FEBS Letters 586 (2012) 2631–2637

journal homepage: www.FEBSLetters.org

Review

Elucidation of the binding preferences of peptide recognition modules: SH3 and PDZ domains

Joan Teyra [a], Sachdev S. Sidhu [a,b,c], Philip M. Kim [a,b,c,d,*]

[a] Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Canada ON M5S 3E1
[b] Banting and Best Department of Medical Research, University of Toronto, Toronto, Canada ON M5S 3E1
[c] Department of Molecular Genetics, University of Toronto, Toronto, Canada ON M5S 3E1
[d] Department of Computer Science, University of Toronto, Toronto, Canada ON M5S 3E1

ARTICLE INFO

Article history:
Received 29 March 2012
Accepted 15 May 2012
Available online 9 June 2012

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

Keywords:
Domain–peptide interactions
Protein interaction networks
Protein interfaces
Peptide recognition modules
Machine learning

ABSTRACT

Peptide-binding domains play a critical role in regulation of cellular processes by mediating protein interactions involved in signalling. In recent years, the development of large-scale technologies has enabled exhaustive studies on the peptide recognition preferences for a number of peptide-binding domain families. These efforts have provided significant insights into the binding specificities of these modular domains. Many research groups have taken advantage of this unprecedented volume of specificity data and have developed a variety of new algorithms for the prediction of binding specificities of peptide-binding domains and for the prediction of their natural binding targets. This knowledge has also been applied to the design of synthetic peptide-binding domains in order to rewire protein–protein interaction networks. Here, we describe how these experimental technologies have impacted on our understanding of peptide-binding domain specificities and on the elucidation of their natural ligands. We discuss SH3 and PDZ domains as well characterized examples, and we explore the feasibility of expanding high-throughput experiments to other peptide-binding domains.

1. Introduction

Cellular processes are dynamically orchestrated by a precise interplay of protein–protein interactions. Many proteins contain peptide recognition domains that mediate the assembly of diverse stable or transient biological complexes to coordinate specific biochemical functions in a wide variety of processes. These modular domains recognize relatively short peptide sequences containing a core structural motif. For example, WW domains recognize proline-rich peptides, EH domains bind to peptides containing the NPF motif, and SH2 and PTB domains bind to peptides containing a phosphorylated tyrosine [1]. Most domains within the same family recognize distinct binding partners. These distinct specificities are usually determined by key residues flanking the core binding motif [2,3]. PSD95-Discs large_ZO1 (PDZ) and Src-homology-3 (SH3) domains are two of the most extensively studied peptide recognition modules (1657 and 2527 hits in PubMed, respectively).

PDZ domains assemble intracellular complexes principally by recognizing certain C-terminal sequences. The specificity is mediated by interactions between ligand side chains and the PDZ domain binding surface [4]. Early studies grouped PDZ domains into two main specificity classes based on distinct ligand signatures: Class I (X[T/S]Xϕ$_{COOH}$) and Class II (XϕXϕ$_{COOH}$), where X is any residue and ϕ is a hydrophobic amino acid [5,6]. In addition, less common classes of PDZ domain binding specificities, such as Class III recognizing the motif X[ED]Xϕ$_{COOH}$, were also identified [7]. However, subsequent studies have shown that the PDZ binding cleft can interact specifically with up to seven C-terminal ligand residues, enabling differentiation between biologically diverse ligands [8].

SH3 domains bind to proline-rich sequences containing a core PXXP motif flanked by a positively charged residue [9,10]. Class I domains bind to ligands conforming to the consensus +XXPXXP (where + is either Arg or Lys), while Class II domains recognize PXXPX+ motifs and bind to ligands in the opposite orientation [11,12]. More recently, a number of alternative SH3 domain binding motifs have also been identified, highlighting a wider breath of SH3 specificities [13–16].

PDZ and SH3 domains are widespread. In the human proteome alone, 364 PDZ domains and 536 SH3 domains have been identified, and these domains mediate diverse cellular functions and compete for thousands of potential ligands [17]. The understanding of selective ligand recognition requires the discovery and comparative analysis of binding motifs for a comprehensive set of these

* Corresponding author at: Department of Computer Science, University of Toronto, Toronto, Canada ON M5S 3E1. Fax: +1 416 978 8287.
E-mail address: pi@kimlab.org (P.M. Kim).

0014-5793/$36.00 © 2012 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
http://dx.doi.org/10.1016/j.febslet.2012.05.043

recognition modules. Here, we review the advances in our understanding of peptide-binding specificities of modular domains, with special emphasis on PDZ and SH3 domains. We also introduce methods for binding specificity and natural ligand prediction, and show the improvements achieved by including the peptide-binding domain specificity motifs. Finally, we highlight examples for practical usage of domain–peptide interactions for rewiring networks and protein inhibition by synthetic design.

2. Technological advances in the study of peptide-binding specificities

The analysis of peptide-binding domain specificities on a large scale has been enabled by the development of high-throughput strategies to complement microarray and phage display technology. These techniques can work in two directions, either fixing the peptides to the plate to be interrogated with solution-phase proteins, or the opposite. Microarrays can explore hundreds of natural peptidic sequences at once, and give semi-quantitative readouts [18]. Alternatively, phage display can explore a much larger diversity of sequences using bacteriophages to display libraries of up to 10 billion random peptides as genetic fusions to phage coat proteins. In this section, we will highlight the most relevant research that has taken advantage of these two powerful techniques to study the peptide-binding specificities of PDZ and SH3 domains.

Microarrays provide a multiplexed approach for the parallel examination of peptide substrates of large numbers of proteins. Originally, this technique was developed with the aim of identifying binding partners. For example, Hall and coworkers developed a proteomic array of 96 putative Class I PDZ domains derived from cytoplasmatic proteins to identify those domains that recognize the C-terminal PDZ binding motifs of GPCR proteins [19]. The positive PDZ-mediated protein–protein interactions were subsequently confirmed by co-immunoprecipitation and inmunofluorescence co-localization. In a later study, MacBeath and coworkers developed a strategy for constructing a multidomain selectivity model for mouse PDZ domains [20]. They prepared microarrays of 157 mouse PDZ domains and used them to survey interactions with over 200 fluorescently labeled synthetic peptides representing C-terminal sequences of mouse proteins. The positive hits were retested and quantified by fluorescence polarization assays, thereby correcting array false positives. The resulting data were used to train a predictive model of PDZ domain selectivity. The model highlighted putative array false negatives, which were tested by fluorescence polarization, and the corrected data were used to retrain the model. After three cycles of prediction, testing, and retraining, the refined model was used to predict PDZ domain–protein interactions across the mouse proteome.

As an alternative approach, microarrays of short peptides can be prepared and interrogated with solution-phase proteins. Standardized methods now exist to synthesize in parallel thousands of peptides bound to a cellulose membrane in a microarray format, such as 'SPOT' synthesis [21]. Peptide microarrays are particularly useful when the objective is to screen one or a few proteins against a large number of potential binding peptides. In an interesting variation of SPOT synthesis, Boisguerin et al. developed an efficient way to prepare microarrays of inverted peptides displaying their C-termini [22]. This method permitted them to study interactions mediated by the PDZ domains of AF6, SNA1 and ERBIN protein against a peptide library comprising 6223 C-termini of human proteins [23]. On the basis of the ligand preferences detected for these PDZ domains, they quantified the binding affinity contribution of each amino acid position, and they predicted their putative natural binding partners. SPOT synthesis has also been applied in combination with phage display to discover all peptides in the yeast proteome that had the potential to bind to eight SH3 domains [24]. Five classes with partially overlapping specificities were identified, where domains bind to a large number of common targets with comparable affinity [25].

Finally, phage display technology provides an accurate way of studying the specificity of peptide-binding domains [26,27]. This technology uses bacteriophages to display libraries of up to 10 billion random peptides as genetic fusions to phage coat proteins [28]. After repeatedly incubating the phage particles with a domain, and washing away non-interacting phages, a specificity profile consisting of a set of strongly interacting peptides can be retrieved by sequencing the phage-encapsulated DNA. The sequences of the binding clones are then aligned to create a position weight matrix (PWM) that describes the domain binding specificity. Each matrix column captures the amino acid binding preference of a domain at a ligand position as a probability distribution. Residues are usually assumed to contribute independently to binding, simplifying our understanding of domain–peptide interactions. However, peptides interacting at one specific binding region may display many correlated positions, where an amino acid at one position can influence another one [29]. While all the above efforts were aimed at predicting single specificities usually expressed as PWM, recent work has showed that many domains exhibit multiple specificities that can be expressed as mixture models of several PWMs [30] (Fig. 1a).

The possibility of using large-scale phage-displayed libraries has increased the throughput of phage display by several orders of magnitude [28]. In addition, recent advances in next-generation sequencing has drastically extended the number of different ligand sequences obtained, improving the generation of high resolution binding profiles [31,32]. In consequence, these technological improvements have permitted the possibility to catalog and derive specificity maps for many of the SH3 domains and PDZ domains from different proteomes: *Saccharomyces cerevisiae* [26,33], *Caenorhabditis elegans* [27], mouse [20] or human [27,33]. These maps have revealed the versatility and highly specific nature of these modules and their interactions. Moreover, these high resolution analyses have revealed that each domain exhibits specificity across multiple ligand positions, including not only the core motif but also flanking positions. Accordingly, Sidhu and coworkers found that approx. 90% of PDZ domains fit into 16 distinct specificity classes, and the remainder represent unique specificities. Similarly, MacBeath and coworkers had proposed that the domains lie on a functional continuum and it appears that their binding selectivity has been optimized across the proteome in order to minimize cross-reactivity [20]. Tonikian et al. also revealed that essentially all PDZ domains recognize the last three ligand positions (0, −1, −2), the majority recognize positions −3 and −4 and some recognize positions −5 and −6 [27]. In the case of SH3 domains, Tonikian et al. have found that most of yeast SH3 domains fall into the canonical classes I and II [33]. However, they also uncovered several SH3 domains with specificity profiles that clearly deviate from the two canonical classes. These Class III binders show a preference for poly-proline ligands, without the requirement for flanking charged residues (Fig. 1b). Finally, they identified other domains that show class promiscuity such that they cannot be simply classified because they exhibit unique specificity profiles that differ from the canonical binding motifs [33].

3. Binding specificity predictions: machine learning and biophysics

Thousands of PDZ and SH3 domains are spread across eukaryotic and eubacterial genomes [17,34]. Obtaining experimental binding specificities for all domains in a family is unfeasible due to common cloning, expression, solubility or phage-related

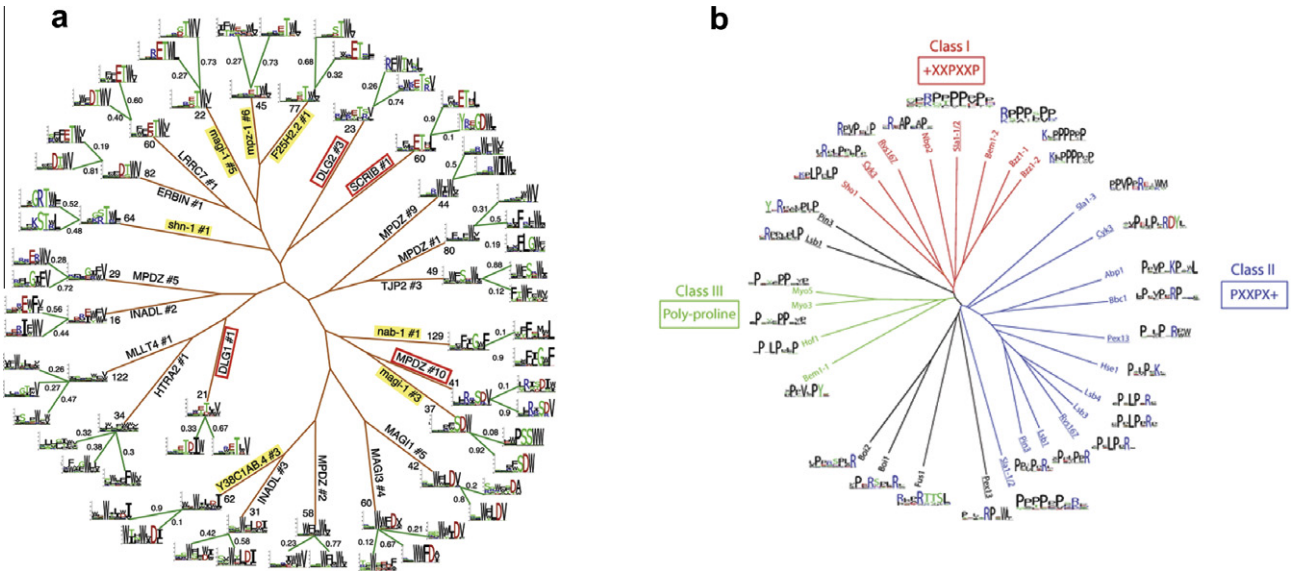

Fig. 1. Specificity maps obtained for (a) PDZ domains and (b) SH3 domains. PDZ signatures are obtained from Gfeller et al. [30] and the SH3 ones from Tonikian et al. [33].

problems. Therefore, the development of good computational predictors of peptide-binding preferences for domains without binding data is indispensable. Large scale experiments have revealed a relationship between binding site identity and specificity for PDZ and SH3 domain members [27,33]. Domain pairs with binding site sequence identities greater than roughly 70% have specificity profiles with equivalent similarity to those within a specificity map class, whereas the relationship is unclear for pairs with lower sequence identity. These observations suggest that, given a representative set of domain binding specificities, a large set of the family interaction space could be inferred. Although this might be a general feature for positive amino acid selection, the underlying basis for selectivity is not only the ability of domains to recognize permissive amino acids but also non-permissive ones that oppose binding in the vicinity of the core motif. Therefore, single amino acid changes in a binding site may also change the specificity map [35].

Computational methods for peptide-binding specificity prediction fall on a spectrum spanning from statistical and machine-learning approaches to biophysical and computational chemistry methods. Computational machine learning methods rely on the previously observed behavior of a molecular system, which is then described in a predictive mathematical model. In order to predict binding partners, data from actual peptide-binding experiments are used to train a classification algorithm to discriminate between binding and non-binding peptides. The models are trained using a set of known binding sequences, where the domain contact residues involved in binding have to be determined. This is achieved by sequence alignment of the query domains with a reference structure of a domain–peptide complex with a well understood binding mode.

One of the first studies developed a variant of the PWM that contained weights describing the relative preference for amino acids at positions in the ligand compared to the other domains they modeled [36]. Later, they used a naive Bayesian model that included several biological features to predict binding partners of 83 PDZ domains in the mouse proteome [37]. The features used in this work comprise weighted scoring matrices for all combinations of the possible PDZ and ligand residues, binding affinity data, and other binary information. They were able to identify genuine ligands for PDZ domains lower than 33% sequence similarity to the training data. However, the method did not perform well at predicting the binding energies of the interactions. Another group used a machine learning method called a support vector machine to predict PDZ domain interactions and achieved an accuracy of 80% with a false positive rate of only 39% [38], highlighting the success of these approaches. Finally, Hawkins et al. generated structural models for a dataset of PDZ–peptide complexes using threading techniques in order to infer the contact residues structurally. They showed that PDZ binding predictions improved, specially for low sequence similarity domains [39].

In the last DREAM4 competition, Zaslavsky et al. were able to predict the most accurate PWMs for five PDZ domains with unknown specificity profiles [40]. They combined linear regression-based prediction for ligand positions whose specificity is known to be determined by relatively few PDZ domain positions, and single-mutant PWM averaging for all other ligand columns. However, more sophisticated machine learning-based specificity predictors are still needed that allow for modeling pair-wise or even higher order positional dependencies for the ligand and the domain. To increase the accuracy of these predictive models, a set of non-binding peptides may be required in the process of model generation. The necessity for a good negative training set has been shown to be especially important for the prediction of relative binding affinities. A recent study of PDZ domains has shown that a set of known non-binding peptides is required to train the models since it improves the prediction of relative binding affinities [41]. However, it is difficult to accurately identify non-binders by using large-scale techniques, and more stringent and tedious experimental techniques, such as fluorescence polarization affinity assays, are required [20].

Machine learning methods have the advantage of being fast and sometimes extremely accurate; however, they typically require large amounts of experimental training data, and thus may fail for systems that have not been well characterized experimentally. By contrast, physical/structural methods rely on basic principles of chemistry and physics to predict the relative binding affinities of different peptide ligands from the precise three-dimensional structure of the protein–peptide complexes. Prediction of binding affinity is often based on ab initio free energy calculations as per classical molecular mechanics or semi-empirical force fields [42,43]. These ab initio methods can be accurate even in the absence of experimental binding data or when the binding mode is unknown. However, they require large computational resources,

rendering the exploration of a large number of peptide sequences a challenge.

Many methods for peptide specificity prediction fall somewhere between these two approaches. For instance, Hou et al. used a combination of homology modeling, molecular dynamics, binding energies based on semi-empirical force fields, and machine learning to characterize interactions and to predict first the substrates for SH3 [42], and later for PDZ domains [43]. They trained a support-vector machine on such molecular interaction energy components (MIEC) to effectively predict binding partners of SH3 domains that bind to Class I peptides and 11 PDZ domains that bind to C-terminal peptides. In another approach, Serrano and coworkers combined structure homology modeling of SH3–peptide complexes and in silico mutagenesis scanning to construct PWMs that were used to predict the binding specificities of *S. cerevisiae* SH3 domains [44]. They were able to achieve high accuracy in predicting the binding specificities by using the FoldX empirical forcefield [45]. A similar methodology has been recently applied to PDZ domains using the Rosetta energy function to score interactions in order to predict changes in binding specificities in a set of point mutants [46]. They were able to predict binding preferences for a large set of natural PDZ domains as well as single and multiple Erbin–PDZ domains obtained from phage display experiments. Interestingly, they demonstrated that incorporation of backbone flexibility in the Rosetta module increased prediction accuracy. Although far from perfect, the results from both methods show great promise for the potential of structure-based approaches for predicting binding specificity.

4. Identification of motifs in the proteome: models and expressions

Complex signalling networks are normally mediated by the interactions of modular domains with short linear sequence motifs. The detection of these motifs is of crucial importance for enhancing our understanding of the molecular and cellular function of these proteins. These regions can be predicted by scanning the specificity profiles of peptide-binding domains across proteomes in order to identify putative natural binders. However, in most cases, the motifs are described by regular expressions that define important residues based on a combination of experimental, structural and evolutionary evidences (Fig. 2). Unlike consensus models, regular expressions are less accurate for describing a linear motif. Nevertheless, consensus models and expressions alone are often insufficient for ligand prediction because they tend to be short and degenerate, and matches are expected to occur frequently by chance in random sequences. Thus, it is required to narrow down the proteome search space in order to reduce the number of false positive hits by using local biological information. General discriminatory features are accessibility, disorder and conservation. Motif search is normally restricted in globular domains and enhanced in intrinsically disordered regions with a lack of tertiary structure. In addition, evolutionary conservation of a motif correlates strongly with functionality and many experimental motifs are seen as islands of strong constraint in regions of weak conservation.

Computational methods can identify putative domain binding partners by scanning the motif along the protein sequence in search of optimal hits. Scansite, developed by Yaffe and co-workers, was the first algorithm that predicts binding sites within protein sequences. Their approach used a sliding window method based on a normalized position weight matrix (PWM) that evaluated the residue conservation at each position [47,48]. More recently, the MOTIPS algorithm has incorporated conservation and structural features in conjunction with domain specificity information in a Bayesian framework to predict binding partners for SH3 domains [49]. The MOTIPS algorithm improves the domain binding predictions compared to using only profile-matching scan.

Linear motifs can also be inferred by mining the proteome alone. In this case, they are recognized as motifs, but cannot be immediately linked to a binding domain, which is assumed to exist. The SliM (short linear motif) approach searches for evolutionarily conserved motifs in biologically related proteins [50]. Unstructured protein regions have also been prioritized to improve prediction performance, since it has been observed that sequential motifs are more enriched in unfolded regions than in globular domains [51,52]. As a related concept, a database of molecular recognition features (MoRFs) has been created [53]. Such MoRFs are regions in disordered stretches that undergo disorder-to-order transition upon binding and largely coincide with linear motifs. In addition, evolutionary and disordered features have also been combined to distinguish functional binding sites by measuring the conservation of the motifs [54] and their flanking regions [55]. A related approach searches for local "islands" of evolutionary conservation in stretches of fast evolving disordered regions [56]. The largest collection of manually curated linear motifs in eukaryotic proteins is eukaryotic linear motif (ELM) that uses patterns

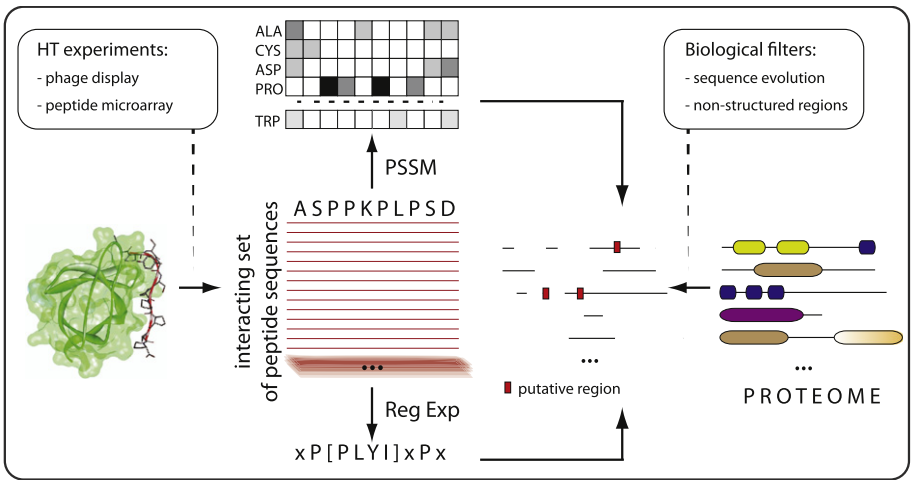

Fig. 2. Work-flow of the motif scanning of a proteome. High-throughput technologies provide a long list of sequences that can recognize a specific domain. After aligning the sequences, a proteome can be scanned either by a simple regular expression or by position weight matrix (PWM). The proteome to be scanned can be previously filtered using relevant biological features.

with context-based rules and logical filters [57]. The ELM instances, together with structural, biophysical, and biochemical features derived from the protein primary sequence are used by SLiMPred (short linear motif predictor) to predict new motifs in the proteome, using a hybrid approach based on machine learning techniques [58]. Recently, Moses and coworkers developed a new comparative genomic approach to identify short linear motifs within structurally disordered regions. Based on a phylogenetic hidden Markov model, they have been able to detect evolutionarily conserved regions that matched to known motifs and discovered other new ones, some of which were validated experimentally [59].

5. Network perturbations by synthetic design or targeted peptides

Nature has exploited loops and binding sites of protein interaction domains to evolve a wide spectrum of specificities. Although an enormous repertoire of specificities have been created by the combinatorial evolution of these regions, nature might only have sampled a fraction of this potential domain specificity space [27]. Therefore, it is possible to obtain a wide range of synthetic domain and peptide variants with desired specificities and affinities for biochemical, cellular, prognostic, diagnostic, or therapeutic applications. An interesting application of synthetic engineering is to rewire protein networks to experimentally change the links and parameters of the network with the final aim of identifying properties that are crucial for function. In this regard, peptide-binding domains are a useful system to perturb protein–protein interaction networks by synthetic design of either the domain or peptide site, as explained in this section.

Several studies have focused on investigating how the diverse specificities exhibited by different members of the PDZ and SH3 families are encoded in a common scaffold [60]. Experiments using phage-displayed libraries of SH3 domains have demonstrated that the binding properties of an SH3 domain could be profoundly changed by modifications in the non-conserved sequence in the RT loop [61]. The human Erbin PDZ domain was also subjected to combinatorial mutagenesis within 10 core positions that make contact with the peptide [62]. Screening of phages that displayed two of such combinatorial libraries yielded 288 structurally stable Erbin PDZ variants. Subsequent screening of phage-displayed peptide libraries using 237 purified Erbin PDZ variants revealed that many of these PDZ variants recognize C-terminal peptides and are as specific as natural domains. Thus the family of synthetic Erbin PDZ variants is as diverse as the natural PDZ family [62]. The diversity of residue types that could be accommodated at each of these 10 positions suggests that the PDZ fold is extremely robust and that co-evolution of ligand-binding residues might provide a rapid and effective means to generate the diverse specificities of natural PDZ domains. Loop engineering presents an attractive approach for creating domain variants with tailor-made specificities for research or therapeutic applications. For instance, loop randomization studies have been carried out on the RT and n-Src loops of several SH3 domains, and modifications of these loops have altered ligand preferences of the ABL1 and hemopoietic cell kinase (HCK) SH3 domains [63,61]. Moreover, loop randomization of the Fyn SH3 domain has yielded a high-affinity fibronectin-binding protein that could be used as an in vivo marker of angiogenesis [64].

Several groups exploited the possibility to engineer domain–peptide pairs that are simultaneously optimized to interact with their correct partner while avoiding cross-interaction with other members of the family [65,66]. In this regard, Lim and coworkers recombined the output domain of an N-WASP mammalian protein with an auto-inhibitory PDZ–peptide motif [67]. In this way, the N-WASP protein became a switch that could be activated by an external signal sequestering the PDZ domain from the peptide. The addition of another auto-inhibitory domain (SH3–peptide) "evolved" the synthetic switch into an AND gate. In more recent work, they also showed that by placing multiple copies of the only SH3–peptide motif on the N-WASP output domain, the activation response became much faster and the switch showed ultrasensitivity [68]. Finally, they also applied the same construction principle to a GEF protein and rendered it activatable [69]. To this aim, they used a modified PDZ–peptide motif, where the peptide had been mutated in order to be activated by a protein kinase. Hence, they were able to use this protein modularity to engineer an artificial signaling cascade that coupled the filopodial (Cdc42) and lamellipodial (Rac1) regulating GTPases in series.

In an increasing number of cases, proteins that contain peptide-binding domains have been found to be direct targets for regulation. In these cases, pathways can be turned on or off by inputs that modify the modular protein rather than the actual signaling enzymes [70–72]. These domains could be targeted by engineering peptide binders specifically optimized to interact with their correct partner while avoiding cross-interaction with other members of the family. For instance, Sidhu and coworkers designed high affinity and specificity peptide inhibitors against the Dishevelled-PDZ domain, which belongs to an unusual recognition class within the PDZ family. By targeting this domain, they were able to block Wnt/beta-catenin signaling in cells [73]. To increase binding affinity and specificity, Imperiali and coworkers chemically designed a bivalent peptide made of two natural C-terminal regions that was able to bind simultaneously to the two neighboring PDZ domains of PSD-95 protein. They used this biomimetic peptide to disrupt PSD-95 native interaction to AMPAR–Stargazin complex mediated by multiple class I PDZ domain-binding motifs, and studied the perturbation effects to excitatory synaptic transmission in the mammalian central nervous system [74]. Stromgaard and coworker have recently been able to develop an improved bivalent peptide inhibitor of PSD-95 PDZ1–2 that increased affinity from 25 to 400-fold relative to the monomeric ligand [75].

These examples show the potential of peptides as inhibitors of protein interactions. In fact, peptides have a long history as therapeutic molecules, and there are currently a total of over 60 peptides as approved drugs on the market [76]. Therapeutic peptides have seen resurgence in interest, partly because of the under-utilization of potential cytosolic targets [77,78]. Novel drugs developed in the past decade have almost solely focused on G protein-couple receptors (GPCRs) and protein kinases, presumably because of the high cost and risk associated with novel drug development [79].

6. Future perspectives

Large-scale analyses using phage-displayed random peptide libraries and other techniques have established specificity maps and comprehensive classification system for modular domain families, expanding significantly the canonical PDZ and SH3 domain classification. This massive information has permitted a significant improvement of computational methods for the prediction of novel peptide-binding domains for which no binding data is known, and for ligand prediction through an accurate scanning of the binding motifs in the proteomes.

Most of what is currently known about protein–peptide interactions is compiled in the ELM database. Currently, there are annotated 179 classes of literature-curated motifs that are interacting to 87 distinct Pfam domains. However, analysis of the PDB repository has shown that the peptide-binding domain families are much more abundant [80,81]. For example, PepX database currently

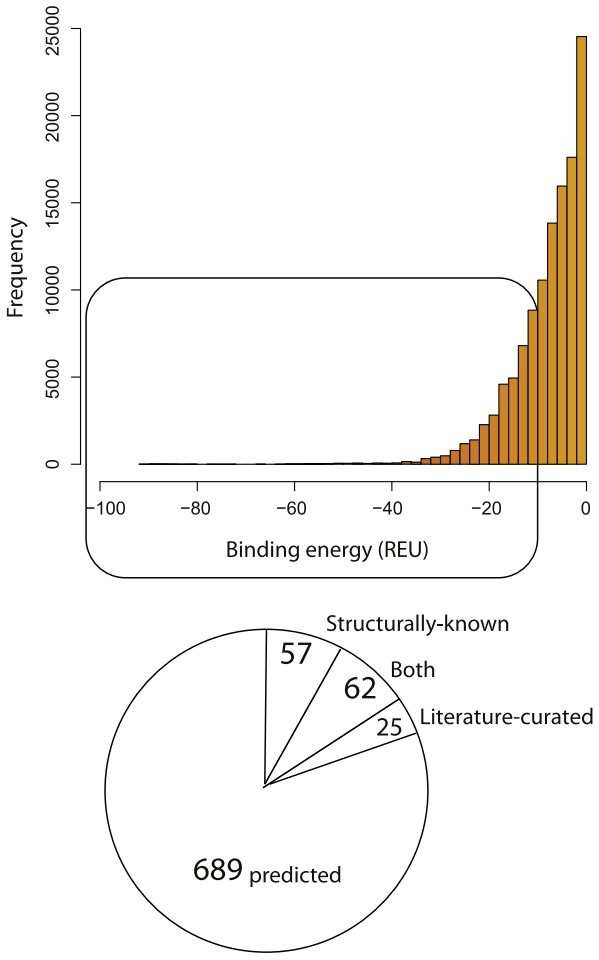

Fig. 3. Distribution of the number of putative peptide-binding domains based on predicted binding energies. This information has been computed for all the PDB using the PeptideDeriver program [82]. A pie graph shows the number of families distributed between natural, structurally-known and predicted peptide-binding domain families applying a −10 REU binding energy cutoff.

contains 119 Pfam domains structurally solved in complex to peptides, suggesting that the biologically relevant peptide-binding domains might be larger. Interestingly, a recent computational study suggest that a significant percentage of the structurally known globular protein–protein interactions are dominated by one short interfacial linear segment that can be used to derive self-inhibitory peptides [82]. These results imply that the number of peptide-binding domains could be much higher. For this reason, we have run their Rosetta-based program for all the protein–protein interactions in the PDB repository in order to estimate the number of putative peptide-binding domain families. Setting up the domain cut length to 16 amino acids, and applying a −10 REU binding energy cutoff, we identified putative peptide-binding domains for 689 new families, which represent almost 80% peptide-binding domain increase relative to the known families (Fig. 3). Although these numbers await experimental validation, previous success stories show that this strategy works for the design of peptide inhibitors [83–85]. In fact, the decline in productivity of drug discovery in the last years has produced an increased interest in pharmacological peptide-based inhibitors [77,78].

The current high-throughput technologies have evolved to a point that the generation of specificity profiles for all or most peptide-binding domains within several proteomes is conceivable. It is very likely that the concept ideas used for SH3 and PDZ domain families will be extended to study the other known peptide-binding modules [86]. In fact, several efforts have already been made in order to explore the specificities of WW domains (Sidhu et al., unpublished results). Such information may provide insights into the specificities of each module, and into the general principles governing protein–peptide recognition. It may also prove instrumental in deciphering the complex processes mediated by the numerous protein–protein interactions within the different proteomes.

References

[1] Pawson, T. and Scott, J.D. (1997) Signaling through scaffold, anchoring, and adaptor proteins. Science 278, 2075–2080.
[2] Sudol, M. (1998) From Src homology domains to other signaling modules: proposal of the "protein recognition code". Oncogene 17, 1469–1474.
[3] Aasland, R. et al. (2002) Normalization of nomenclature for peptide motifs as ligands of modular protein domains. FEBS Lett. 513, 141–144.
[4] Harris, B.Z. and Lim, W.A. (2001) Mechanism and role of PDZ domains in signaling complex assembly. J. Cell Sci. 114, 3219–3231.
[5] Songyang, Z. et al. (1997) Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. Science 275, 73–77.
[6] Nourry, C., Grant, S.G. and Borg, J.P. (2003) PDZ domain proteins: plug and play! Sci. STKE, RE7.
[7] Stricker, N.L. et al. (1997) PDZ domain of neuronal nitric oxide synthase recognizes novel C-terminal peptide sequences. Nat. Biotechnol. 15, 336–342.
[8] Zhang, Y. et al. (2006) Convergent and divergent ligand specificity among PDZ domains of the LAP and zonula occludens (ZO) families. J. Biol. Chem. 281, 22299–22311.
[9] Sparks, A.B., Quilliam, L.A., Thorn, J.M., Der, C.J. and Kay, B.K. (1994) Identification and characterization of Src SH3 ligands from phage-displayed random peptide libraries. J. Biol. Chem. 269, 23853–23856.
[10] Simon, J.A. and Schreiber, S.L. (1995) Grb2 SH3 binding to peptides from Sos: evaluation of a general model for SH3–ligand interactions. Chem. Biol. 2, 53–60.
[11] Mayer, B.J. (2001) SH3 domains: complexity in moderation. J. Cell Sci. 114, 1253–1263.
[12] Zarrinpar, A., Bhattacharyya, R.P. and Lim, W.A. (2003) The structure and function of proline recognition domains. Sci. STKE 2003, re8.
[13] Pires, J.R., Hong, X., Brockmann, C., Volkmer-Engert, R., Schneider-Mergener, J., Oschkinat, H. and Erdmann, R. (2003) The ScPex13p SH3 domain exposes two distinct binding sites for Pex5p and Pex14p. J. Mol. Biol. 326, 1427–1435.
[14] Jia, C.Y., Nie, J., Wu, C., Li, C. and Li, S.S. (2005) Novel Src homology 3 domain-binding motifs identified from proteomic screen of a pro-rich region. Mol. Cell. Proteomics 4, 1155–1166.
[15] Tian, L., Chen, L., McClafferty, H., Sailer, C.A., Ruth, P., Knaus, H.G. and Shipston, M.J. (2006) A noncanonical SH3 domain binding motif links BK channels to the actin cytoskeleton via the SH3 adapter cortactin. FASEB J. 20, 2588–2590.
[16] Kim, J., Lee, C.D., Rath, A. and Davidson, A.R. (2008) Recognition of non-canonical peptides by the yeast Fus1p SH3 domain: elucidation of a common mechanism for diverse SH3 domain specificities. J. Mol. Biol. 377, 889–901.
[17] Letunic, I., Doerks, T. and Bork, P. (2011) SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res. 40, D302–D305.
[18] Stoll, D., Templin, M.F., Bachmann, J. and Joos, T.O. (2005) Protein microarrays: applications and future challenges. Curr. Opin. Drug Discov. Devel. 8, 239–252.
[19] Fam, S.R. et al. (2005) P2Y1 receptor signaling is controlled by interaction with the PDZ scaffold NHERF-2. Proc. Natl. Acad. Sci. U S A 102, 8042–8047.
[20] Stiffler, M.A., Chen, J.R., Grantcharova, V.P., Lei, Y., Fuchs, D., Allen, J.E., Zaslavskaia, L.A. and MacBeath, G. (2007) PDZ domain binding selectivity is optimized across the mouse proteome. Science 317, 364–369.
[21] Hilpert, K., Winkler, D.F. and Hancock, R.E. (2007) Peptide arrays on cellulose support: SPOT synthesis, a time and cost efficient method for synthesis of large numbers of peptides in a parallel and addressable fashion. Nat. Protocols 2, 1333–1349.
[22] Boisguerin, P., Leben, R., Ay, B., Radziwill, G., Moelling, K., Dong, L. and Volkmer-Engert, R. (2004) An improved method for the synthesis of cellulose membrane-bound peptides with free C termini is useful for PDZ domain binding studies. Chem. Biol. 11, 449–459.
[23] Wiedemann, U., Boisguerin, P., Leben, R., Leitner, D., Krause, G., Moelling, K., Volkmer-Engert, R. and Oschkinat, H. (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. J. Mol. Biol. 343, 703–718.
[24] Landgraf, C., Panni, S., Montecchi-Palazzi, L., Castagnoli, L., Schneider-Mergener, J., Volkmer-Engert, R. and Cesareni, G. (2004) Protein interaction networks by proteome peptide scanning. PLoS Biol. 2, e14.
[25] Carducci, M. et al. (2012) The protein interaction network mediated by human SH3 domains. Biotechnol. Adv. 30, 4–15.
[26] Tong, A.H. et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295, 321–324.
[27] Tonikian, R. et al. (2008) A specificity map for the PDZ domain family. PLoS Biol 6, e239.

[28] Tonikian, R., Zhang, Y., Boone, C. and Sidhu, S.S. (2007) Identifying specificity profiles for peptide recognition modules from phage-displayed peptide libraries. Nat. Protocols 2, 1368–1386.
[29] Stein, A. and Aloy, P. (2008) Contextual specificity in peptide-mediated protein interactions. PLoS ONE 3, e2524.
[30] Gfeller, D. et al. (2011) The multiple-specificity landscape of modular peptide recognition domains. Mol. Syst. Biol. 7, 484.
[31] Ernst, A., Gfeller, D., Kan, Z., Seshagiri, S., Kim, P.M., Bader, G.D. and Sidhu, S.S. (2010) Coevolution of PDZ domain–ligand interactions analyzed by high-throughput phage display and deep sequencing. Mol. BioSyst. 6, 1782–1790.
[32] Fowler, D.M., Araya, C.L., Fleishman, S.J., Kellogg, E.H., Stephany, J.J., Baker, D. and Fields, S. (2010) High-resolution mapping of protein sequence–function relationships. Nat. Meth. 7, 741–746.
[33] Tonikian, R. et al. (2009) Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. PLoS Biol. 7, e1000218.
[34] Punta, M. et al. (2012) The Pfam protein families database. Nucleic Acids Res. 40, D290–D301.
[35] Liu, B.A., Jablonowski, K., Shah, E.E., Engelmann, B.W., Jones, R.B. and Nash, P.D. (2010) SH2 domains recognize contextual peptide sequence information to determine selectivity. Mol. Cell. Proteomics 9, 2391–2404.
[36] Stiffler, M.A., Grantcharova, V.P., Sevecka, M. and MacBeath, G. (2006) Uncovering quantitative protein interaction networks for mouse PDZ domains using protein microarrays. J. Am. Chem. Soc. 128, 5913–5922.
[37] Chen, J.R., Chang, B.H., Allen, J.E., Stiffler, M.A. and MacBeath, G. (2008) Predicting PDZ domain–peptide interactions from primary sequences. Nat. Biotech. 26, 1041–1045.
[38] Kalyoncu, S., Keskin, O. and Gursoy, A. (2010) Interaction prediction and classification of PDZ domains. BMC Bioinform. 11, 357.
[39] Hawkins, J.C., Zhu, H., Teyra, J. and Pisabarro, M.T. (2012) Reduced false positives in PDZ binding prediction using sequence and structural descriptors. IEEE/ACM Trans. Comput. Biol. Bioinform., in press.
[40] Zaslavsky, E., Bradley, P. and Yanover, C. (2010) Inferring PDZ domain multi-mutant binding preferences from single-mutant data. PLoS ONE 5, e12787.
[41] Shao, X., Tan, C.S.H., Voss, C., Li, S.S.C., Deng, N. and Bader, G.D. (2011) A regression framework incorporating quantitative and negative interaction data improves quantitative prediction of PDZ domain–peptide interaction from primary sequence. Bioinformatics 27, 383–390.
[42] Hou, T., Xu, Z., Zhang, W., McLaughlin, W.A., Case, D.A., Xu, Y. and Wang, W. (2009) Characterization of domain–peptide interaction interface. Mol. Cell. Proteomics 8, 639–649.
[43] Li, N., Hou, T., Ding, B. and Wang, W. (2011) Characterization of PDZ domain–peptide interaction interface based on energetic patterns. Proteins: Struct. Funct. Bioinform. 79, 3208–3220.
[44] Fernandez-Ballester, G., Beltrao, P., Gonzalez, J.M., Song, Y.H., Wilmanns, M., Valencia, A. and Serrano, L. (2009) Structure-based prediction of the *Saccharomyces cerevisiae* SH3–ligand interactions. J. Mol. Biol. 388, 902–916.
[45] Guerois, R., Nielsen, J.E. and Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J. Mol. Biol. 320, 369–387.
[46] Smith, C.A. and Kortemme, T. (2010) Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. J. Mol. Biol. 402, 460–474.
[47] Yaffe, M.B., Leparc, G.G., Lai, J., Obata, T., Volinia, S. and Cantley, L.C. (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways. Nat. Biotechnol. 19, 348–353.
[48] Obenauer, J.C., Cantley, L.C. and Yaffe, M.B. (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res. 31, 3635–3641.
[49] Lam, H.Y., Kim, P.M., Mok, J., Tonikian, R., Sidhu, S.S., Turk, B.E., Snyder, M. and Gerstein, M.B. (2010) MOTIPS: automated motif analysis for predicting targets of modular protein domains. BMC Bioinform. 11, 243.
[50] Davey, N.E., Shields, D.C. and Edwards, R.J. (2006) SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. Nucleic Acids Res. 34, 3546–3554.
[51] Neduva, V. et al. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol. 3, e405.
[52] Mészáros, B., Simon, I. and Dosztányi, Z. (2009) Prediction of protein binding regions in disordered proteins. PLoS Comput. Biol. 5, e1000376.
[53] Mohan, A., Oldfield, C.J., Radivojac, P., Vacic, V., Cortese, M.S., Dunker, A.K. and Uversky, V.N. (2006) Analysis of molecular recognition features (MoRFs). J. Mol. Biol. 362, 1043–1059.
[54] Mok, J. et al. (2010) Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. Sci. Signal. 3, ra12.
[55] Chica, C., Diella, F. and Gibson, T.J. (2009) Evidence for the concerted evolution between short linear protein motifs and their flanking regions. PLoS ONE 4, e6052.
[56] Lai, A.C.W., Nguyen Ba, A.N. and Moses, A.M. (2012) Predicting Kinase Substrates using conservation of local motif density. Bioinformatics.
[57] Gould, C.M. et al. (2010) ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic Acids Res. 38, D167–180.
[58] Mooney, C., Pollastri, G., Shields, D.C. and Haslam, N.J. (2012) Prediction of short linear protein binding regions. J. Mol. Biol. 415, 193–204.
[59] Nguyen Ba, A.N., Yeh, B.J., van Dyk, D., Davidson, A.R., Andrews, B.J., Weiss, E.L. and Moses, A.M. (2012) Proteome-wide discovery of evolutionary conserved sequences in disordered regions. Sci. Signal. 5, rs1.
[60] Kaneko, T., Sidhu, S. and Li, S. (2011) Evolving specificity from variability for protein interaction domains. Trends Biochem. Sci., 183–190.
[61] Hiipakka, M. and Saksela, K. (2007) Versatile retargeting of SH3 domain binding by modification of non-conserved loop residues. FEBS Lett. 581, 1735–1741.
[62] Ernst, A., Sazinsky, S.L., Hui, S., Currell, B., Dharsee, M., Seshagiri, S., Bader, G.D. and Sidhu, S.S. (2009) Rapid evolution of functional complexity in a domain family. Sci. Signal. 2, ra50.
[63] Panni, S., Dente, L. and Cesareni, G. (2002) In vitro evolution of recognition specificity mediated by SH3 domains reveals target recognition rules. J. Biol. Chem. 277, 21666–21674.
[64] Grabulovski, D., Kaspar, M. and Neri, D. (2007) A novel, non-immunogenic Fyn SH3-derived binding protein with tumor vascular targeting properties. J. Biol. Chem. 282, 3196–3204.
[65] Bashor, C.J., Horwitz, A.A., Peisajovich, S.G. and Lim, W.A. (2010) Rewiring cells: synthetic biology as a tool to interrogate the organizational principles of living systems. Ann. Rev. Biophys. 39, 515–537.
[66] Grünberg, R., Ferrar, T.S., Van Der Sloot, A.M., Constante, M. and Serrano, L. (2010) Building blocks for protein interaction devices. Nucleic Acids Res. 38, 2645–2662.
[67] Dueber, J.E., Yeh, B.J., Chak, K. and Lim, W.A. (2003) Reprogramming control of an allosteric signaling switch through modular recombination. Science 301, 1904–1908.
[68] Dueber, J.E., Mirsky, E.A. and Lim, W.A. (2007) Engineering synthetic signaling proteins with ultrasensitive input/output control. Nat. Biotechnol. 25, 660–662.
[69] Yeh, B.J., Rutigliano, R.J., Deb, A., Bar-Sagi, D. and Lim, W.A. (2007) Rewiring cellular morphology pathways with synthetic guanine nucleotide exchange factors. Nature 447, 596–600.
[70] Strickfaden, S.C., Winters, M.J., Ben-Ari, G., Lamson, R.E., Tyers, M. and Pryciak, P.M. (2007) A mechanism for cell-cycle regulation of MAP kinase signaling in a yeast differentiation pathway. Cell 128, 519–531.
[71] McKay, M.M., Ritt, D.A. and Morrison, D.K. (2009) Signaling dynamics of the KSR1 scaffold complex. Proc. Natl. Acad. Sci. U S A 106, 11022–11027.
[72] Won, A.P., Garbarino, J.E. and Lim, W.A. (2011) Recruitment interactions can override catalytic interactions in determining the functional identity of a protein kinase. Proc. Natl. Acad. Sci. U S A 108, 9809–9814.
[73] Zhang, Y., Appleton, B.A., Wiesmann, C., Lau, T., Costa, M., Hannoush, R.N. and Sidhu, S.S. (2009) Inhibition of Wnt signaling by Dishevelled PDZ peptides. Nat. Chem. Biol. 5, 217–219.
[74] Sainlos, M. et al. (2010) Biomimetic divalent ligands for the acute disruption of synaptic AMPAR stabilization. Nat. Chem. Biol. 7, 81–91.
[75] Bach, A. et al. (2012) A high-affinity, dimeric inhibitor of PSD-95 bivalently interacts with PDZ1-2 and protects against ischemic brain damage. Proc. Natl. Acad. Sci. U S A. 109, 3317–3322.
[76] Vlieghe, P., Lisowski, V., Martinez, J. and Khrestchatisky, M. (2010) Synthetic therapeutic peptides: science and market. Drug Discovery Today 15, 40–56.
[77] Gaestel, M. and Kracht, M. (2009) Peptides as signaling inhibitors for mammalian MAP kinase cascades. Curr. Pharm. Des. 15, 2471–2480.
[78] Marchioni, F. and Zheng, Y. (2009) Targeting rho GTPases by peptidic structures. Curr. Pharm. Des. 15, 2481–2487.
[79] Overington, J.P., Al-Lazikani, B. and Hopkins, A.L. (2006) How many drug targets are there? Nat. Rev. Drug Discovery 5, 993–996.
[80] Vanhee, P., Reumers, J., Stricher, F., Baeten, L., Serrano, L., Schymkowitz, J. and Rousseau, F. (2010) PepX: a structural database of non-redundant protein–peptide complexes. Nucleic Acids Res. 38, D545–551.
[81] Teyra, J., Samsonov, S.A., Schreiber, S. and Pisabarro, M.T. (2011) SCOWLP update: 3D classification of protein–protein, -peptide, -saccharide and -nucleic acid interactions, and structure-based binding inferences across folds. BMC Bioinform. 12, 398.
[82] London, N., Raveh, B., Movshovitz-Attias, D. and Schueler-Furman, O. (2010) Can self-inhibitory peptides be derived from the interfaces of globular protein–protein interactions? Proteins 78, 3140–3149.
[83] Huang, Z. (2000) Bcl-2 family proteins as targets for anticancer drug design. Oncogene 19, 6627–6631.
[84] Walensky, L.D. et al. (2004) Activation of apoptosis in vivo by a hydrocarbon-stapled BH3 helix. Science 305, 1466–1470.
[85] LaCasse, E.C., Mahoney, D.J., Cheung, H.H., Plenchette, S., Baird, S. and Korneluk, R.G. (2008) IAP-targeted therapies for cancer. Oncogene 27, 6252–6275.
[86] Huang, H. and Sidhu, S.S. (2011) Studying binding specificities of peptide recognition modules by high-throughput phage display selections. Methods Mol. Biol. 781, 87–97.