Coevolution of PDZ domain–ligand interactions analyzed by high-throughput phage display and deep sequencing†

Andreas Ernst,[a] David Gfeller,[a] Zhengyan Kan,[b] Somasekar Seshagiri,[b] Philip M. Kim,[a] Gary D. Bader[a] and Sachdev S. Sidhu*[a]

Received 21st June 2010, Accepted 13th July 2010
DOI: 10.1039/c0mb00061b

The determinants of binding specificities of peptide recognition domains and their evolution remain important problems in molecular systems biology. Here, we present a new methodology to analyze the coevolution between a domain and its ligands by combining high-throughput phage display with deep sequencing. First, from a library of PDZ domains with diversity introduced at ten positions in the binding site, we evolved domains for binding to 15 distinct peptide ligands. Interestingly, for a given peptide many different functional domains emerged, which exhibited only limited sequence homology, showing that many different binding sites can recognize a given peptide. Subsequently, we used peptide-phage libraries and deep sequencing to map the specificity profiles of these evolved domains at high resolution, and we found that the domains recognize their cognate peptides with high affinity but low specificity. Our analysis reveals two aspects of evolution of new binding specificities. First, we were able to identify some common features amongst domains raised against a common peptide. Second, our analysis suggests that cooperative interactions between multiple binding site residues lead to a diversity of binding profiles with considerable plasticity. The details of intramolecular cooperativity remain to be elucidated, but nonetheless, we have established a general methodology that can be used to explore protein evolution in a systematic yet rapid manner.

Introduction

Protein–protein interactions are involved in essentially all cellular processes, and hence, there is intense interest in understanding how these interactions have evolved.[1] A better characterization of these evolutionary processes would enhance our fundamental understanding of cell function, and also, would aid the design of synthetic proteins with novel functions.[2,3] However, efforts to understand natural protein–protein interactions have been stymied by the complex nature of these systems, which often include dozens of residues on either side of the interaction interface and usually depend on intricate structural features of the interacting partners.[4,5]

In this regard, peptide recognition modules (PRMs) have served as useful model systems, because they represent some of the simplest types of protein–protein interactions, and yet, they are central to the regulation of numerous cellular functions.[1,6,7] PRMs recognize short linear stretches of primary sequence in other proteins and serve to assemble multi-protein complexes. Dozens of PRM families have been identified in the human genome and each family may contain up to hundreds of distinct members. Each family is defined by a common fold and a core recognition motif, but sequence differences amongst family members confer differences in specificity which in turn adapt each member to a distinct functional niche.

Amongst PRM families, the PDZ domain family is particularly noteworthy, because it is amongst the most prevalent PRM folds in metazoans (*e.g.* the human genome contains over 250 PDZ domains in over 100 proteins).[8] Furthermore, most PDZ domains recognize C-terminal sequences by using a common binding cleft that lies between a β-strand (β2) and an α-helix (α2) and is blocked at one end by a "carboxylate-binding" loop.[9] The peptide ligand inserts between strand β2 and helix α2 as an additional β-strand, and the C terminus interacts with the carboxylate-binding loop.[10] We and others have studied PDZ domain specificity using peptide libraries to gain insights into the functions of individual domains, and also, in the hopes that specificity analyses may reveal general rules about relationships between structure and function.[11–14] Eventually, if enough information is obtained about a large number of PDZ domains, it may be possible to predict specificity on the basis of primary sequence and to engineer novel specificities by computational design, and on a broader scale, knowledge gained from the PDZ domain family may help to elucidate the fundamental principles that govern protein–protein interactions.

With these goals in mind, two recent large-scale studies have tackled the issue of PDZ domain specificity on a family-wide scale. In one study, MacBeath and colleagues used synthetic peptides to map interactions between 85 mouse PDZ domains and several hundred natural ligands.[15] In our own study, we

[a] *Banting and Best Department of Medical Research, Department of Molecular Genetics, and the Terrence Donnelly Center for Cellular and Biomolecular Research, University of Toronto, 160 College Street, Toronto, Ontario, Canada M5S 3E1. E-mail: sachdev.sidhu@utoronto.ca*
[b] *Department of Molecular Biology, Genentech, 1 DNA Way, South San Francisco, CA 94080, USA*

† Electronic supplementary information (ESI) available: Supplementary Figures. See DOI: 10.1039/c0mb00061b

used peptide-phage libraries to map specificity profiles for 82 worm and human PDZ domains.[16] Each of these studies shed light on the functional diversity of the natural PDZ domain family and enabled the development of algorithms that can predict specificities for binding sites that are similar to those of mapped PDZ domains.[16,17] However, even these large data sets do not provide a fully predictive understanding of the relationships between PDZ domain sequence and specificity, and unfortunately, the findings do not provide general insights into protein–protein interactions.

Having conducted a large-scale analysis of PDZ domain specificity by phage display, we have arrived at the sobering conclusion that even exhaustive knowledge of natural PDZ domain specificity may not be sufficient to achieve a complete understanding of the relationships between PDZ structure and function. This is due to a fundamental limitation of natural protein sequence space. Even amongst large domain families with hundreds of members, the natural diversity represents only a sparse sampling of the sequence space that is compatible with the family fold. In the case of human PDZ domains, the average sequence identity is less than 30% and there are only a few positions conserved among all domains. Consequently, it has thus far proven impossible to accurately define the minimum changes necessary to interconvert between different natural PDZ domain specificities, aside from some simple changes that are mediated by single amino acid residues,[16] and furthermore, natural PDZ sequence space provides little insight into non-natural specificities that could be supported by the PDZ domain fold.

Confronted by the limitations of natural protein sequence space, we have turned to "synthetic" PDZ domains designed to specifically address key questions about the evolution of protein–protein interactions. Recently, we adapted the Erbin PDZ domain (Erbin-PDZ) as a model system to explore protein evolution. We displayed Erbin-PDZ on phage and constructed a large combinatorial library by randomizing 10 positions that we defined as the core peptide-binding site.[18] The library contained approximately 10^9 members and was subjected to a selection for protease resistance to enrich for structured domains. Remarkably, 61 out of 237 randomly chosen domains proved to be functional for C-terminal peptide recognition. Thus, the PDZ domain fold is "hardwired" for C-terminal peptide recognition, because one-quarter of our structured repertoire was functional, despite being heavily mutated and not being subjected to any selective pressure for function.

Detailed analysis of our family of 61 synthetic domains revealed at least 14 distinct specificity classes,[18] which was comparable to the 16 classes defined for the 82 natural domains mapped by phage display.[16] Approximately half of the specificity classes for the synthetic domains matched those of natural domains and the other half represented novel specificities not observed in nature. Thus, the functional diversity of the synthetic PDZ domain family, derived in weeks without any selection for function, was equal to that of the natural family, which has evolved for function over more than one billion years.[19] Importantly, the synthetic domain family was derived from a well-defined library restricted to mutations at only ten positions in the peptide-binding site, and thus, the synthetic system is as functionally complex as the natural system but structurally much simpler. Our synthetic approach established a general methodology that can be used to expand our database of synthetic PDZ domains and corresponding specificity profiles designed to address particular questions about protein structure, function and evolution.

Here, we present a further study designed to address how PDZ domains and peptide ligands co-evolve. We used high-throughput phage display methods to analyze numerous domains and peptides in a rapid manner, and in addition, we adapted deep sequencing methods to exhaustively explore the peptide ligand sequence space. First, we assembled a set of 15 peptides representing optimal ligands for members of our original synthetic PDZ domain family and selected for additional binding domains from our phage-displayed Erbin-PDZ library. Subsequently, we purified a panel of these newly selected synthetic domains and determined their specificity profiles using peptide-phage libraries. By exploring the evolution of both sides of the PDZ domain–ligand interface in a controlled, large-scale manner, we provide important insights into the process of domain–ligand co-evolution.

Results

Evolution of peptide-binding synthetic PDZ domains

Our panel of 61 previously characterized synthetic PDZ domains constitutes a family of "unevolved" reference domains, because they were not subjected to any selective pressure for function.[18] Nonetheless, each family member is capable of specific recognition of defined peptide sequences. Since this family was generated by sampling only a few hundred of the $\sim 10^9$ Erbin-PDZ library members, these results raise an interesting question: what kinds of domains would evolve if the entire library was subjected to selective pressure for binding to optimal ligands for the reference domains? Based on an inspection of the specificity profiles for the unevolved synthetic PDZ domain family, we designed a set of 15 peptides that included an optimal ligand for wild type Erbin-PDZ (P-WT) and 14 ligands for 12 synthetic PDZ domains (Fig. 1A). These peptides were chosen to represent a diverse sequence space so as to explore different specificities at the last four ligand positions.

We verified that each peptide bound to the PDZ domain for which it was designed (the reference PDZ domain), and thus confirmed that our Erbin-PDZ library contained at least one binding domain for each peptide (data not shown). Next, we cycled the Erbin-PDZ library (Fig. 1B) through rounds of independent binding selections against each of the 15 peptides. All of the selections were successful and sequencing of approximately 700 binding clones revealed 162 unique PDZ domains evolved for binding to the various peptide ligands. Peptides P-37a and P-26 yielded only two or one unique binding domains, respectively, but all other peptides yielded at least six unique binding domains (ESI†, Fig. S1). A comparison of domains selected for binding to different peptides showed that most domains were unique, but 17 domains were selected for binding to two or more peptides. These multi-specific domains likely arose due to cross-contamination

A

Peptide	Position -6	-5	-4	-3	-2	-1	0
P-WT	T	G	W	E	T	W	V
P-35	S	K	L	R	F	W	V
P-39	F	F	G	E	R	W	V
P-41	G	N	R	D	G	W	V
P-47	S	L	D	G	W	W	V
P-12	S	N	H	G	R	W	V
P-19	R	Y	G	G	F	R	V
P-19a	S	L	W	K	F	W	V
P-14	R	W	E	S	D	W	L
P-48	W	V	F	G	P	W	L
P-55	H	Y	R	F	W	D	V
P-37	P	Y	R	Y	F	D	V
P-37a	L	Y	R	H	W	R	V
P-26	F	W	D	G	R	R	V
P-24	W	F	T	S	G	W	M

B

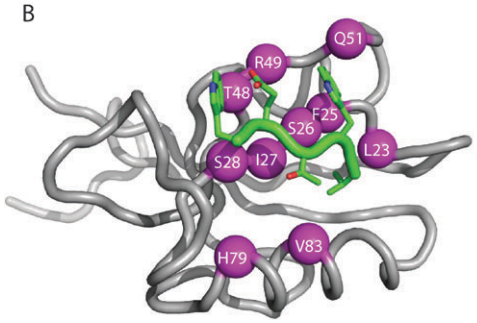

Fig. 1 The peptide ligands and the Erbin-PDZ library used for the selection of synthetic PDZ domains. (A) Each peptide was designed as an optimal ligand for a reference synthetic PDZ domain, based on specificity profiles from a previous study.[18] The numerical designation for each peptide corresponds to the numerical designation of its reference PDZ domain (*e.g.* peptide P-35 was designed to match the specificity profile of domain E-35). (B) Design of the Erbin-PDZ library.[18] Erbin-PDZ (grey) is shown with a bound peptide ligand ($WETWV_{COOH}$, green).[24] Binding site positions that were varied are depicted as magenta spheres; the wt sequences and position numbers are shown.

between selection pools for different peptides, but nonetheless, each domain was verified for binding to each peptide.

We compared the sequence identity between each reference domain and the evolved domains selected for binding to its optimal peptide ligand (ESI†, Fig. S2). At the ten binding site positions that were varied in the library, the mean identity between the reference and evolved domains is only 29%. Furthermore, a comparison between domains evolved for binding to the same peptide reveals only a slightly greater mean identity of 41%. Notably, a comparison amongst domains evolved for binding to different peptides reveals a mean identity of 35%, which is only slightly lower than the identity amongst domains evolved for binding to the same peptide. Thus, there is a remarkable lack of sequence conservation within the binding sites of domains selected for identical binding function, and even lower conservation when comparing evolved domains to their reference domains. Taken together, these results show that the Erbin-PDZ library contains multiple members that are capable of recognizing each peptide ligand by using binding sites that differ greatly from each other.

Specificity profiling of evolved synthetic PDZ domains by deep sequencing

Although the evolved PDZ domains bind to the peptide ligands they were selected against, it is not clear whether these peptides are optimal ligands for the domains. To address this question, we purified 22 PDZ domains that were evolved for binding to nine different peptides. Each of these domains was used as the target for binding selections with a phage-displayed library of random heptapeptides with free C termini. To obtain high resolution mapping of each specificity profile, we used deep sequencing methods and obtained a total of 44 097 sequence reads representing 26 566 unique peptides, which constitutes an order of magnitude more data than the largest previous phage display analyses.[16,18,20] We aligned the unique binding peptides for each domain using a new computational algorithm that is capable of detecting pairwise cooperativity amongst ligand positions (see Experimental).[21] This analysis revealed that most of the domains exhibit two or more specificity profiles, indicating that the binding sites are capable of accommodating several distinct types of peptides (Fig. 2).

We next compared the specificity of the evolved synthetic PDZ domains to those of the unevolved synthetic domains[18] and the natural domains[16] in our previous data sets, using the specificity potential (SP) metric, which is based on the position weight matrix (PWM), calculated from the entire set of binding peptides for each domain, and which varies from one (most specific) to zero (least specific).[16] As described previously,[18] the SP values of the unevolved synthetic domains are comparable to those of the natural domains across the last seven ligand positions, indicating that the unevolved synthetic domains are as specific as the natural domains. In contrast, the evolved synthetic domains exhibit comparable SP values for the last two ligand positions but are much less specific than the natural and unevolved synthetic domains for ligand positions further upstream (Fig. 3A). Consequently, the total specificity potential (the sum of SP values across all ligand positions) of the evolved synthetic domains is much lower than that of the unevolved synthetic domains and is comparable to that of the least specific natural domains (Fig. 3B). Previously, we also showed that the unevolved synthetic domains typically recognize optimal ligands with lower affinities than the natural PDZ domains.[18] Affinity measurements for three evolved synthetic PDZ domains show that they recognize their optimal ligands with affinities that are much greater than those of the unevolved synthetic domains and are comparable to those of natural domains.

In sum, the evolved synthetic domains recognize peptides with interactions that are of higher affinity but lower specificity than those of the unevolved synthetic domains, and these traits resemble those of the least specific natural domains. Consequently, most of the specificity profiles of the evolved synthetic domains appear to be less specific versions of the specificity profiles of the corresponding reference domains (Fig. 2). Thus, each evolved domain is able to bind to the

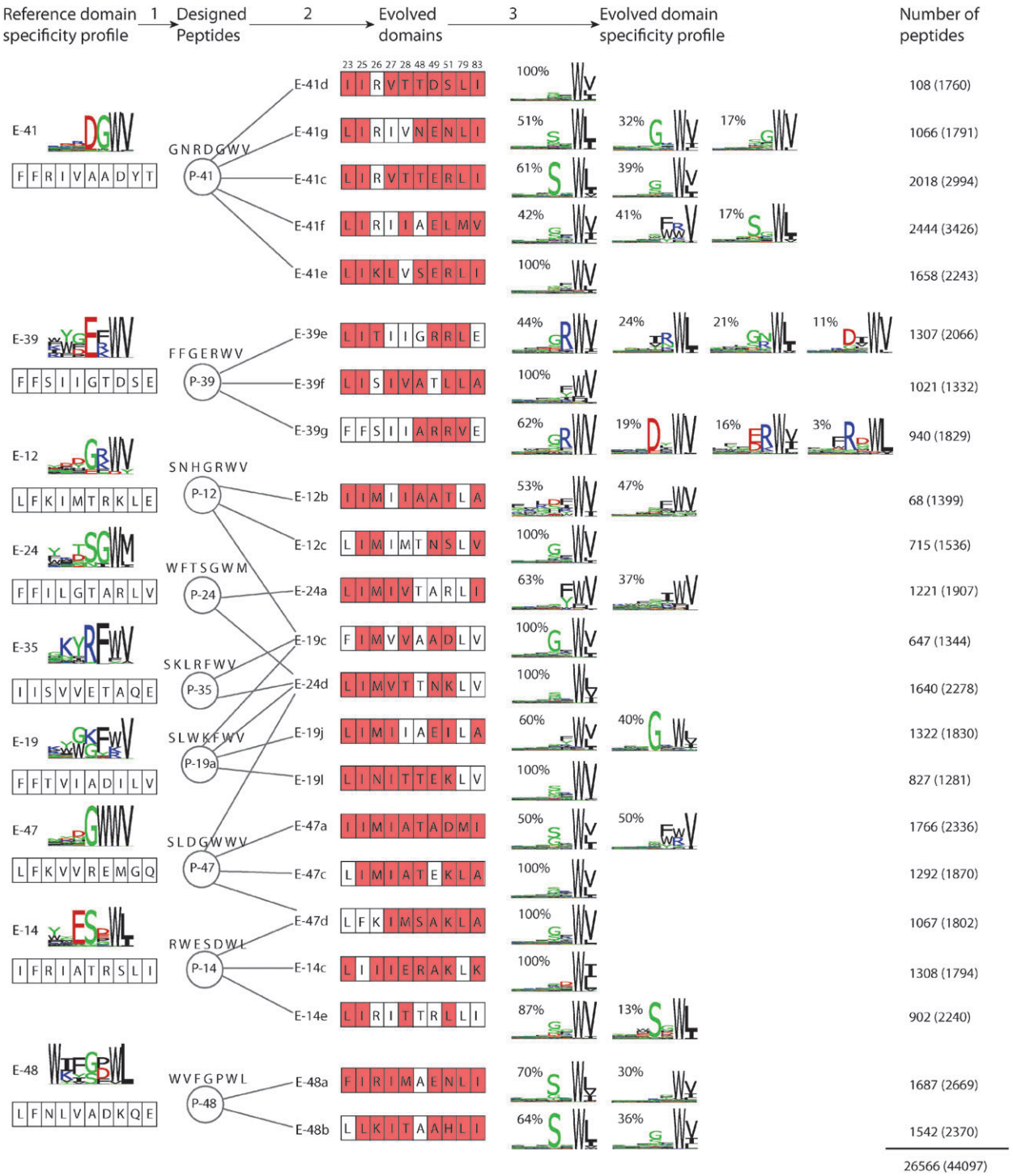

Fig. 2 Specificity profiles for evolved synthetic domains selected for binding to peptide ligands. The aligned peptide ligand set for each domain was used to create a PWM, and the specificity of each ligand position was visualized as a sequence logo. In step 1, the specificity profile of a previously analyzed unevolved reference domain[18] was used to design an optimal peptide ligand. The sequence logo for the specificity of each reference domain is shown, and below is shown the sequence of the reference domain binding site at positions that were varied in the Erbin-PDZ library. The name of each domain is shown to the left of the sequence logo. In step 2, each peptide was used to evolve binding domains from the Erbin-PDZ library. The sequence of the binding site of each evolved domain is shown and residues that differ from the corresponding reference domain are coloured red. The name of each evolved domain is shown to the left of the binding site sequence. In step 3, the evolved domains were used as targets in binding selections with a random peptide-phage library to select binding peptides. DNA from the binding pools was subjected to deep sequencing and the binding peptide sequences were analyzed to define specificity profiles. Owing to the extremely large number of unique peptide sequences that were retrieved, multiple distinct specificities could be detected for many domains. The numbers of unique peptides used to derive the sequence logos are shown to the right and the total numbers of sequence reads are shown in parentheses. The number above each sequence logo is the percentage of unique peptides used to derive the logo.

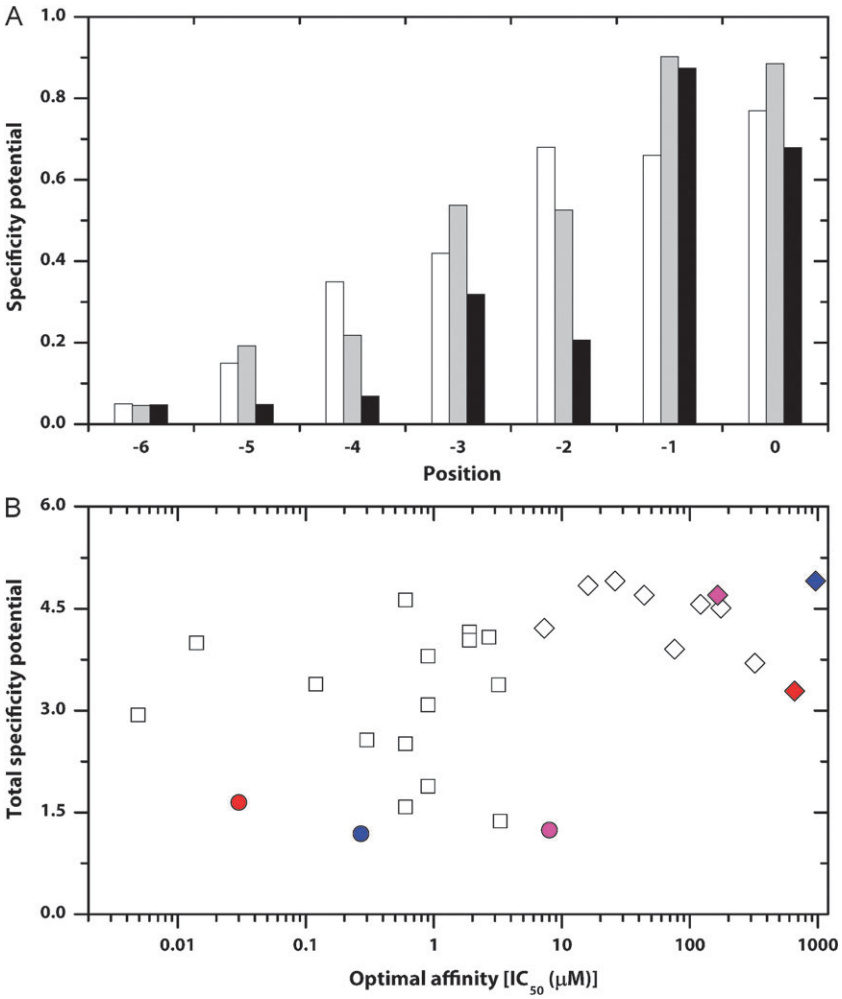

Fig. 3 Specificity and affinity of synthetic and natural PDZ domain–ligand interactions. (A) The mean specificity potential (SP) value at each ligand position is shown for 73 natural domains (white bars), 51 unevolved synthetic domains (grey bars), and 22 evolved synthetic domains (black bars). (B) The plot shows the total specificity potential summed over all ligand positions (*y* axis) and the affinities for optimal ligands (*x* axis) for 11 unevolved synthetic domains (diamonds), 15 natural domains (squares) and three evolved synthetic domains (circles). The filled symbols that are coloured the same indicate pairs of unevolved and evolved domains that were assayed for binding to the same peptide, as follows: magenta (peptide P-37, domains E-37 and E-37a), red (peptide P-12, domains E-12 and E-12a), blue (peptide P-19, domains E-19 and E-19j). Data for the natural and unevolved synthetic domains were reported previously.[18]

peptide ligand against which it was selected, but in comparison with the reference domain, it does so by recognizing fewer features of the ligand, and in extreme cases, many evolved domains recognize only the last two residues of the ligand. It appears that selection for binding to peptide ligands has led to the evolution of PDZ domains that can achieve high affinity through optimized recognition of only a few ligand side chains. Thus, somewhat paradoxically, the selection pressure has produced domains that are less specific than the unevolved synthetic domains, because unlike the unevolved domains, most of the evolved domains do not require interactions with many ligand side chains to achieve productive binding.

Correlation analysis of PDZ domain–ligand interactions

Overall, our results show that many different PDZ domains can recognize a given peptide, and binding sites that are selected for the same peptide are only slightly more similar to each other than they are to binding sites that are selected for different peptides (ESI†, Fig. S2). Consequently, even though we varied only ten positions in the PDZ domain, it is difficult to discern whether common changes in the binding site are responsible for improved recognition of particular peptides. Nonetheless, we wondered whether it might be possible to identify changes at specific PDZ domain positions that are correlated with changes at specific ligand positions. In particular, the 15 peptide ligands that were used in the selection experiments were designed such that at least some peptides contain positively-charged or negatively-charged side chains at the −1, −2 or −3 positions. Consequently, we analyzed our data set of evolved PDZ domains to identify sequence features that may be correlated with these features of the peptide ligands.

First, we aligned the sequences of the ten varied positions for all 162 unique PDZ domains that were selected against any of the 15 peptide ligands, and we represented this alignment as

a PWM (Fig. 4A). Next, we generated PWMs for the subsets of PDZ domains that were selected against peptide ligands that contain an R^{-1} (P-19, P-37a and P-26), a D^{-1} (P-55 and P-37), an R^{-2} (P-39, P-12 and P-26), a D^{-2} (P-14), an R/K^{-3} (P-35 and P-19a), or a D/E^{-3} (P-WT, P-39 and P-41). We then compared the PWM for each subset to the PWM for the entire set to identify differences that might represent sequence changes in the PDZ domain that favour the ligand feature of each subset.

We observed significant correlated changes at four positions in the PDZ domain binding site (Fig. 4A and B), and overall, the changes seem to alter the electrostatic potential of the PDZ domain binding site in a manner that would favour interactions with the different charged ligand features. In particular, there appears to be a depletion or enrichment of positively-charged residues at position 26 in response to positively-charged or negatively-charged residues, respectively, at any of the three ligand positions. This observation is consistent with general electrostatic effects whereby positive charge at position 26 would be expected to repel or attract positively-charged or negatively-charged ligands, respectively. At position 51, negatively-charged residues are enriched in response to ligands that contain an R^{-1}. These results are consistent with the structures of two natural PDZ domains, in each of which, a negatively-charged side chain at the corresponding position in the binding site makes favourable electrostatic interactions with an R^{-1} in a peptide ligand.[22,23] Similarly, at position 49, enrichment for positive or negative charge is correlated with ligands containing D^{-1} or R^{-3}, respectively. Position 49 is proximal to both the −1 and −3 positions of the ligand, and thus, these changes are consistent with the introduction of favourable electrostatic interactions between opposite charges in the PDZ domain and the ligand. Finally, positively-charged K residues at position 83 are enriched in response to ligands that contain D^{-2}, and this change is expected to establish favourable electrostatic interactions, because the ligand side chain at the −2 position projects towards the PDZ domain side chain at position 83.[10,24]

Discussion

Previously, we conducted a large-scale analysis of the natural PDZ domain family,[16] and subsequently, we developed a high-throughput phage display methodology that enables the specificity profiling of hundreds of domains in parallel.[18] Now, we have combined the high-throughput selection methodology with deep sequencing techniques[25] that enable exhaustive sampling of selection pools, and the combined approach enables the study of *in vitro* protein evolution with unprecedented speed and precision. In the current work, we have used this new approach to explore the coevolution of PDZ domain–ligand interactions.

By selecting a phage-displayed PDZ domain library for binding to a panel of 15 distinct peptide ligands, we evolved a large panel of functional synthetic domains. Comparison of domains raised against the same peptide revealed binding site identities that are only slightly greater than those amongst domains raised against different peptides. As observed previously for natural domains,[16] these results show that a

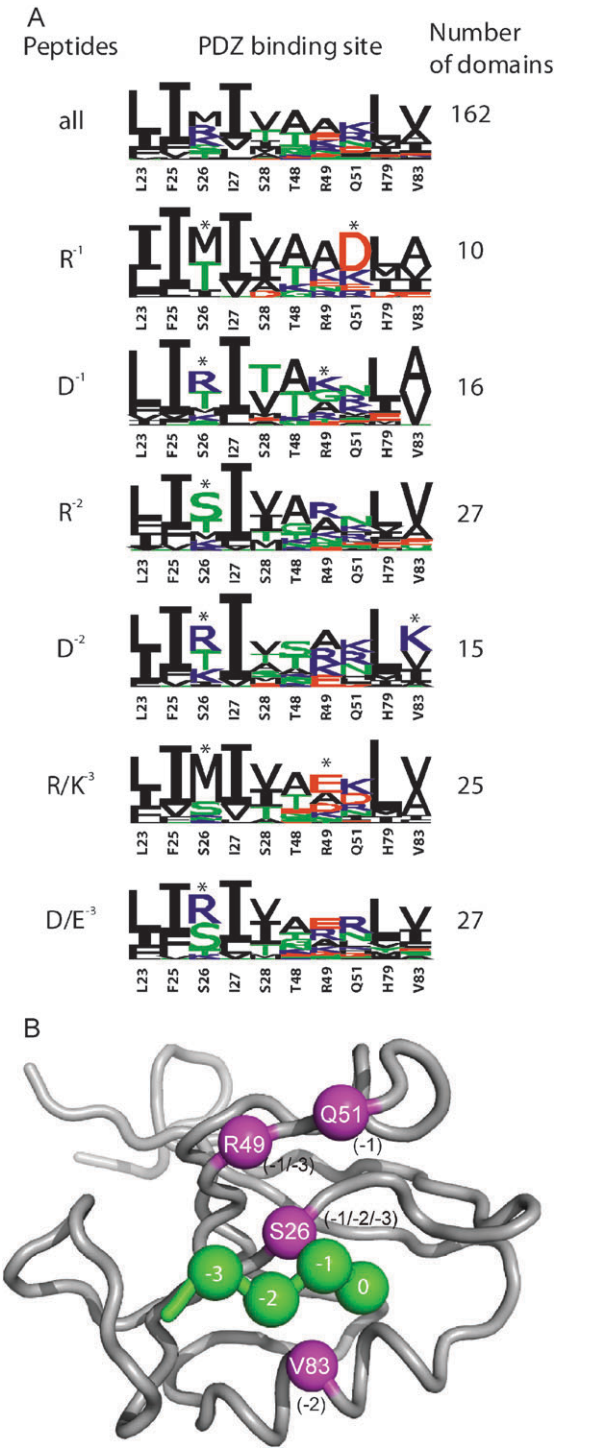

Fig. 4 Comparison of synthetic PDZ domain binding sites selected for binding to different types of peptides. (A) The aligned sequences at binding site positions that were varied in the Erbin-PDZ library were used to create a PWM, which was visualized as a sequence logo. At the left is indicated the nature of the peptides that were used to select the domains included in each PWM. At the right is indicated the number of unique domains included in each PWM. Asterisks (*) indicate binding site positions that differ in each sequence logo compared to the logo for the domains selected against all peptides. (B) Correlated positions in Erbin-PDZ and peptide ligands. Erbin-PDZ (grey) is shown with a bound peptide ligand (green). Magenta spheres represent Erbin-PDZ positions at which changes are correlated with changes in particular peptide positions (shown in parenthesis).

single peptide can be recognized by very different binding sites. This property has likely aided the rapid evolution of functional diversity in the PDZ domain family,[18] but it complicates our efforts to understand the molecular basis for PDZ domain function and evolution. As expected, the evolved synthetic domains recognize their cognate peptides with higher affinities than reference domains that were not evolved for function, but surprisingly, the evolved domains do so with lower specificities than the unevolved synthetic domains and most natural domains. This result likely originates from a key difference between our *in vitro* evolution process and natural selection, namely, the lack of competition for ligands. In our case, the system was driven purely by affinity, and thus, we obtained domains that bound to peptides with high affinity but often did so by utilizing highly optimized contacts with only a few ligand residues. In contrast, during natural selection inside the cell, each domain is exposed to thousands of other proteins and selection for biological function is driven not only by affinity but also by specificity. Consequently, it appears that natural domains recognize extended features of their cognate ligands, because this mode of recognition provides not only affinity but also specificity by enabling discrimination across multiple ligand positions.

Despite the considerable differences amongst our panel of evolved synthetic domains, correlation analysis did reveal trends that suggest the evolution of charge complementarity amongst binding sites and ligands. However, it is clear that cooperative interactions between multiple mutations are required for drastic changes in specificity, and the details of these intramolecular relationships remain to be elucidated.

Importantly, our synthetic domain family is considerably simpler than the natural PDZ domain family, as all members contain a maximum of only ten residue differences, which occur directly in the binding site. Furthermore, unlike the natural family, our synthetic family is expandable and the stage is now set for further large-scale *in vitro* evolution studies that will enlarge the synthetic family to address additional questions about the process of evolution. For example, the selection process could be repeated in the presence of competitor peptides to more closely mimic natural selection and to observe differences amongst selected populations in the presence or absence of competition. Also, the set of varied positions in the binding site could be reduced to narrow the degenerate sequence space, or alternatively, the set could be expanded to investigate the influence of second sphere residues. Finally, our methods are general and could be readily applied to other protein families.

In closing, it is worth noting that there have been impressive advances in computational methods that use statistical coupling analysis of natural sequence families to identify cooperative interactions amongst protein residues.[26] In the case of WW domains, these analyses have revealed the minimum information necessary to encode structure, and they have enabled the design of functional synthetic domains.[27,28] Our empirical methods are complementary to these computational methods, because we can use *in vitro* evolution to generate large yet defined datasets for computational analysis. In future experiments, we envision close interplay between high-throughput *in vitro* evolution and computational simulations in large-scale experiments designed to elucidate the rules governing protein structure and function.

Experimental

Selection and analysis of peptide-binding synthetic PDZ domains

A previously described Erbin-PDZ library (library 1, Fig. 1B) was used in the selection experiments.[18] N-Terminally biotinylated peptides (Fig. 1A) were immobilized on 96-well Maxisorp Immunoplates (NUNC, Rochester, NY) coated with neutravidin (Pierce, Rockford, IL) and phage from the Erbin-PDZ library were cycled through five rounds of binding selection, as described.[18] After the fifth round, individual phage clones were assayed for binding to the peptide by phage ELISA,[29] 48 positive clones from each peptide selection were subjected to DNA sequence analysis, and unique sequences were aligned.

High-throughput expression and purification of synthetic PDZ domains

Individual variants were expressed and purified as glutathione *S*-transferase (GST) fusion proteins, as described.[18] In brief, individual variants from the binding selections were pooled and used as the template for a PCR that amplified DNA fragments encoding the PDZ domains. The DNA fragments were ligated into an expression phagemid to produce an open reading frame encoding a fusion protein consisting of a hexa-histidine tag, followed by GST followed by an Erbin-PDZ variant. Bacteria harbouring individual clones were grown in a 96-well format in duplicate plates. In one plate, phage production was induced by the addition of M13-KO7 helper phage (New England Biolabs, Beverly MA) and the phage particles were used as templates for a PCR that amplified a DNA fragment that was subjected to DNA sequence analysis. In the other plate, glycerol was added to a final concentration of 10% (v/v) and the cultures were frozen and stored as stocks for protein expression. Protein expression and purification was also performed in a 96-well format. Bacteria were pelleted and lysed, and the lysates were loaded onto Phynexus tips containing Ni–NTA resin (PhyNexus, San Jose CA). After washing, bound protein was eluted with elution buffer (50 mM NaPi, pH 8.0, 300 mM NaCl and 250 mM imidazole). Protein concentrations were measured using a Bradford Assay (Biorad, Hercules, CA) and the yields of purified protein ranged from 0.05–0.150 mg per 1.5 ml expression culture.

Selection of peptide ligands for PDZ domains

Peptide-phage selections were performed using a library of random heptapeptides (10^{11} unique members) fused to the C terminus of the gene-8 major coat protein of M13 phage, as described.[18] The binding selections were performed in a 96-well format with one well dedicated to each target protein from the high-throughput purification. Phages from the peptide-phage library were cycled through five rounds of binding selection against each PDZ domain. After five rounds of selection, specific binding of the selected phage pool to the PDZ domain was verified by phage ELISA[29] and the positive

phage pools were subjected to DNA sequence analysis using 454 sequencing (454 Life Sciences, Branford, CT).

Deep sequence analysis by 454 sequencing

Each phage peptide pool was used as the template for a PCR with an individual forward primer comprising the 454 compatible 5′ end (5′-GCCTCCCTCGCGCCATCAG), 8 base pair barcode sequence and an annealing site for a constant region of the peptide-phage display vector (GCGCCCCCGGTGGCGGA-3′). For all phage peptide pools, we used a generic reverse primer comprising a 454 compatible 3′ end (5′-GCCTTGCCAGCCCGCTCA) and an annealing site for a constant region of the peptide-phage display vector (GCACTGAGTTTCGTCACCA-3′). Successful amplification of the correct 320 base pair DNA fragment from each phage pool was verified by agarose gel electrophoresis. The amplified DNA fragments were pooled and subjected to 454 DNA sequencing.[25] Each sequencing read was assigned to its correct pool on the basis of its unique barcode sequence. The DNA sequences were translated to decode the sequence of each selected peptide.

Specificity profiling

For each domain, the binding peptides were first aligned from the C terminus. The multiple PWMs (displayed as multiple logos in Fig. 2) were generated with a recent computational method,[21] based on the machine learning framework of mixture models, which enables uncovering multiple specificity in peptide profiles. This approach is particularly useful when different ligand positions display cooperativity (*i.e.* residues are not contributing independently to the binding). To determine the optimal number of PWMs, we split the set of peptides into clusters using standard hierarchical clustering algorithm until ligand positions are no longer correlated within each cluster. As a measure of correlation or cooperativity, we used the *Z*-score of the Mutual Information and a threshold linearly increasing from 3.5 to 10.5 was applied to account for the variable size of different peptide sets.

The specificity potential (Fig. 3) measures how specific a ligand position is. It is defined as $1 + \Sigma_{i=1}^{20} p_i \log_{20}(p_i)$, where p_i is the frequency of the amino acid i at this position in the phage profile of a given domain.

Affinity assays

The binding affinities of peptides for PDZ domains were determined as IC_{50} values with a competition ELISA, as described.[30] For each domain, optimal peptide pairs were synthesized with either a biotinylated or an acetylated N terminus. GST-PDZ protein was immobilized on assay plates coated with an anti-GST antibody (Sigma-Aldrich, St. Louis, MO). A fixed concentration of biotinylated peptide (10 μM) in PBS, 0.5% BSA, 0.1% Tween 20 (PBT buffer) was mixed with serial dilutions of the acetylated peptide and the mixture was transferred to the assay plates. After incubation for 1 h, the plates were washed with PBS and 0.05% Tween 20, incubated with horse radish peroxidase conjugated to neutravidin (Pierce, Rockford, IL) (1 : 10 000 dilution in PBT buffer), washed again, and detected with TMB (3,3′,5,5′-tetramethylbenzidine) peroxide substrate (KPL, Gaithersburg, MD). The IC_{50} was defined as the concentration of acetylated peptide that blocked 50% of biotinylated peptide binding to the immobilized GST-PDZ protein.

Acknowledgements

This work was supported by funding of the Canadian Institutes of Health Research for S. S. S (grant MOP-93684) and G. D. B (grant MOP-84324). D. G. acknowledges the financial support of the Swiss National Science Foundation (grant PBELA-120936).

References

1 T. Pawson and P. Nash, *Science*, 2003, **300**, 445–452.
2 H. K. Binz and A. Plückthun, *Curr. Opin. Biotechnol.*, 2005, **16**, 459–469.
3 S. S. Sidhu and S. Koide, *Curr. Opin. Struct. Biol.*, 2007, **17**, 481–487.
4 B. C. Cunningham and J. A. Wells, *J. Mol. Biol.*, 1993, **234**, 554–563.
5 G. Pal, J.-L. K. Kouadio, D. R. Artis, A. A. Kossiakoff and S. S. Sidhu, *J. Biol. Chem.*, 2006, **281**, 22378–22385.
6 W. A. Lim, *Curr. Opin. Struct. Biol.*, 2002, **12**, 61–68.
7 C. Vogel and C. Chothia, *PLoS Comput. Biol.*, 2003, **2**, e48.
8 M. Sheng and C. Sala, *Annu. Rev. Neurosci.*, 2001, **24**, 1–29.
9 B. Z. Harris and W. A. Lim, *J. Cell Sci.*, 2001, **114**, 3219–3231.
10 D. A. Doyle, A. Lee, J. Lewis, E. Kim, M. Sheng and R. MacKinnon, *Cell*, 1996, **85**, 1067–1076.
11 Z. Songyang, A. S. Fanning, C. Fu, J. Xu, S. M. Marfatia, A. H. Chishti, A. Crompton, A. C. Chan, J. M. Anderson and L. C. Cantley, *Science*, 1997, **275**, 73–77.
12 N. L. Stricker, K. S. Christopherson, B. A. Yi, P. J. Schatz, R. W. Raab, G. Dawes, D. E. Bassett, Jr., D. S. Bredt and M. Li, *Nat. Biotechnol.*, 1997, **15**, 336–342.
13 Y. Zhang, S. Yeh, B. A. Appleton, H. A. Held, P. J. Kausalya, D. C. Y. Phua, W. L. Wong, L. A. Lasky, C. Wiesmann, W. Hunziker and S. S. Sidhu, *J. Biol. Chem.*, 2006, **281**, 22299–22311.
14 P. Vaccaro, B. Brannetti, L. Montecchi-Palazzi, S. Philipp, M. H. Citterich, G. Cesareni and L. Dente, *J. Biol. Chem.*, 2001, **276**, 42122–42130.
15 M. A. Stiffler, J. R. Chen, V. P. Grantcharova, Y. Lei, D. Fuchs, J. E. Allen, L. A. Zaslavskaia and G. Macbeath, *Science*, 2007, **317**, 364–369.
16 R. Tonikian, Y. Zhang, S. L. Sazinsky, B. Currell, J.-H. Yeh, B. Reva, H. A. Held, B. A. Appleton, M. Evangelista, Y. Wu, X. Xin, A. C. Chan, S. Seshagiri, L. A. Lasky, C. Sander, C. Boone, G. D. Bader and S. S. Sidhu, *PLoS Biol.*, 2008, **6**, e239.
17 J. R. Chen, B. H. Chang, J. E. Allen, M. A. Stiffler and G. Macbeath, *Nat. Biotechnol.*, 2008, **26**, 1041–1045.
18 A. Ernst, S. L. Sazinsky, S. Hui, B. Currell, M. Dharsee, S. Seshagiri, G. D. Bader and S. S. Sidhu, *Sci. Signaling*, 2009, **2**, ra50.
19 D. Y.-C. Wang, S. Kumar and S. B. Hedges, *Proc. R. Soc. London, Ser. B*, 1999, **266**, 163–171.
20 R. Tonikian, X. Xin, C. P. Toret, D. Gfeller, C. Landgraf, S. Panni, S. Paoluzi, L. Castagnoli, B. Currell, S. Seshagiri, H. Yu, B. Winsor, M. Vidal, M. B. Gerstein, G. D. Bader, R. Volkmer, G. Cesareni, D. G. Drubin, P. M. Kim, S. S. Sidhu and C. Boone, *PLoS Biol.*, 2009, **7**, e1000218.
21 D. Gfeller, A. Ernst, E. Verschueren, P. Vanhee, N. Dar, L. Serrano, S. S. Sidhu, G. D. Bader and P. M. Kim, 2010, submitted.
22 J. M. Elkins, E. Papagrigoriou, G. Berridge, X. Yang, C. Phillips, C. Gileadi, P. Savitsky and D. A. Doyle, *Protein Sci.*, 2007, **16**, 683–694.
23 S. Karthikeyan, T. Leung and J. A. A. Ladias, *J. Biol. Chem.*, 2001, **276**, 19683–19686.

24 N. J. Skelton, M. F. T. Koehler, K. Zobel, W. L. Wong, S. Yeh, M. T. Pisabarro, J. P. Yin, L. A. Lasky and S. S. Sidhu, *J. Biol. Chem.*, 2003, **278**, 7645–7654.
25 M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. T. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. G. Yu, R. F. Begley and J. M. Rothberg, *Nature*, 2005, **437**, 376–380.
26 S. W. Lockless and R. Ranganathan, *Science*, 1999, **286**, 295–299.
27 W. P. Russ, D. M. Lowery, P. Mishra, M. B. Yaffe and R. Ranganathan, *Nature*, 2005, **437**, 579–583.
28 M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner and R. Ranganathan, *Nature*, 2005, **437**, 512–518.
29 K. Deshayes, M. L. Schaffer, N. J. Skelton, G. R. Nakamura, S. Kadkhodayan and S. S. Sidhu, *Chem. Biol.*, 2002, **9**, 495–505.
30 G. Fuh, M. T. Pisabarro, Y. Li, C. Quan, L. A. Lasky and S. S. Sidhu, *J. Biol. Chem.*, 2000, **275**, 21486–21491.