Science Signaling
AAAS

Rapid Evolution of Functional Complexity in a Domain Family

Andreas Ernst, Stephen L. Sazinsky, Shirley Hui, Bridget Currell, Moyez Dharsee, Somasekar Seshagiri, Gary D. Bader and Sachdev S. Sidhu (8 September 2009)

Science Signaling **2** (87), ra50. [DOI: 10.1126/scisignal.2000416]

Science Signaling (ISSN 1937-9145) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Copyright 2008 by the American Association for the Advancement of Science; all rights reserved.

PROTEIN DOMAINS

Rapid Evolution of Functional Complexity in a Domain Family

Andreas Ernst,[1]* Stephen L. Sazinsky,[2] Shirley Hui,[3,4] Bridget Currell,[5] Moyez Dharsee,[3] Somasekar Seshagiri,[5] Gary D. Bader,[3,4] Sachdev S. Sidhu[1]*†

(Published 8 September 2009; Volume 2 Issue 87 ra50)

Multicellular organisms rely on complex, fine-tuned protein networks to respond to environmental changes. We used in vitro evolution to explore the role of domain mutation and expansion in the evolution of network complexity. Using random mutagenesis to facilitate family expansion, we asked how versatile and robust the binding site must be to produce the rich functional diversity of the natural PDZ domain family. From a combinatorial protein library, we analyzed several hundred structured domain variants and found that one-quarter were functional for carboxyl-terminal ligand recognition and that our variant repertoire was as specific and diverse as the natural family. Our results show that ligand binding is hardwired in the PDZ fold and suggest that this flexibility may facilitate the rapid evolution of complex protein interaction networks.

INTRODUCTION

The evolution of protein interaction networks at the molecular level results in phenotype changes at the cellular level. In multicellular organisms, molecular networks have evolved to support increasingly complex signal processing functions that respond to the environment and regulate cell growth, polarity, and replication (*1*). Peptide recognition modules (PRMs) are key players in the evolution of multicellular complexity (*1*–*3*) and recognize short linear stretches of primary sequence in other proteins. Metazoan protein complexes in signaling systems are assembled by scaffolding proteins containing multiple PRMs, each recognizing a distinct set of ligands. The human genome encodes dozens of PRM families with up to several hundred members each. Each family is characterized by a common fold and core recognition motif, but individual family members possess distinct structural features that impart distinct specificities and functions within the broad structure and function of the family.

The observation that PRM families grow with organism complexity (*3*) prompts a simple hypothesis for how complex functions mediated by protein signaling networks arose during metazoan evolution. It is likely that each PRM family evolved from a progenitor domain with a characteristic fold and core recognition function. PRMs can usually fold and function autonomously, and thus, a progenitor gene could expand to form a large and diverse PRM family through a process of duplication (new family members), mutation (divergence of function), and shuffling (function in different contexts) (*4*, *5*). PRM folds that are tolerant of mutations that alter function could evolve rapidly through this process. However, natural genomes and PRM families provide only circumstantial evidence for this hypothesis, and it is difficult to address questions of how rapidly and efficiently the process might operate to generate families from progenitors.

To examine how PRM families evolve, we studied the prototypical and large PDZ (PSD-95, Discs-large, ZO-1) domain family. PDZ domains typically recognize specific C-terminal sequences to assemble protein complexes and are often used in scaffold proteins that organize specialized subcellular sites, such as epithelial junctions, neuronal postsynaptic densities, and immunological synapses (*6*–*8*). Large-scale specificity profiling of the PDZ domain family revealed 16 distinct specificity classes amongst the several hundred human and worm PDZ domains and showed that the PDZ fold is robust to mutations in the binding site and versatile in function (*9*). However, the average sequence identity between 250 human PDZ domains is less than 30%, and, consequently, it is difficult to trace the generation of family member specificities from a progenitor (*9*, *10*). This situation is common to natural PRM families, so we developed a precisely defined repertoire of PDZ domain variants to more easily study the process of family evolution.

RESULTS

A repertoire of PDZ domain variants reveals that its structure is highly stable

To investigate evolution in the PDZ domain family, we constructed a repertoire of domains with mutant binding sites displayed on phage, using the Erbin PDZ domain (Erbin-PDZ) (*11*) as the starting template or progenitor. We combinatorially mutated the core binding site, defined as 10 positions that contact peptide ligands in nine distinct PDZ-ligand structures (*11*). At each position, we allowed the wild-type Erbin-PDZ residue and other residues that are common in natural PDZ domains (Fig. 1A). We constructed two libraries to increase the diversity of the repertoire. Library 1 displayed 6×10^8 PDZ variants with an N-terminal epitope tag (the naïve population), and binding of phage particles to an antibody that recognizes the tag (anti-tag) was used to select for resistance to proteolysis, which can be used as a proxy for stable, structured proteins (the structure-selected population) (*11*–*13*). Sequencing of 158 "structure-selected" clones showed no bias for the wild-type sequence in 9 out of 10 positions across the binding site (Fig. 1B). At position 25, four amino acids were allowed, but the wild-type phenylalanine occurs in 80% of the population. In contrast, the wild-type

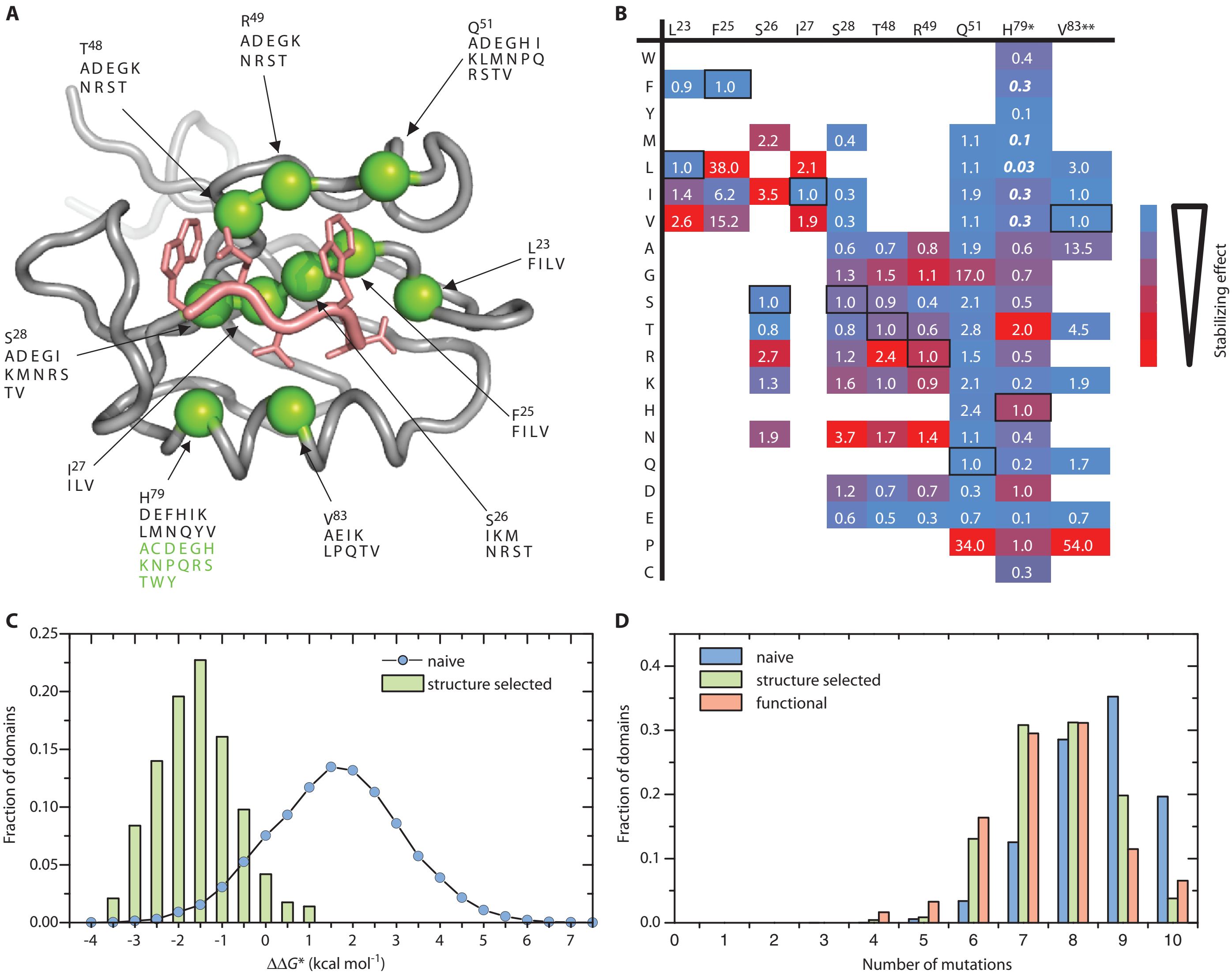
[1] Department of Protein Engineering, Genentech, 1 DNA Way, South San Francisco, CA 94080, USA. [2] Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, E19-563, Cambridge, MA 02139, USA. [3] Banting and Best Department of Medical Research, University of Toronto, Donnelly CCBR, 160 College Street, Toronto, Ontario, Canada M5S 3E1. [4] Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [5] Department of Molecular Biology, Genentech, 1 DNA Way, South San Francisco, CA 94080, USA.

*Present address: Banting and Best Department of Medical Research and Department of Molecular Genetics, University of Toronto, Donnelly CCBR, 160 College Street, Toronto, Ontario, Canada M5S 3E1.

†To whom correspondence should be addressed. E-mail:sachdev.sidhu@utoronto.ca

histidine at position 79 was replaced by hydrophobic residues in 94% of the population. To sample greater diversity at this position, we constructed a second library that did not encode hydrophobes at position 79. Library 2 contained 7 × 10^8 members (the naïve population) and was subjected to the same selection and analysis as library 1, resulting in 130 unique clones (the structure-selected population). The data from the two libraries were combined to form a set of 288 structure-selected PDZ domains for further analysis (table S1).

To assess the stabilities of the domains within the naïve or structure-selected populations from both libraries, at each mutated position, we compared the frequency of the occurrence of the wild-type residue of Erbin-PDZ (wt frequency) relative to the non–wild-type frequency (mutant frequency).



	L^{23}	F^{25}	S^{26}	I^{27}	S^{28}	T^{48}	R^{49}	Q^{51}	H^{79*}	V^{83**}
W									0.4	
F	0.9	1.0							***0.3***	
Y									0.1	
M			2.2		0.4			1.1	***0.1***	
L	1.0	38.0		2.1				1.1	***0.03***	3.0
I	1.4	6.2	3.5	1.0	0.3			1.9	***0.3***	1.0
V	2.6	15.2		1.9	0.3			1.1	***0.3***	1.0
A					0.6	0.7	0.8	1.9	0.6	13.5
G					1.3	1.5	1.1	17.0	0.7	
S			1.0		1.0	0.9	0.4	2.1	0.5	
T			0.8		0.8	1.0	0.6	2.8	2.0	4.5
R			2.7		1.2	2.4	1.0	1.5	0.5	
K			1.3		1.6	1.0	0.9	2.1	0.2	1.9
H								2.4	1.0	
N			1.9		3.7	1.7	1.4	1.1	0.4	
Q								1.0	0.2	1.7
D					1.2	0.7	0.7	0.3	1.0	
E					0.6	0.5	0.3	0.7	0.1	0.7
P								34.0	1.0	54.0
C									0.3	

Fig. 1. Structural and functional fitness of the Erbin-PDZ variant repertoire. (**A**) Design of the Erbin-PDZ variant repertoire. Erbin-PDZ (gray) is shown with a bound peptide ligand (WETWV$_{COOH}$; pink) (*26*) (PDB entry 1N7T) (*11*). Binding-site positions that were subjected to combinatorial mutagenesis are depicted as green spheres; the wild-type sequences and position numbers are shown, and the library-encoded mutations are listed below. For position 79, the mutations encoded in library 1 or 2 are listed in black or green text, respectively. (**B**) Wt/mutant ratios for sequences of structure-selected Erbin-PDZ variants. Mutations predicted to stabilize (wt frequency/mutant frequency < 1) or destabilize (wt frequency/mutant frequency > 1) the protein are highlighted in blue or red, respectively. The ratios were derived from the sequences of 288 variants. * For F, M, L, I, and V, the wt frequency/mutant frequency ratio is based on 157 sequences from library 1 (italic and bold) and for the remaining amino acids on 130 sequences from library 2. ** At position V^{83}, no Pro occurred in the data set. We assumed a count of 1 Pro to highlight the destabilizing effect. (**C**) Distribution of ΔΔ*G** relative to Erbin-PDZ. Distributions are shown for the naïve repertoire before selection for structure (blue circles) and for 288 structure-selected domains in the population after selection for stable domains displayed on phage particles (green bars). (**D**) Distribution of mutations in synthetic binding sites. The following populations were analyzed: naïve (blue bars), 237 structure-selected domains (green bars), and 61 functional domains that recognize C-terminal peptides (red bars). The sequences of the domains used in the analysis are shown in table S2.

As shown previously for Erbin-PDZ (*11*) and several other proteins (*14–16*), these data can be used to estimate the effects of mutations on stability as statistical free-energy values ($\Delta\Delta G^*$), by substituting the wt frequency/mutant frequency ratio in place of the equilibrium constants in the standard free-energy equation, as follows: $\Delta\Delta G^* = RT \ln(K_{\mathrm{wt}}/K_{\mathrm{mutant}}) = RT \ln$(wt frequency/mutant frequency) (Fig. 1B and table S1). For each mutation, the $\Delta\Delta G^*$ is calculated by averaging data from many independent PDZ domain variants, and consequently, any slight nonadditive interactions between particular residue pairs in a particular variant are likely to be minimal contributors to the overall $\Delta\Delta G^*$, as shown previously in similar combinatorial studies of other proteins (*15*, *16*). Thus, we assumed additivity between positions and calculated the stability of each variant relative to wild type by summing the $\Delta\Delta G^*$ values across mutated positions (Fig. 1C). This analysis reveals that the naïve population of Erbin-PDZ variants is only moderately destabilized, as the population forms a Gaussian distribution centered at $\Delta\Delta G^*$ ~1.5 kcal/mol. The structure-selected population of 288 variants was stabilized relative to the naïve population, with the distribution centered at $\Delta\Delta G^*$ ~–1.5 kcal/mol. Within the synthetic repertoire, the large proportion of stable PDZ domains with diverse binding site sequences shows that the fold is highly tolerant to mutations that could alter specificity.

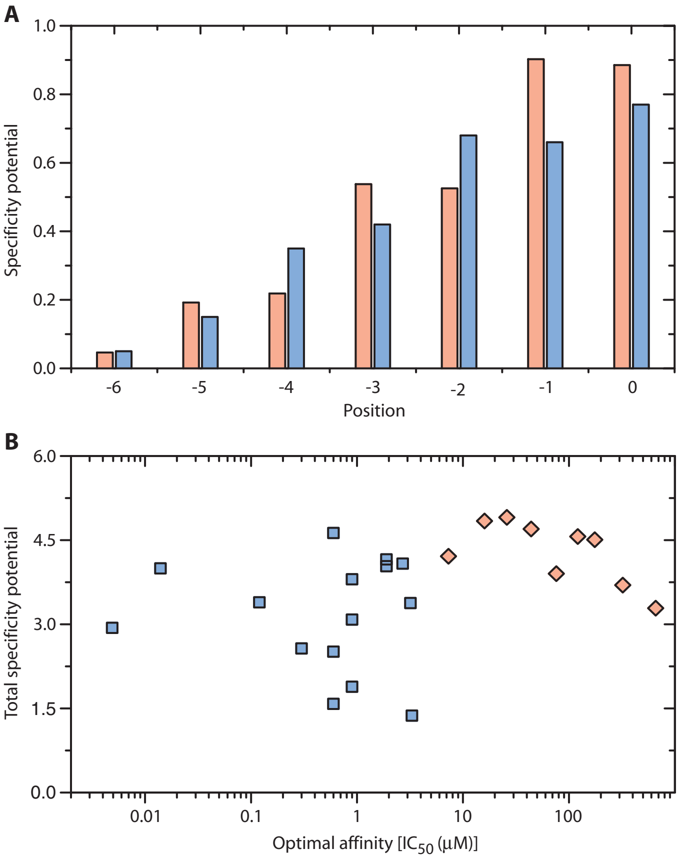


Fig. 2. Specificity and affinity of variant and natural PDZ domain–ligand interactions. (**A**) The mean specificity potential (SP) value at each ligand position is shown for 51 functional Erbin-PDZ variants (red bars) and 73 natural domains (blue bars) (*9*). (**B**) The plot shows the total SP (SP^t) summed over all ligand positions (*y* axis) and the affinities for optimal ligands designed to match the PWM specificity profiles (*x* axis) for nine Erbin-PDZ variants (red diamonds) and 15 natural domains (blue squares). The affinity data are shown in table S3.

Ligand-binding function is hardwired in the PDZ domain fold

We next investigated ligand-binding functionality within the structure-selected repertoire of PDZ domain variants. We purified 237 of the 288 structure-selected variants and used a phage-displayed library of random C-terminal heptapeptides (~10^{11} members) to ascertain whether the domains recognized C termini. Strikingly, we found that 61 domains (~25%) were functional by this criterion (table S1). A comparison of mean number of mutations shows only minor differences when the functional population (7.5 mutations per domain) is compared to the structure-selected population (7.7 mutations per domain) or the naïve population (8.5 mutations per domain) (Fig. 1D). Moreover, even if we conservatively assume that only those domains with stabilities close to that of the wild type are structured ($\Delta\Delta G^*$ <1.0 kcal/mol), this amounts to 30% of the naïve repertoire. Given that one-fourth of structured domains are functional, we estimate that ~10^8 domains in the naïve repertoire would be functional (7.5% of 1.3 × 10^9). Even without any selection for structure or function, our randomly generated repertoire of PDZ domain binding sites constitutes a rich source of functional domains, and thus, ligand-binding function appears to be hardwired in the PDZ domain fold.

Erbin-PDZ variants are as specific as natural domains

We compared the specificity of the functional Erbin-PDZ variants to those of natural PDZ domains, using the specificity potential (SP) metric, which is based on the position weight matrix (PWM), calculated from the set of binding peptides for each domain, and which varies from one (most specific) to zero (least specific) (*9*). Across the last seven positions of the peptide ligand, the Erbin-PDZ variants exhibit SP values that are comparable to those of worm and human domains (Fig. 2A) (*9*). To assess ligand affinities, we designed optimal peptide ligands for nine Erbin-PDZ variants on the basis of their specificity profiles and compared their affinities to those of optimal ligands for 15 natural domains (table S2). For eight of the nine variant ligands, we could not detect any interaction with Erbin-PDZ (fig. S1). In contrast, each domain variant recognized its cognate ligand, but the optimal binding interactions are generally of lower affinity than those for natural domains (Fig. 2B). Thus, the domain variants recognize C-terminal ligands that are not recognized by the wild-type Erbin-PDZ. Although these variant-ligand interactions are relatively weak, they are as specific as those of natural domains.

The Erbin-PDZ variant family is as diverse as the natural PDZ family

We compared the specificity profiles of our Erbin-PDZ variants to those of 52 previously mapped human domains (*9*), yielding a matrix of all pairwise profile distances. A comparison of the distances of human and variant profiles to that of the Erbin-PDZ profile reveals that the variant specificities are as diverse as the human specificities (Fig. 3A). A two-dimensional representation of the complete distance matrix shows that natural and variant domains form distinct clusters, indicating that the variant domains are more closely related to each other than to Erbin-PDZ or other natural domains (fig. S3). Within the plot, the Erbin-PDZ variants cover an area comparable to that occupied by the human domains, indicating that the variant repertoire, derived without any selection for ligand-binding function, covers a functional diversity space that is comparable to that of the human PDZ family, which has evolved over more than a billion years. Indeed, clustering of the Erbin-PDZ variants reveals 14 distinct specificity clusters (fig. S4), of which 7 match natural clusters (*9*) and 7 are not known to occur in nature.

To further investigate Erbin-PDZ variants with specificities resembling those of natural domains, which are located at the interface between the two groups in the two-dimensional plot (fig. S3), we examined the binding sites of matched variant-natural pairs (Fig. 3B). Binding site comparison reveals that the Erbin-PDZ variants establish natural specificities with binding sites that are different from those of their natural counterparts. Indeed, it is

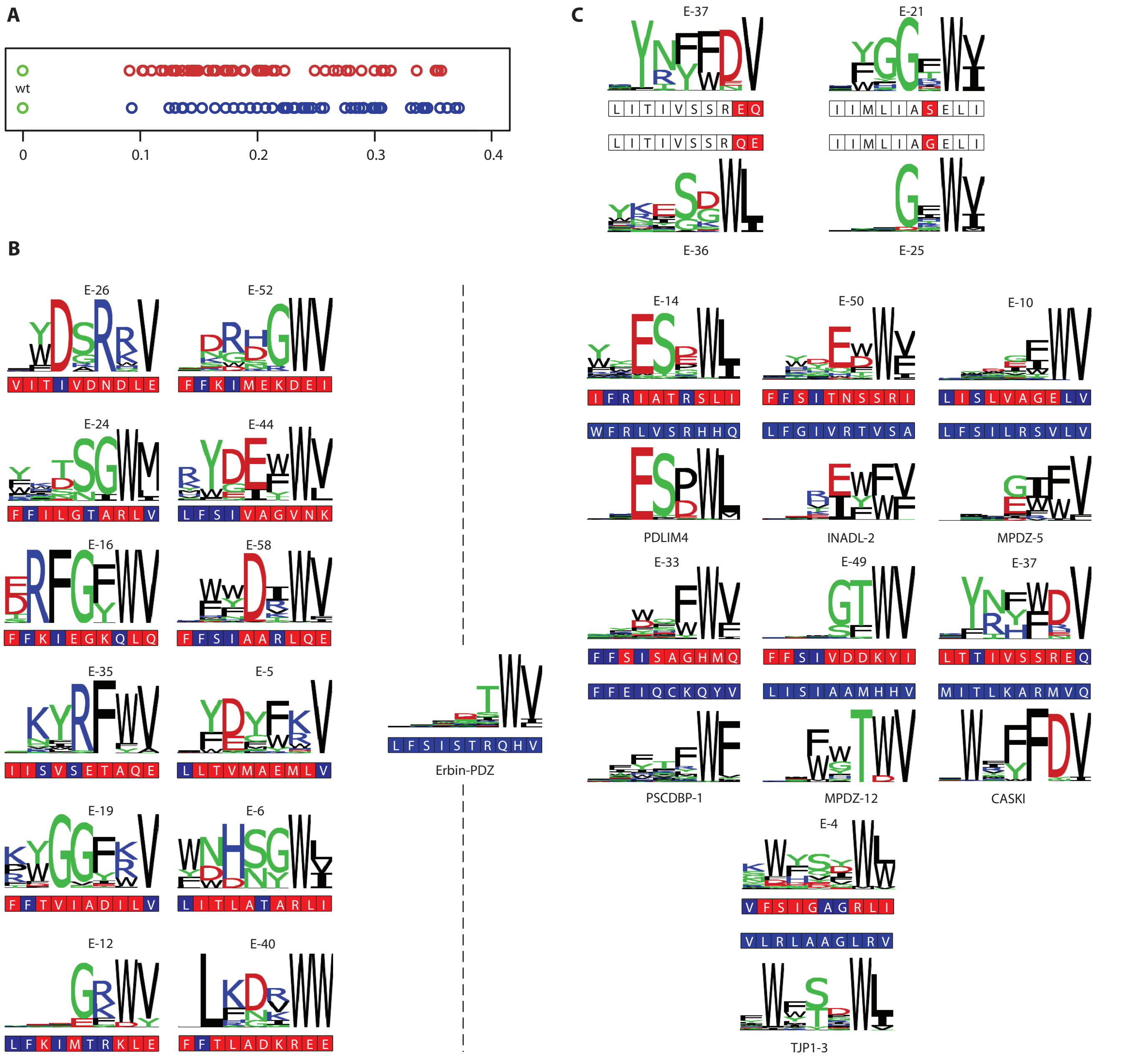


Fig. 3. The functional diversity of the Erbin-PDZ variant repertoire and natural PDZ domains. (A) Comparison of the distances of the Erbin-PDZ specificity profile (green) to the variant (red) or human (blue) (*9*) specificity profiles. (B) Erbin-PDZ variants exhibit natural and specificity profiles not found in nature. In the middle is the Erbin-PDZ specificity profile represented as a sequence logo; shown below is the sequence of the 10 binding-site positions that were subjected to mutagenesis. Right of the dashed line are pairs of logos for Erbin-PDZ variants (upper) and natural domains (lower) with similar specificity profiles; the binding site alignments are shown between the logos and are colored blue for sequences that match or red for those that do not match the natural sequence (*30*). Left of the dashed line are Erbin-PDZ variants with specificity profiles that have no natural counterpart; the binding site sequences that match (blue) or do not match (red) that of Erbin-PDZ are indicated. The name of the natural domain (*9*) is shown below the logo and the number of the Erbin-PDZ variant (table S2) is shown above the logo. (C) Similar Erbin-PDZ variants can exhibit similar or divergent ligand specificities. The specificity profiles and sequence alignments as in (B) except differences are highlighted in red. The binding sites of domains E-37and E-36 differ at only two positions, yet the specificity profiles are different. The binding sites of domains E-21 and E-25 differ at only one position, and the specificity profiles are similar.

possible to encode similar specificities with completely different binding site compositions (Fig. 3B), or in contrast very different specificities with similar binding sites (Fig. 3C). Consequently, two variants that connect to the same natural PDZ domain may or may not have similar binding site compositions. Furthermore, many domain variants exhibit specificities that have not been observed amongst natural domains (Fig. 3B). The diversity of ligand-binding function and binding site sequence in Erbin-PDZ variants suggests that natural domains do not saturate the specificity potential of the PDZ fold.

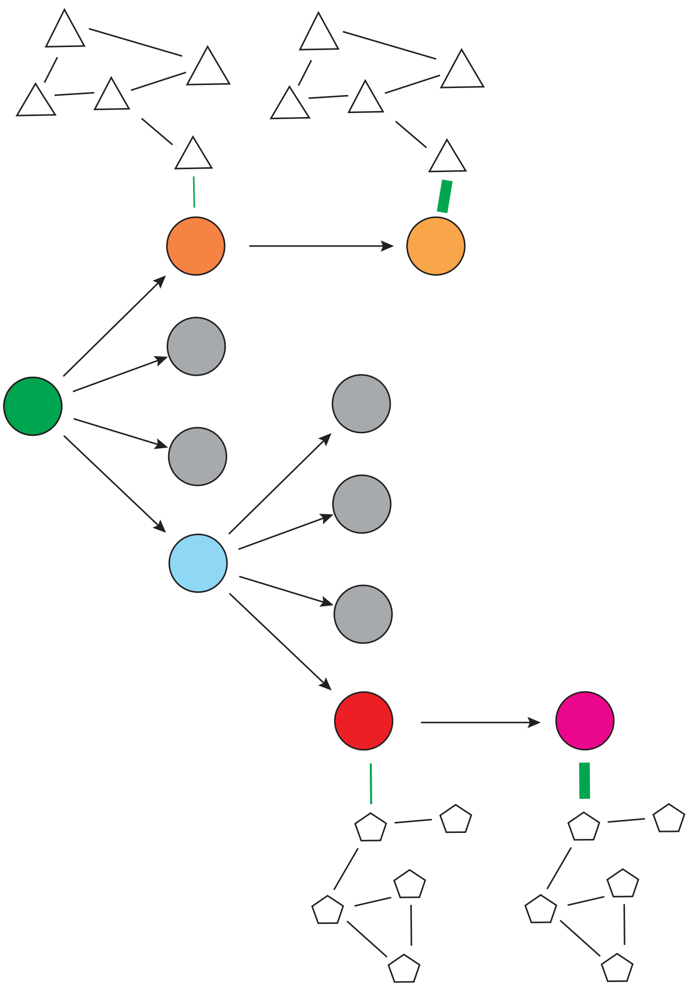


Fig. 4. Model for the evolution of protein interaction networks mediated by PDZ domains. A progenitor domain (green) undergoes duplications and mutations (black arrows) to yield descendants with diverse binding sites within the same structural fold. One-quarter of the structured descendants (orange) interact with the C termini of proteins and establish weak but specific connections with proteins in networks (thin green line). If the new interactions confer favorable phenotypes, the connection may be stabilized by further mutations that increase affinity (thick green line). Other descendants (blue) may undergo further duplications and mutations and some of these descendants (pink) may also establish new network connections. Structured descendants that lose the ability to recognize C-terminal sequences (gray) may be discarded, or they may diverge and acquire other functions associated with the PDZ domain fold (for example, recognition of lipids or protein regions other than C termini) (*6*, *35*, *36*).

DISCUSSION

We present a large-scale study of mutational tolerance and the in vitro evolution of a PRM family. Previous studies of another PRM, the WW domain, have shown that only limited sequence information is required to specify the domain fold (*17*, *18*). In the PDZ domain, we show that ligand-binding function is inherent to the fold. Without any selective pressure, a substantial proportion of structured domains are endowed with the ability to bind ligands. Variation at a limited set of positions in the binding site achieves large functional complexity in the fold. Together, these structural and functional properties of PRMs could enable facile evolution of diverse functions within a domain family, which, in turn, would allow PRM-mediated interaction networks to rapidly rewire during evolution.

Our findings provide direct empirical evidence for a model of network evolution at the molecular level (Fig. 4), which is similar to a previous model derived from natural databases of interacting domains (*19*). Gene duplication and mutation could produce many potential binding domains from a single progenitor that is structurally robust but functionally malleable. We find that these newly formed domains can achieve high specificity directly from a naïve repertoire but the initial interactions are relatively weak. Thus, new specificities could evolve through a process of trial without error, because the weak but specific interactions would be unlikely to interfere with existing tight interactions, but beneficial interactions could be tightened by further mutations to optimize new network connections. Thus, PDZ domains and other PRMs could support the rapid evolution of network complexity, consistent with the proposition that PRMs have played a major role in the evolution of eukaryotic interaction networks (*19*).

Functional plasticity, resembling a neutral network of mutationally connected states, has been shown to contribute to the evolution of new enzymatic functions (*20*, *21*). Previous studies suggested that the evolutionary landscape of protein function resembles neutral networks (*22*), allowing covert properties to become more prevalent while maintaining the original functionality (*20*, *21*). However, these studies relied on the application of selection pressure for functionality throughout the selection experiment. Additionally, a third of the time, single–amino acid mutations produced nonfunctional variants or destabilized the enzymes (*23*, *24*), which if extrapolated to eight amino acid changes would produce only 1.5% functional proteins. We show that the PDZ domain fold is highly robust to an average of eight mutations in the ligand-binding site and that, even in the absence of selection pressure for functionality, ~7.5% of the population was predicted to be functional for C-terminal peptide recognition. Additionally, the functional subpopulation was almost as heavily mutated as the naïve population. Furthermore, it is apparent that the PDZ fold can support specificities that do not exist in nature, suggesting that the full potential of the PDZ family has not been exhausted and that novel domains could be exploited to wire new networks naturally through continuing natural evolution or intentionally through engineered synthetic biology.

METHODS

Erbin-PDZ phage display library construction

Phage display libraries were constructed from previously described methods (*25*) with a phagemid that was designed for the phage display of Erbin-PDZ with a peptide epitope tag (gD-tag: SMADPNRFRGKDL) (*26*) fused to the N terminus (*11*). In the binding site of Erbin-PDZ, 10 positions were randomized to encode residues that occur frequently in the curated PFAM alignment of 58 PDZ domains (*27*). In both libraries, nine positions were mutated by replacing the wild-type codons with the following degenerate codons (N = A/C/G/T, V = A/C/G, R = A/G, S = C/G, W = A/T, H =

A/C/T): Leu^{23} (NTC), Phe^{25} (NTC), Ser^{26} (ANS), Ile^{27} (VTC), Ser^{28} (RNS), Thr^{48} (RVS), Arg^{49} (RVS), Gln^{51} (VNS), Val^{83} (VHA). The two libraries differed in terms of the degenerate codons used to replace the codon for the 10th position (His^{79}), which was replaced by an NWS or NVS codon in library 1 or 2, respectively. Each constructed library contained >10^{10} unique members, which greatly exceeded the theoretical diversities of all possible combinations encoded by the degenerate codons (library 1, 5.6×10^8combinations; library 2, 7.0×10^8 combinations). Thus, we expect the libraries to contain all possible sequence combinations. The amino acid position numbering for the binding site corresponds to the numbers used in the previously reported structure of Erbin-PDZ [Protein Data Bank (PDB) entry 1N7T] (*11*).

Assessment of the effects of binding site mutations on Erbin-PDZ stability

The phage-displayed Erbin-PDZ libraries were used to assess the effects of mutations on protein stability, as described (*11*). Briefly, phage pools representing each Erbin-PDZ library were subjected to rounds of selection for binding to an antibody recognizing the N-terminal gD-tag epitope to select for variants that are displayed on the phage particles. The enrichment process was monitored by determining the phage titer in each round of selection. After five rounds of selection, the enrichment reached a plateau, and 158 or 130 sequences of unique variants were compiled for libraries 1 and 2, respectively (table S2). At each randomized position, the frequencies of amino acids were normalized to correct for bias in the degenerate codons (*15*), and the normalized frequencies were used to calculate the wild type/mutant (wt/mutant) ratio for each mutation at each position (Fig. 1B). The effect of each mutation on the stability of Erbin-PDZ relative to the that of the wild type was estimated as a statistical $\Delta\Delta G^*$ value by using each wt/mutant ratio in the standard free-energy equation, as follows: $\Delta\Delta G^* = RT\ln(K_{\mathrm{wt}}/K_{\mathrm{mutant}}) = RT\ln(\mathrm{wt/mutant})$. For each unique selected variant, it was assumed that the effects of mutations were additive, and the change in stability relative to that of wt Erbin-PDZ was estimated as a $\Delta\Delta G^*$ value by summing the calculated statistical $\Delta\Delta G^*$ for all of the individual mutations in the binding site (table S2). The $\Delta\Delta G^*$ values were calculated with the combined sequences from libraries 1 and 2 for all positions except His^{79}, for which the sequences from library 1 were used for mutations to Phe, Leu, Ile, Met, or Val, and the sequences from library 2 were used for all other mutations. For the naïve library (not enriched for structural stability), the $\Delta\Delta G^*$ distribution was calculated for a computer-generated set of Erbin-PDZ variants matching the naïve library codon distribution using the above calculated $\Delta\Delta G^*$ values for individual mutations.

High-throughput expression and purification of the Erbin-PDZ variants

Phage representing the 288 sequenced Erbin-PDZ variants from libraries 1 and 2 subjected to gD-tag binding selection (table S2) were pooled and used as the template for a polymerase chain reaction (PCR) that amplified DNA fragments encoding the PDZ domains. The DNA fragments were ligated into an expression phagemid containing the β-lactamase gene and an isopropyl-β-D-thiogalactopyranoside (IPTG)–inducible lac promoter for heterologous protein expression. The ligation produced an open reading frame encoding a fusion protein consisting of a hexahistidine tag, followed by glutathione *S*-transferase (GST), followed by an Erbin-PDZ variant. The ligation was transformed into *Escherichia coli* XL1-blue (Stratagene), and individual colonies were used to inoculate 450 μl of 2xYT (yeast extract–tryptone) medium supplemented with carbenicillin (100 μg/ml) in a 96–deep-well plate. Each 96–deep-well plate was used to inoculate a duplicate plate containing 2xYT medium supplemented with carbenicillin (100 μg/ml), kanamycin (25 μg/ml), and M13-K07 helper phage (10^9 phage/ml) to induce the production of phage particles containing the expression phagemid DNA (New England Biolabs, Beverly, MA). The plates were grown with shaking at 37°C for 8 hours. To the original plate, glycerol was added to a final concentration of 10% (v/v) and the cultures were frozen and stored as stocks for protein expression. In the duplicate plate, the cultures were centrifuged and the phage supernatants were used as templates for a PCR with primers that amplified a DNA fragment encoding the Erbin-PDZ variant and added M13fwd (TGTAAAACGACGGCCAGT) and M13rev (CAGGAAACAGCTATGACC) sequencing primers to the 5′ and 3′ ends, respectively. The PCR fragments were subjected to DNA sequencing and the analysis of four 96-well plates yielded the sequences of 237 unique Erbin-PDZ variants representing a good sampling of the pool of 288 Erbin-PDZ variants from the gD-tag binding selection.

For protein expression, frozen bacterial stocks from an original 96-well plate were used to inoculate wells containing 450 μl of 2xYT medium supplemented with carbenicillin (50 μg/ml). After overnight growth, each inoculated plate was used to inoculate a 96-well expression plate with wells containing 1.5 ml of LB medium supplemented with carbenicillin (50 μg/ml) and 0.4 mM IPTG (7 μl of inoculum per well). The expression plates were incubated for 48 hours at 37°C with vigorous shaking. The bacterial cultures were centrifuged at 3500 rpm (2951*g*) for 10 minutes and the bacterial pellets were stored overnight at –20°C. The frozen pellets were lysed by resuspension with vigorous shaking at 30°C in 300 μl of resuspension buffer [50 mM NaPi (pH 8.0), 300 mM NaCl, 5 mM imidazol, Protease Inhibitor (Roche, 1 tablet per 100 ml), Lysozyme (0.2 mg/ml; Sigma), stock solution of DNase I (deoxyribonuclease I, 0.15 μl/ml of 5000 U/ml; Roche)]. Bacterial lysates were loaded onto Phynexus tips (PhyNexus, San Jose CA) containing 10 μl of Ni-NTA resin equilibrated with 150 μl of loading buffer [50 mM NaPi (pH 8.0), 300 mM NaCl, 5 mM imidazol]. The tips were washed with 150 μl of loading buffer and subsequently with 150 μl of wash buffer [50 mM NaPi (pH 8.0), 300 mM NaCl, 10 mM imidazol]. Bound protein was eluted with 100 μl of elution buffer [50 mM NaPi (pH 8.0), 300 mM NaCl, 250 mM imidazol]. To ensure efficient loading, washing, and elution of the Phynexus tips, pipetting was repeated seven times at each step. All steps were performed on an LM 600 liquid handler with a 96-channel pipette head (Dynamic Devices, Apex, NC). Protein concentrations were measured with a Bradford assay (Bio-Rad Laboratories, Hercules, CA) and the yields of purified protein ranged from 50 to 150 μg per 1.5-ml expression culture. Protein purities were assessed by SDS-PAGE and were typically >90%.

Selection and sequencing of the peptide ligands for the Erbin-PDZ and variants

Peptide-phage selections were performed with 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 (*28*). Phage pools representing the naïve peptide library were produced from *E. coli* SS320 cultures grown overnight at 37°C in superbroth medium supplemented with kanamycin (25 μg/ml), carbenicillin (100 μg/ml), and 0.4 mM IPTG. Phage were harvested by precipitation with polyethylene glycol–NaCl and resuspended at a final concentration of 10^{13} phage/ml in assay buffer [phosphate-buffered saline (PBS), proclin (15 parts per million), 0.5% bovine serum albumin (BSA), 0.5% Tween 20].

The binding selections were performed in a 96-well format with one well dedicated to each target protein. Target proteins from the above-described high-throughput purification were diluted 1:11 in PBS (5 to 15 μg/ml final concentration) and coated overnight at 4°C on a Maxisorp microtiter plate (NUNC, Denmark) (100 μl per well). The wells were blocked for 2 hours with blocking buffer (PBS, 0.2% BSA). In the first selection round, 100 μl of the phage pool representing the naïve peptide library was added to each well and incubated for 2 hours at 4°C. The plate was washed eight times with cold wash buffer (PBS and 0.5% Tween-20), and bound phage were directly infected into bacteria by the addition of 100 μl of *E. coli* XL1 blue (A_{600} = 0.8) in 2xYT to each well and incubation for 25 min at 37°C with shaking. M13K07 helper phage (NEB, Beverley MA) were added to each well to a final concentration of 10^{10} phage/ml to enable phage production and the incubation was continued for 45 min. For storage of samples from each selection round, 10 μl from each well was transferred to a well containing 90 μl of 2xYT supplemented with 15% glycerol (v/v); the samples were frozen in liquid nitrogen and stored at –80°C. The remaining culture volume was transferred to 1.4 ml of 2xYT, kanamycin (25 μg/ml), carbenicillin (100 μg/ml), and 0.4 mM IPTG, and the incubation was continued overnight for phage production. The plates were centrifuged at 3000 rpm (2100*g*) and 500-μl aliquots of the phage supernatants were transferred to a fresh 96-well plate and incubated at 65°C for 20 min to kill remaining bacteria. Four additional rounds of selection were performed with methods identical to those for round 1, except that sterilized phage supernatants from preceding rounds were used directly for binding to immobilized target proteins. Sterilized phage supernatants from each round were stored at 4°C or −20°C for short- or long-term storage, respectively.

The progress of the selection at each round was also followed by analyzing aliquots of sterilized phage supernatants in a phage enzyme-linked immunosorbent assay (ELISA) to detect specific binding of the phage pool to immobilized target protein (*29*). Phage pools from round 5 with strong ELISA signals were subjected to DNA sequence analysis. For DNA sequence analysis, phage from each pool of interest were infected at a low multiplicity of infection into 100 μl of *E. coli* XL1 blue in a 96-well format. After 30 min of incubation at 37°C with shaking, serial dilutions of the cultures were spread on LB agar plates supplemented with carbenicillin (100 μg/ml) and grown overnight at 37°C. Single colonies were picked into 450 μl of 2xYT, kanamycin (25 μg/ml), carbenicillin (100 μg/ml), and 10^{10} phage/ml M13K07 helper phage in 96-well plates, and the cultures were grown overnight at 37°C with shaking. The phage supernatants were harvested and sterilized as described above and were used as the template for a PCR that amplified a fragment that was subjected to DNA sequencing analysis to obtain the sequence of the displayed peptide, as described (*28*). For each Erbin-PDZ variant, 48 to 96 positive clones were sequenced and, for clones that were unique at the DNA level, peptide sequences were aligned, with the C terminus as an anchor position. We identified an average of 36 unique peptides per domain.

Specificity profiling of the Erbin-PDZ and variants

For each PDZ domain, the set of aligned peptide ligands was used to derive a PWM that represents a simple statistical model of the binding motif (the specificity profile) of each domain. The PWMs were constructed by calculating the distribution of amino acid residues found at each of the seven positions of the ligand and correcting for codon bias in the naïve library with an NNK codon correction, as described (*28*). PWMs were visualized as sequence logos (*30*).

For each position of each PWM, the SP was calculated as the total information content in bits expressed as the log to the base 20. An SP value of 1 means the domain is absolutely specific for a single amino acid at that position, and a value of 0 means that there is no preferred amino acid at that position. The SP value for each position was corrected for codon bias with an NNK codon correction, as described (*9*). Because SP scores can be artificially high when there are only a few peptides used to create a PWM, a saturation analysis was performed to determine which PWMs contained sufficient peptide numbers. For each synthetic domain, 1000 sets of peptides were created by randomly sampling *x* number of peptides from the original set of peptides. Sets were created for *x* = 5, 10, 15, …*N* (where *N* = maximal number of peptides) or, for domains with less than 25 peptides, *x* = 2, 4, 6, …*N* to more finely sample the space. For each of the sample sets, SP scores were calculated to determine if the number of peptides affected the SP score. On the basis of this analysis, 51 domains with more than 12 peptides each were kept for further analysis.

Calculating the distance between binding site sequences and PWMs

The distance between two domain binding site sequences was calculated as the number of mismatched positions/(aligned sequence length – gap positions). The sequence identity between two sequences was calculated as 1.0-distance. This was used for binding-site and whole-domain sequences for all relevant calculations.

The distance between two PWMs is calculated as the average of the normalized Euclidean distance of the columns of the PWMs (Eq. 1). The similarity between two PWMs was calculated as 1.0-distance with the following equation (Eq. 2):

$$\mathrm{Dist_{SP}}(a,b) = \frac{1}{\sqrt{2}}\sum_{i=1}^{w}\sqrt{\sum_{\mathrm{L}\in(20\ \mathrm{aa's})}(a_{i\mathrm{L}}-b_{i\mathrm{L}})^2} \qquad (1)$$

$$\mathrm{Sim_{SP}}(a,b) = 1.0 - \mathrm{Dist_{SP}}(a,b) \qquad (2)$$

where *a* is PWM A, *b* is PWM B, *w* is the number of columns in the PWM, and L is amino acids (aa's).

This metric is normalized such that 0 represents perfectly similar PWMs and 1 represents perfectly dissimilar PWMs.

Relationship between binding site sequence identity and PWM similarity

For each Erbin-PDZ variant and natural PDZ domain pair, the PWM similarity and binding site sequence identity were calculated. At each binding site sequence identity (0.0, 0.1, 0.2, …1.0), box plots were created to summarize the PWM similarities between pairs (fig. S2). Box plots were created with the R statistical software.

Two-dimensional specificity distance plot

PWM similarities between all combinations of domain pairs were calculated with Eq. 2 and stored in an *n* × *n* matrix, where *n* is the total number of Erbin-PDZ variants and natural PDZ domains. The similarity relationships were visualized as a fully connected directed network with the Cytoscape network visualization and analysis software (*31*). Nodes in the network represent the Erbin-PDZ variants or natural PDZ domains. Edge lengths attempt to represent the similarity, in terms of ligand sequence preference, between nodes, with shorter edges representing more similar nodes. This was accomplished by applying an edge-weighted spring embedded layout to the network with the weights set to be the similarity attribute, the minimum edge weight set to 0.0, the maximum edge weight set to be 1.0, spring rest length set to be 80, and the spring strength set to be 20. To reduce the network complexity and highlight interactions of interest, we displayed only edges connecting Erbin-PDZ variant nodes to their three most similar (Eq. 2) neighbors that are natural nodes. This was accomplished by importing nearest-neighbor edge attribute information into the network. These edge attributes represented the neighbor rank of a node *i* with respect to node *j* for each directed $\mathrm{edge}_{(i,j)}$. Edges connecting natural to natural nodes or Erbin-PDZ variant to Erbin-PDZ variant nodes and edges with nearest neighborhood attributes >4 were not visualized. A customized Cytoscape visual style was then applied to yield the final visualization.

Two-dimensional maps of the similarity relationships between Erbin-PDZ variants and natural PDZ domains were created with Cytoscape, multidimensional scaling, and principal component analysis using the similarity matrix, distance matrix, and vectorized PWMs, respectively. The maps were then compared by plotting Shepard diagrams and calculating the Spearman rank correlations between the original distance matrices and the distance matrices created from the two-dimensional maps. Maps produced with Cytoscape had more accurate and better correlated distance matrices and were, therefore, used to create the final two-dimensional maps.

Affinity assays

The binding affinities of peptides for PDZ domains were determined as IC_{50} values with a competition ELISA, as described (*32*). The IC_{50} value was defined as the concentration of peptide that blocked 50% of PDZ domain binding to immobilized peptide. For each domain, optimal peptide pairs were synthesized with either a biotinylated or an acetylated N terminus. Assay plates were prepared by immobilizing the biotinylated peptide on Maxisorp immunoplates coated with neutravidin and blocked with BSA. A fixed concentration of GST-PDZ fusion protein in PBS, 0.5% BSA, 0.1% Tween 20 (PBT buffer) was preincubated for 3 hours with serial dilutions of the acetylated peptide and then transferred to the assay plates. After 15 min of incubation, the plates were washed with PBS and 0.05% Tween 20, incubated with a mixture of an antibody that recognizes GST (0.5 μg/ml) and horseradish peroxidase conjugated to a rabbit antibody against mouse immunoglobulin G (1:2000 dilution) in PBT buffer, washed again, and detected with TMB (3,3′,5,5′-tetramethylbenzidine) peroxidase substrate. Data for PDZ domains from the following proteins were reported previously: Erbin (*11*), Scrib and TJP1 (*33*), HtrA2 (*28*), HtrA1 and HtrA3 (*34*).

SUPPLEMENTARY MATERIALS

www.sciencesignaling.org/cgi/content/full/2/87/ra50/DC1
Table S1. Structure-selected Erbin-PDZ variants.
Table S2. IC_{50} values for synthetic peptides binding to PDZ domains.
Fig. S1. Binding of Erbin-PDZ to synthetic peptides.
Fig. S2. Comparison of binding site similarity and specificity profile similarity for Erbin-PDZ variants.
Fig. S3. Two-dimensional representation of PDZ domain specificity space.
Fig. S4. Tree representation of the specificity profiles of Erbin-PDZ variants.

REFERENCES AND NOTES

1. T. Pawson, P. Nash, Assembly of cell regulatory systems through protein interaction domains. *Science* **300**, 445–452 (2003).
2. W. A. Lim, The modular logic of signaling proteins: Building allosteric switches from simple binding domains. *Curr. Opin. Struct. Biol.* **12**, 61–68 (2002).
3. C. Vogel, C. Chothia, Protein family expansions and biological complexity. *PLoS Comput. Biol.* **2**, e48 (2006).
4. G. P. Karev, Y. I. Wolf, E. V. Koonin, Simple stochastic birth and death models of genome evolution: Was there enough time for us to evolve? *Bioinformatics* **19**, 1889–1900 (2003).
5. E. V. Koonin, Y. I. Wolf, G. P. Karev, The structure of the protein universe and genome evolution. *Nature* **420**, 218–223 (2002).
6. B. Z. Harris, W. A. Lim, Mechanism and role of PDZ domains in signaling complex assembly. *J. Cell Sci.* **114**, 3219–3231 (2001).
7. E. Kim, M. Sheng, PDZ domain proteins of synapses. *Nat. Rev. Neurosci.* **5**, 771–781 (2004).
8. M. Sheng, C. Sala, PDZ domains and the organization of supramolecular complexes. *Annu. Rev. Neurosci.* **24**, 1–29 (2001).
9. 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, S. S. Sidhu, A specificity map for the PDZ domain family. *PLoS Biol.* **6**, e239 (2008).
10. M. A. Stiffler, J. R. Chen, V. P. Grantcharova, Y. Lei, D. Fuchs, J. E. Allen, L. A. Zaslavskaia, G. MacBeath, PDZ domain binding selectivity is optimized across the mouse proteome. *Science* **317**, 364–369 (2007).
11. N. J. Skelton, M. F. T. Koehler, K. Zobel, W. L. Wong, S. Yeh, M. T. Pisabarro, J. P. Yin, L. A. Lasky, S. S. Sidhu, Origins of PDZ domain ligand specificity. Structure determination and mutagenesis of the Erbin PDZ domain. *J. Biol. Chem.* **278**, 7645–7654 (2003).
12. P. Kristensen, G. Winter, Proteolytic selection for protein folding using filamentous bacteriophages. *Fold. Des.* **3**, 321–328 (1998).
13. S. Jung, A. Honegger, A. Plückthun, Selection for improved protein stability by phage display. *J. Mol. Biol.* **294**, 163–180 (1999).
14. C. J. Bond, C. Wiesmann, J. C. Marsters Jr., S. S. Sidhu, A structure-based database of antibody variable domain diversity. *J. Mol. Biol.* **348**, 699–709 (2005).
15. G. Pál, J.-L. K. Kouadio, D. R. Artis, A. A. Kossiakoff, S. S. Sidhu, Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. *J. Biol. Chem.* **281**, 22378–22385 (2006).
16. F. F. Vajdos, C. W. Adams, T. N. Breece, L. G. Presta, A. M. de Vos, S. S. Sidhu, Comprehensive functional maps of the antigen-binding site of an anti-ErbB2 antibody obtained with shotgun scanning mutagenesis. *J. Mol. Biol.* **320**, 415–428 (2002).
17. W. P. Russ, D. M. Lowery, P. Mishra, M. B. Yaffe, R. Ranganathan, Natural-like function in artificial WW domains. *Nature* **437**, 579–583 (2005).
18. M. Socolich, S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner, R. Ranganathan, Evolutionary information for specifying a protein fold. *Nature* **437**, 512–518 (2005).
19. P. Beltrao, L. Serrano, Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comput. Biol.* **3**, e25 (2007).
20. S. Bershtein, K. Goldin, D. S. Tawfik, Intense neutral drifts yield robust and evolvable consensus proteins. *J. Mol. Biol.* **379**, 1029–1044 (2008).
21. G. Amitai, R. D. Gupta, D. S. Tawfik, Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP J.* **1**, 67–78 (2007).
22. R. Wroe, H. S. Chan, E. Bornberg-Bauer, A structural model of latent evolutionary potentials underlying neutral networks in proteins. *HFSP J.* **1**, 79–87 (2007).
23. H. H. Guo, J. Choe, L. A. Loeb, Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 9205–9210 (2004).
24. N. Tokuriki, F. Stricher, L. Serrano, D. S. Tawfik, How protein stability and new functions trade off. *PLoS Comput. Biol.* **4**, e1000002 (2008).
25. S. S. Sidhu, H. B. Lowman, B. C. Cunningham, J. A. Wells, Phage display for selection of novel binding peptides, in *Applications of Chimeric Genes and Hybrid Proteins* (Academic Press Inc., San Diego, 2000), Pt C, pp. 333–363.
26. Abbreviations for the amino acids are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
27. R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H.-R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, A. Bateman, The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288 (2008).
28. R. Tonikian, Y. Zhang, C. Boone, S. S. Sidhu, Identifying specificity profiles for peptide recognition modules from phage-displayed peptide libraries. *Nat. Protoc.* **2**, 1368–1386 (2007).
29. K. H. Pearce Jr., B. J. Potts, L. G. Presta, L. N. Bald, B. M. Fendly, J. A. Wells, Mutational analysis of thrombopoietin for identification of receptor and neutralizing antibody sites. *J. Biol. Chem.* **272**, 20595–20602 (1997).
30. T. D. Schneider, R. M. Stephens, Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).
31. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
32. G. Fuh, M. T. Pisabarro, Y. Li, C. Quan, L. A. Lasky, S. S. Sidhu, Analysis of PDZ domain-ligand interactions using carboxyl-terminal phage display. *J. Biol. Chem.* **275**, 21486–21491 (2000).
33. 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, S. S. Sidhu, Convergent and divergent ligand specificity among PDZ domains of the LAP and zonula occludens (ZO) families. *J. Biol. Chem.* **281**, 22299–22311 (2006).
34. S. T. Runyon, Y. Zhang, B. A. Appleton, S. L. Sazinsky, P. Wu, B. Pan, C. Wiesmann, N. J. Skelton, S. S. Sidhu, Structural and functional analysis of the PDZ domains of human HtrA1 and HtrA3. *Protein Sci.* **16**, 2454–2471 (2007).
35. T. Balla, Inositol-lipid binding motifs: Signal integrators through protein-lipid and protein-protein interactions. *J. Cell Sci.* **118**, 2093–2104 (2005).
36. B. J. Hillier, K. S. Christopherson, K. E. Prehoda, D. S. Bredt, W. A. Lim, Unexpected modes of PDZ domain scaffolding revealed by structure of nNOS-syntrophin complex. *Science* **284**, 812–815 (1999).
37. G.D.B. acknowledges funding from the Canadian Institutes of Health Research (MOP-84324).

Submitted 7 May 2009
Accepted 20 August 2009
Final Publication 8 September 2009
10.1126/scisignal.2000416
Citation: A. Ernst, S. L. Sazinsky, S. Hui, B. Currell, M. Dharsee, S. Seshagiri, G. D. Bader, S. S. Sidhu, Rapid evolution of functional complexity in a domain family. *Sci. Signal.* **2**, ra50 (2009).