110]. To date, several hundred interacting proteins have been reported including Raf and BCR [99], however direct binding has yet to be confirmed for many of them [100]. One perplexing problem, however, is that 14-3-3 interactomes produced by many studies display remarkably little overlap (20–40%). This may be attributed to differing experimental conditions such as enrichment methods, cell type, in vivo vs. in vitro, direct vs. indirect binding, or type of bait used [110–113].

Another plausible explanation is isoform dependence. 14-3-3 isoforms have been implicated to play unique roles in human disease, suggesting that each may bind a distinct set of ligands [114,115]. Intriguingly, experimental data to date seem to be insufficient to explain apparent isoform selectivity in vivo. All isoforms share high affinity motifs RXSXP and RXD/YXpSXP with some variation [116], and a peptide scanning study failed to detect any isoform-specific phosphopeptide binding [117]. Structural observations also demonstrated that the conformation of 14-3-3 hetero- and homo-dimers were largely similar indicating their structural rigidity [116]. Thus, it has been hypothesized that 14-3-3 dimers act as rigid molecular scaffolds keeping an enzyme in its active conformation through a bidentate interaction [118]. Nevertheless, isoform specific ligands and minor conformational variability and flexibility have been reported [116,119]. Therefore, it is likely 14-3-3 proteins largely share their ligands to exert broad regulatory functions, while playing selective roles by interacting with a few...
isoform-specific ligands. To address this issue, multiplex scanning domainomics, comparing all isoforms in equivalent conditions, would be informative.

BRCT, FHA, and 14-3-3 work cooperatively in the DNA damage response. DNA strand breaks activate the recruitment of serine/threonine kinases to the damage site, where they phosphorylate DNA repair proteins, transcription factors, protein kinases, and scaffolds [120,121]. Phosphorylation signals are then “read” by proteins containing BRCT-1, FHA, and 14-3-3 domains. This process is organized as a hierarchy of domain–phosphosite binding interactions where specificity of kinases and domains are tightly regulated to prevent molecular signals from getting crossed [122]. For example, ATM kinase preferentially phosphorylates serine sites that are favored by MDC1 BRCT domains, while Chk kinase phosphorylates the motif recognized by 14-3-3 [123,124]. In addition, other PTMs such as ubiquitination and methylation also play roles in the DNA damage response [125].

The link between BRCT domains and pS/pT-motifs was initially illuminated by a well-designed multiplex scanning study [106]. To identify novel phosphorylation-dependent domains involved in the DNA damage response, Manke et al. performed a pull-down assay against a cDNA expression library using an oriented phosphopeptide library mimicking ATM kinase substrates as bait. Proteins that bound in a phosphorylation-dependent manner were visualized by gel electrophoresis, which resulted in identification of the multi-BRCT domain-containing protein PTIP. Interestingly, they found that C-terminal tandem BRCT domains of PTIP function as a single module for phospho-specific binding. This “library to library pull-down” method is advantageous in that identification of PTM-recognizing domains is feasible without specifying a motif sequence and types of domains. Recently, this method has been combined with SILAC-MS technology leading to identification of a novel pTyr-binding module [89].

Highly dynamic and selective interactions at domain-PTM sites in the DNA damage response raise many questions: How many tandem BRCT domains are phosphorylation-dependent? How many domains are highly selective or promiscuous? Why are BRCT domains generally confined to DNA damage response signaling? While some studies have begun to address these questions [126], systematic domain binding analyses should provide more comprehensive answers.

5.4. Carboxyl terminus binding/PDZ domains

PDZ (PSD-95, DLG-4, ZO-1) domains are carboxyl terminal binding modules found in proteins generally considered to play a role in maintenance of cell polarity and adhesion [127]. Domains were initially divided into two major classes based on oriented peptide screening [128]. A third outlier class has also been proposed based on its lack of consensus with groups I and II [129]. Recently two large-scale studies have been undertaken to more explicitly define the specificity space of PDZ domains. Tonikian et al. screened a phage display library harboring 10 billion random C-terminal peptides against 82 PDZ domains from human and Caenorhabditis elegans, revealing that greater than 90% of PDZ domains fit into one of sixteen classes [21]. Stiffler et al. performed a domain microarray screen of 157 mouse PDZ domains probed with 217 C-terminal peptides. Surprisingly, they found that PDZ domains do not fall into discrete specificity classes; rather they lie on a continuum, evenly dispersed across the specificity space [2]. These data suggest that simplified classes based on consensus motifs are not suitable to precisely predict novel PDZ domain–ligand interactions.

Several groups have applied these binding datasets to develop prediction models. Using the microarray data, Chen et al. developed a modified PSSM that integrated domain/motif structural contact map information and was capable of predicting novel PDZ domain interactions across multiple species [31]. The sensitivity of this model was confirmed by another group using an independent human data set, although the false positive rate was shown to be higher than originally described [32]. Hui et al. incorporated Chen’s contact map information and trained a machine learning method using both phage and microarray data [130]. This approach significantly reduced the false positive rate. Instead of structural information, Li et al. used evolutionary relatedness of domain sequences and a PSSM to compare different domains, which resulted in increased predictive accuracy [30]. The model, referred to as DomPep, is publicly available online. Together, multiple efforts have been made to improve PDZ interaction prediction models. It is not possible to rank different prediction tools without a comparable value such as sensitivity or specificity, although we observed that ROC area under the curves for these models generally fell between 0.75 and 0.9, indicating good prediction accuracy. However, the obvious problem is that all models have been trained using two limited data sets based on PDZ domains from only three species. Because models are greatly affected by their training data set, broad validated experimental data is a requisite for further development. In addition, integration of protein expression and PTM data [131] could also provide more meaningful predictions of protein–protein interactions in vivo.

5.5. Methyllysine binding domains

PHD (Plant Homeodomain) fingers and Royal Family domains (Chromo, Tudor, MTBD, and PWWP) recognize methlyllysine and methylarginine on histone tails in the context of flanking residues [132]. These domains are thought to act as “readers” of the histone code which translate histone PTMs into biological outcomes in a process that in some ways parallels reading of tyrosine phosphorylation by SH2 domains [133,134].

Intriguingly, over 392 proteins have been shown to contain at least one of these five domain families, yet only ~10 methylated residues have been identified on human histones [135,136]. This raises the question: How do so many domains specifically interact with so few ligands? Several domainomics studies have provided some clues. A domain scanning approach employing a microarray of 109 PHD fingers and Royal Family domains, probed with histone tail peptides identified Tudor and MBT domains whose binding is dependent on the degree of lysine methylation (mono, di, tri) [20]. Further, peptide pull-down experiments employing a set of PHD finger, Tudor and WD40 domains showed that binding to methyllysine could be significantly affected by nearby arginine methylation [137]. To test the effect of other PTMs, Garke et al. performed a motif scanning assay in which a histone-3 N-terminal peptide bead library containing over 5000 randomized PTM combinations was probed with five PHD fingers [13]. The screen identified a number of PTMs, including arginine methylation and threonine phosphorylation, which modulated PHD-methyllysine binding via switch (binary binding) and rheostat-like (variable binding) mechanisms. Subsequent SILAC-MS (domain scanning) and peptide array (motif scanning) experiments added support to the importance of phosphorylation in methyllysine binding [138,139]. Taken together, these results suggest combinatorial PTMs provide specificity for domain binding to histone tails, adding credence to the theory that neighboring PTMs confer widespread regulation of histone tail effector binding [140]. A complete understanding of this mechanism will require further studies including a greater number of methyllysine binding domains.

5.6. μ-Propeller family

WD40 repeats are one of the most abundant domain families found in eukaryotic proteomes (Table 1). They rank as one of the
top interacting domains in human interactome data sets [141], and are found in proteins which play roles in a diverse set of biological functions, including signal transduction and vesicular trafficking [142,143]. It is hypothesized that WD40 domains act as a scaffolds for multiprotein complexes. These domains differ from “classical” modular protein domains in multiple ways: Typically WD40 repeats fold into a non-catalytic, seven blade beta-propeller that functions as an independent module. These propellers share little sequence conservation, and studies have shown they have PTM-dependent (e.g., phosphorylation, ubiquitination, and methylation) and independent substrates [144,145]. In this sense, WD40 propellers behave in a similar manner as antibodies, providing a framework on which specificity for different ligands can be built.

The unique features of WD repeats could impede application of conventional domainomics. For example, motif scanning, to define binding specificity for poorly understood WD40 proteins, would require an enormous peptide library containing unrelated motifs and different types of PTMs. Further, recombinant WD propeller proteins tend to be insoluble proving a hurdle for in vitro studies [146]. As a result, a systematic inspection of WD40 ligand selectivity has yet to be undertaken to date. Nevertheless, it has been shown that SPOT array scanning was able to define the consensus motif of the Cdc4 WD40 propeller, suggesting a multi-domain study is feasible. In addition, their deep involvement in the human interactome and multiple PTM-dependency suggests a possible role as signaling gatekeepers. For instance, there is a switch-like interaction between WD40 repeats of Cdc4 and multiple phosphorylation sites of Sic1; only when all six sites are phosphorylated, can the WD40 repeats bind Sic1 leading to cell-cycle progression [147]. Thus, it would be interesting to exploit WD40 repeats as molecular probes to distinguish fluctuations in cellular signaling pathways, e.g., as biosensors for a multiply phosphorylated site.

6. Perspectives

6.1. Domainomics-reducing the complexity

The major premise of domainomics is that in vivo protein networks are generally governed by interactions between modular protein domains and short peptide motifs, and that these interactions are generally reproducible in an in vitro binding assay. Unfortunately, some information is lost at the expense of the convenience provided by this overly simplified view. First, non-domain-dependent binding that relies on the reciprocal large surface interactions between proteins will be missed [148]. Second, cooperative interactions involving multiple domains might also be missed [149]. Third, transient enzyme–substrate interactions may not be captured. Thus domainomics does not provide a complete picture of regulatory protein networks. Nevertheless, domains and motifs are valuable tools for surveying specific domain–motif interactions, and thus widely used in motif and domain scanning, as outlined above.

We believe, however, the power of domainomics can be fully exerted in multiplex scanning formats, especially when many domains are employed. By incorporating all functionally similar domains into a high-throughput system, it is possible to dissect individual layers of a complex full-length protein interaction network. For example, the SH2/WW domain–ligand map is a subset of the interactome involving only proline-rich sequences [43,150]. Such a sub-network map is not only useful to overview the distribution of binding specificity and promiscuity, but also to predict effects caused by interruption or overexpression at a particular interaction node and in rewiring networks [151,152]. Likewise, functional layers based on phosphotyrosine, phosphoserine, and methyllysine dependent domains should provide distinct interaction maps. Usefulness of this approach is visible even when using a subset of domains or motifs. For example a SH2 and PTB domain macroarray provided a pTyr-dependent effector map for ErbB/EGFR family members, revealing a qualitative difference among them [22]. Another study compared interaction profiles of selected WW domains and suggested a connection may exist between binding motifs and specific subcellular systems [26]. Taken together, multiplex scanning domainomics offers a unique tool to dissect interactomes in a way not obtainable through non-domain proteomics.

6.2. Genome-wide domainomics

Given the potential of domainomics and the availability of advanced technologies and bioinformatics, is it now possible to construct many different interactomes based on modular protein domains and their binding proteins? Unfortunately, because the domainomics toolbox is insufficient, the answer is No. Publicly available complete domain libraries are limited to only a few domains, e.g., SH2 and SH3, precluding the construction of comprehensive domain-interaction maps. There are, however, independent efforts to create such libraries and interaction maps for a number of domains [2,43,74]. In the long term though, redundant collection of clones by individual research groups is highly inefficient. Ideally, the modular domain research community should work together to bank domain constructs.

Along with modular interaction-dedicated domains, enzyme catalytic domains may be a potential addition to the toolbox. PTM-dependent protein interactions are governed by the specificities of both binding domains (readers, e.g., SH2s) and enzymes (writers and erasers, e.g., tyrosine kinases and phosphatases) [153]. Although enzymes bind substrates transiently, making them inappropriate for use as affinity reagents, stabilization of binding by mutagenesis or cross-linking is feasible at least in part [154,155]. These enzymes, as well as modified interaction domains [155,156] and domain-specific antibodies [157,158], may strengthen domainomics by revealing further functional layers of protein–protein networks.

7. Concluding remarks

In this review, we highlighted a portion of proteomics (domainomics) focusing on how modular protein interacting domains are popularly used as a means to address specific scientific questions. Methods involving domains and their motifs are increasingly playing a role in the proteomics era, powered by the availability of genome data, domain prediction tools, and high-throughput technologies. Accumulation of domain–motif binding data will ultimately lead to a domain-based interactome, providing insights into the underlying structure of protein networks and the selectivity and plasticity of signal transduction.

Acknowledgements

The authors gratefully acknowledge Jonathon Ditlev, Peter Nollau, and Bruce Mayer for critical reading of the manuscript, and Joshua Kritzer for helpful discussions. This work was supported by Neag Comprehensive Cancer Center, American Cancer Society Institutional Research Grant and NIH grants RC1 CA146843 and U01 CA154966.

References


