Evolution of domain–peptide interactions to coadapt specificity and affinity to functional diversity

Abdellali Kelil[a], Emmanuel D. Levy[b], and Stephen W. Michnick[c,1]

[a]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada M5S 3E1; [b]Department of Structural Biology, Weizmann Institute of Science, Rehovot 7610001, Israel; and [c]Department of Biochemistry and Molecular Medicine, University of Montreal, Montreal, QC, Canada H3T 1J4

Edited by James A. Wells, University of California, San Francisco, CA, and approved May 6, 2016 (received for review September 16, 2015)

Evolution of complexity in eukaryotic proteomes has arisen, in part, through emergence of modular independently folded domains mediating protein interactions via binding to short linear peptides in proteins. Over 30 years, structural properties and sequence preferences of these peptides have been extensively characterized. Less successful, however, were efforts to establish relationships between physicochemical properties and functions of domain–peptide interactions. To our knowledge, we have devised the first strategy to exhaustively explore the binding specificity of protein domain–peptide interactions. We applied the strategy to SH3 domains to determine the properties of their binding peptides starting from various experimental data. The strategy identified the majority (~70%) of experimentally determined SH3 binding sites. We discovered mutual relationships among binding specificity, binding affinity, and structural properties and evolution of linear peptides. Remarkably, we found that these properties are also related to functional diversity, defined by depth of proteins within hierarchies of gene ontologies. Our results revealed that linear peptides evolved to coadapt specificity and affinity to functional diversity of domain–peptide interactions. Thus, domain–peptide interactions follow human-constructed gene ontologies, which suggest that our understanding of biological process hierarchies reflect the way chemical and thermodynamic properties of linear peptides and their interaction networks, in general, have evolved.

linear peptides | domain–peptide interactions | binding specificity | binding affinity | functional specificity

Many proteins, particularly in eukaryotes, are composed of modular protein architectures consisting of multiple independently folding domains (1). Specific domains such as SH3 and PDZ domains were repeatedly used throughout evolution in increasingly complex organisms to mediate protein–protein interactions involved in signal transduction and protein targeting (2–5). These domains are associated with a number of human diseases and are targets of virus and other pathogen virulence proteins (6). Functions of these domains include binding to sequence-specific peptides both among themselves and on other proteins. Such interactions can create enormous plasticity in complex signaling and regulatory networks on immediate to evolutionary timescales (7), and are often used for regulating the activities of proteins and the spatiotemporal organization of protein interaction networks (8, 9). However, at the cellular level, we still do not grasp why certain peptides in proteins bind to distinct domains with high specificity whereas others highly cross-react with a number of members of a family of domains, and also what is the relationship between specificity of binding and specificity of functions of domain–peptide interactions. Two extreme examples are peptides of the MAPKK protein Pbs2 (residues 92–106) (10) and the actin assembly protein Las17 (residues 306–336) (11), which both interact with the osmosensor protein Sho1 via its SH3 domain. At one extreme, Pbs2 binds to the Sho1 SH3 domain with absolute specificity, and, at the other extreme, Las17 binds to the Sho1 and the SH3 domains of several other proteins as well. Three crucial questions about the interactions of linear peptides with modular domains are as follows: First, what are the properties of linear peptides that determine their binding to an individual member vs. multiple members of a domain family? Second, how does the binding specificity of domain–peptide interactions relate to the functions they are involved in? That is, can we predict whether a domain–peptide interaction is involved in an individual vs. many functions based on the physical properties of a peptide and its binding specificity? Finally, how does the affinity of domain–peptide interactions relate to the specificity and the function of proteins?

To address these questions, we chose a model system for which among the largest sets of experimental information is available, the *Src* homology-3 (SH3) domain. SH3 domains are peptide recognition modules that mediate protein interactions involved in many cellular functions, including signal transduction and cytoskeletal organization (2, 3). SH3 domains typically recognize proline-rich PXXP peptides, where P is fixed as the amino acid proline and X represents any amino acid (12). SH3 domains fall generally into two classes recognizing peptides with fixed residues proline (P), lysine (K), and arginine (R), conforming, respectively, to [RK]XXPXXP and PXXPX[RK] (13). A number of alternative peptides have, however, been identified (14–25), such as the noncanonical peptides that bind to the Fus1 SH3 domain (15), and the structured β-sheet ubiquitin-like domain of Ubi4 that binds to the Sla1-3 SH3 domain (25). A number of in vitro methods have been applied to study preferences of SH3 domains for specific sequence patterns, i.e., motifs. These methods include protein microarrays (26), synthetic peptide arrays (27), and screening of phage-displayed peptide libraries against

Significance

Today, we understand how short linear peptides bind to distinct recognition domains. However, at the cellular level, we still do not grasp why certain binding peptides are highly specific, whereas others are highly cross-reactive, and also what is the relationship with the functions of domain–peptide interactions. We revealed remarkable relationships among binding peptides, between their binding specificity, binding affinity, structural properties, and evolution. Surprisingly, we found that these properties are also related to functional specificity of domain–peptide interactions. Our results suggest that the structural properties and sequences of binding peptides have coevolved to achieve the levels of binding specificity and binding affinity that are required for the different levels of functional specificity of domain–peptide interactions.

Author contributions: A.K., E.D.L., and S.W.M. designed research; A.K. performed research; A.K. contributed new reagents/analytic tools; A.K., E.D.L., and S.W.M. analyzed data; and A.K., E.D.L., and S.W.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: stephen.michnick@umontreal.ca.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1518469113/-/DCSupplemental.

individual SH3 domains (28). However, in vitro methods are limited in the number, sequence variation, and the length of linear peptides that can be explored, making it hard to fully study the binding specificity of linear peptides and the roles of flanking or structurally distant amino acids. Thus, we cannot be certain that the linear peptide sequences we observe to bind to a domain represent all possible recognizable peptides and consensus motifs, and therefore, we cannot determine the specificity of any given sequence for any family of protein domains.

To our knowledge, we have developed the first strategy to exhaustively enumerate all possible linear peptides and resulting consensus motifs within proteins that are known to bind to some family of binding domains (michnick.bcm.umontreal.ca/dalel/). Whereas other methods for finding motifs are designed to search for known motifs (e.g., Eukaryotic Linear Motif database) (29–32), or motifs with properties previously observed in linear binding sites (e.g., 3D structures, intrinsic disorder, sequence conservation, and solvent accessibility) (33–38), or motifs that are over-represented according to a statistical model (e.g., hidden Markov model, Gibbs sampling, and Nested sampling) (39–42), our method exhaustively enumerates all possible motifs, covering the entire space of peptide variations. Here, we determine the specificity of all of the enumerated motifs for a target peptide-binding domain. The results of these analyses then serve as the essential information needed to address the questions posed above.

To this end, we first determined properties of motifs with high binding specificity and compared them to known SH3 binding sites. We then compared the properties of all motifs at distinct levels of binding specificity. Finally, we explored the relationship between binding specificity, binding affinity, structural properties, and sequence conservation of known SH3 binding sites, including the functions of proteins in which the binding sites are found.

We show here that there are simple linear correlations among physical properties, binding affinity, and sequence conservation of motifs with their binding specificity to SH3 domains. Surprisingly, we discovered that all of these variables correlate with functional specificity as defined by the position of proteins that contain members of linear consensus motifs within the Gene Ontology (GO). We illustrate this relationship between functional and binding specificity with the example of a yeast osmosensory membrane protein (Sho1), which binds with absolute specificity to a linear peptide within its direct osmosensory signaling effector (Pbs2), but binds to a number of other proteins in a less specific manner.

Results

General Strategy for Exhaustive Search of Motifs. We collected experimentally validated interaction data between SH3 domains and proteins of the budding yeast *Saccharomyces cerevisiae*, and, in parallel, we carefully selected negative interactions for these SH3 domains (*Materials and Methods*). In total, we manually curated 890 domain–protein interactions from the literature, involving 24 SH3 domains and 361 proteins, encoding a total of 749 verified SH3 binding sites, each of which was shown to bind to one/multiple SH3 domains through two or more independent methods (henceforth, "known SH3 binding sites") (Datasets S1 and S2).

Our strategy for inferring binding specificity was based on the premise that proteins known to bind to a common target domain should be enriched for amino acid sequences that share particular patterns, i.e., motifs, that mediate binding specifically with that domain, whereas in other proteins these motifs should not exhibit such enrichment. Thus, the binding specificity of a family of linear peptides displaying a common motif could be scored by comparing the enrichment of that motif in binding proteins vs. nonbinding proteins. Here, considering a particular SH3 domain, we define three distinct sets of proteins, the "positives" are proteins known to bind the SH3 domain, the "negatives" are proteins that do not bind to the target domain but bind to other domains of the same family (i.e., SH3 domains), and the "background" is a large set of proteins that do not bind to the target domain nor to any other domain from the same family (i.e., the rest of the proteome). One goal of our strategy is to infer binding specificity of all potential motifs; thus, we exhaustively enumerate all of the motifs of variable length that are present in the positives, and we calculate two *P*-values for each motif: (*i*) P_{NEG} reflects motif enrichment in the positives relative to the negatives, and (*ii*) P_{BAK} scores motif enrichment in the positives relative to the background (Fig. 1 *C–F*). In other words, P_{NEG} aims to evaluate specificity of the motifs for the target domain relative to other domains of the same family, whereas P_{BAK} aims to evaluate binding specificity for the target domain relative to the motifs found in the background. Thus, for a motif mediating binding with a target domain, strong P_{NEG} and P_{BAK} means high specificity, whereas weak P_{NEG} means high cross-reactivity and weak P_{BAK} means high promiscuity (Fig. 1*C*). Exhaustive enumeration of motifs includes the enumeration of all possible combinations of amino acids in each position in the motifs. The goal is to find preferences for multiple amino acids at individual positions, e.g., [RK] in [RK]XXPXXP and PXXPX[RK] (13), and correlated preferences at distinct positions, e.g., [ST] in R[ST][ST]SL recognized by the Fus1 SH3 domain (15). This is important because binding specificity generally depends on the amino acid identities at distinct positions, which can be either independent or correlated (43).

For each SH3 domain, we filtered out proteins with over 95% sequence identity to avoid motif enrichment due to redundancy (Fig. 1*A*, *iii*). Then, motifs present in positives were enumerated and scored through a unique three-step strategy (Fig. 1*B*). In the first step, positives were exhaustively scanned for all possible motifs of 3–15 residues including any number and combination of wildcards (Fig. 1*B*). The rationale behind our choice of searching motifs of 3–15 aa long is based on the 749 experimentally characterized SH3 binding sites that we curated from the literature (Dataset S1), the length of which ranges mostly from 3 to 15. This enumeration thus covered the entire space of all possible motifs present in the positives, for the lengths considered. The two *P*-values, P_{NEG} and P_{BAK}, were computed for each motif to score its enrichment in the positives relative to the negatives and to the background, respectively, based on the cumulative hypergeometric distribution (Fig. 1 *C–F*). The goal of the first step was thus to find motifs and score their enrichment among groups of proteins known to bind to the target domain. The next steps aimed to refine the motifs, which involved searching possible variations of each motif by substitution of wildcards with combinations of amino acids, e.g., [IVL] or [DE]. However, searching all possible variations of the motifs is physically infeasible because the combinatorial space is too vast (*Materials and Methods*). For this reason, variations of each motif were searched iteratively, by substitution of all wildcards, one by one, with all combinations of amino acids, and only new motifs with better *P*-values than the original motif were retained and then refined in their turn, until no more motifs were retained. Therefore, in the second step, motifs obtained in the first step were refined by searching positions with preferences for multiple amino acids (Fig. 1*B*). Then, in the third step, motifs obtained in the second step were iteratively refined to find correlated preferences for amino acids at distinct positions (Fig. 1*B*). The final stage consisted of filtering out overlapping motifs by keeping those with best *P*-values, which considerably reduced the number of discovered motifs (Fig. 1*A*, *v*). After the normalization of P_{NEG} and P_{BAK} distributions (*Materials and Methods*), we assigned a single *P*-value to each motif, corresponding to the least significant of P_{NEG} and P_{BAK}.

Discovered Motifs with High Binding Specificity Overlap with Known SH3 Binding Sites. We started our experiments with the analysis of the motifs with high binding specificity. We assessed here their binding specificity by measuring their overlap with known SH3 binding sites (Dataset S1). For each SH3 domain, in each positive

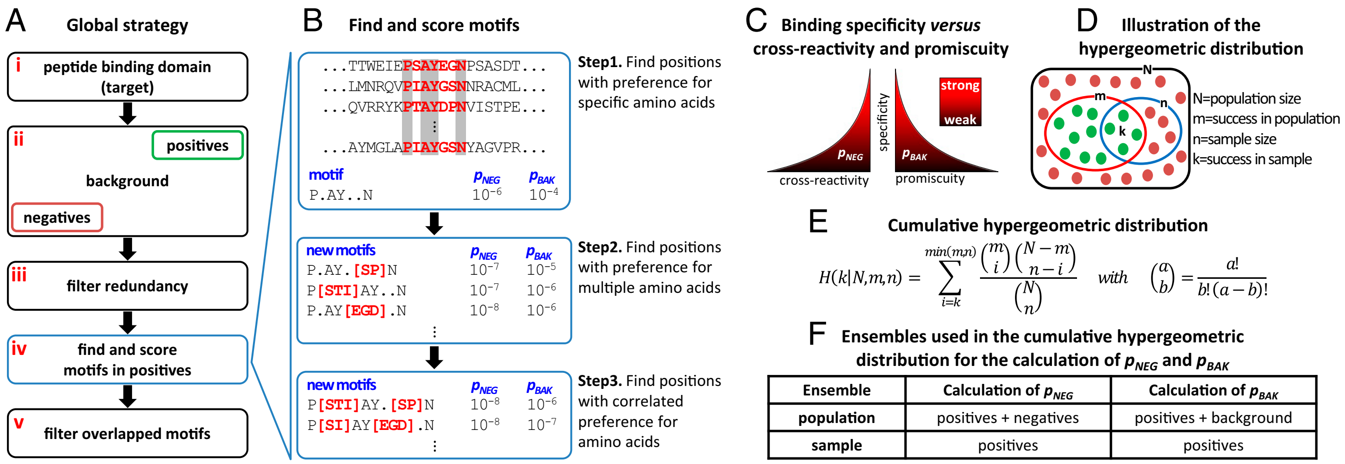

A
Global strategy
i
peptide binding domain (target)
ii
positives
background
negatives
iii
filter redundancy
iv
find and score motifs in positives
v
filter overlapped motifs
B
Find and score motifs
...TTWEIEPSAYEGNPSASDT...
...LMNRQVPIAYGSNNRACML...
...QVRRYKPTAYDPNVISTPE...
...AYMGLAPIAYGSNYAGVPR...
motif
p_{NEG}
p_{BAK}
P.AY..N
10^{-6}
10^{-4}
Step1. Find positions with preference for specific amino acids
new motifs
P.AY.[SP]N 10^{-7} 10^{-5}
P[STI]AY..N 10^{-7} 10^{-6}
P.AY[EGD].N 10^{-8} 10^{-6}
Step2. Find positions with preference for multiple amino acids
new motifs
P[STI]AY.[SP]N 10^{-8} 10^{-6}
P[SI]AY[EGD].N 10^{-8} 10^{-7}
Step3. Find positions with correlated preference for amino acids
C
Binding specificity *versus* cross-reactivity and promiscuity
specificity
strong
weak
p_{NEG}
p_{BAK}
cross-reactivity
promiscuity
D
Illustration of the hypergeometric distribution
N
m
n
k
N=population size
m=success in population
n=sample size
k=success in sample
E
Cumulative hypergeometric distribution
$H(k|N,m,n) = \sum_{i=k}^{\min(m,n)} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$ with $\binom{a}{b} = \frac{a!}{b!(a-b)!}$
F
Ensembles used in the cumulative hypergeometric distribution for the calculation of p_{NEG} and p_{BAK}
Ensemble | Calculation of p_{NEG} | Calculation of p_{BAK}
population | positives + negatives | positives + background
sample | positives | positives

Fig. 1. A general strategy for exhaustive search of amino acid (aa) sequence motifs in domain binding proteins. (*A*) Global strategy. (*i*) We start from a target domain; (*ii*) we then select the positives from available experimental information (i.e., proteins that bind to the target domain), as well as the negatives (i.e., proteins that do not bind to the target domain but that bind to other domains from the same family), and the background (i.e., proteins that do not bind to the target domain nor to any other domain from the same family); (*iii*) we filter out redundant proteins in positives, negatives, and background; (*iv*) we enumerate motifs present in positives and score their enrichment relative to both negatives and background; (*v*) we retain motifs with best scores (*Supporting Information*). (*B*) Motifs discovery strategy. Step 1: Find positions with preference for specific amino acids: we scan positives for all possible motifs of 3–15 residues including wildcards. We then score motif enrichment over the negatives and the background, with P_{NEG} and P_{BAK}. Step 2: Find positions with preference for multiple amino acids: we find new motifs by substituting all combinations of amino acids into each wildcard of motifs obtained in step 1, and retain only those with better *P*-values than the original motifs. Step 3: Find multiple positions with correlated preference for amino acids: we repeat step 2 on the last previously retained motifs until no new motif with a better *P*-value is obtained. (*C*) Binding specificity vs. cross-reactivity and promiscuity: The motifs with the most significant P_{NEG} and P_{BAK} are the most specific. We define enriched motifs in the positives with weak P_{NEG} as cross-reactive (enriched in the positives and the negatives), and those with weak P_{BAK} are promiscuous (enriched in the positives and the proteome). (*D*) Illustration of the cumulative hypergeometric distribution. The cumulative hypergeometric distribution describes the probability to see at least *k* successes in a sample of size *n* picked from a finite population of size *N* containing *m* successes. (*E*) Cumulative hypergeometric distribution. The lower the probability, the higher is the overrepresentation of the successes in the sample in comparison with the population. (*F*) Ensembles used in the cumulative hypergeometric distribution for the calculation of P_{NEG} and P_{BAK}. Two *P*-values are calculated for each motif: P_{NEG} scores motif enrichment in positives over negatives, and P_{BAK} scores motif enrichment in positives over background. For P_{NEG}, the population is the union of positives and negatives, whereas for P_{BAK}, the population is the union of positives and background. However, for both P_{NEG} and P_{BAK}, the sample is the positives.

sequence, we selected motifs with the highest binding specificity by increasing *P*-value cutoff until the covered length was comparable (i.e., equal or below) to that of known SH3 binding sites (Dataset S3). Henceforth, we refer to these motifs as the predicted motifs and their instances in the positives as the predicted SH3 binding sites (Fig. 2). We found that more than 80% of the motifs we predicted overlap or are within a distance of 10 aa from a known binding site. In addition, more than 70% of the amino acid sequences in the predicted SH3 binding sites were within experimentally known SH3 binding sites (Fig. 2*A* and Fig. S1). This is highly unlikely because known SH3 binding sites covered 2.93% of the total length of positives sequences, whereas predicted SH3 binding sites covered only 2.57%. This means that selected motifs with the highest binding specificity cover (2.57% × 70%)/2.93% ~ 60% of the total length of all known binding sites. The probability to find such overlap by chance within the 361 positive sequences is less than 10^{-100} (*Materials and Methods*). The quality of these results thus demonstrate that scoring motif enrichment in binding proteins of a target domain relative to nonbinding proteins enables us to discriminate between motifs that mediate binding to the target domain vs. random motifs. These results also confirm the supposition that SH3 binding sites have distinct sequences compared with the rest of the sequences of the proteins where they are located and are also rare or absent in the rest of the proteome. The ~20% of motifs that were not within previously determined SH3 binding sites also did not have amino acid sequences consistent with the canonical PXXP motif (Dataset S1).

Predicted SH3 Binding Sites Have Distinct Structural and Evolutionary Properties from Flanking Sequences. We next compared the three key structural properties of predicted SH3 binding sites to those of their flanking sequences (Fig. 2*B*), including binding energy, solvent accessibility, intrinsic disorder, and sequence conservation (*Materials and Methods*). We found that the amino acid sequences covered by predicted SH3 binding sites are highly conserved compared with their flanking sequences, and are localized within intrinsically disordered regions in proteins (Fig. 2*B*). Furthermore, we observed a striking contrast between structural and evolutionary properties of predicted SH3 binding sites compared with their flanking regions, suggesting functional importance (44). In cases where structures of SH3 protein binding partners were available, the SH3 binding sites we predicted were indeed in conserved yet unstructured and solvent-exposed loops (Fig. 2*C*). Flanking regions were usually also unstructured, but the other structural properties were distinct (Fig. 2*B*). We also predicted stark exceptions, such as the unconventional SH3 binding site surrounding the Ile44 in the compact and globular ubiquitin Ubi4 that binds to the SH3 domain Sla1-3 (Fig. S2). The ubiquitin SH3 binding site has little in common with conventional SH3 binding sites; it forms a structured β-sheet and does not carry a PXXP motif, yet it binds to the same hydrophobic groove on SH3 domains as PXXP binding peptides (25).

Consensus Residues in Predicted SH3 Binding Sites Exhibit Expected Structural and Evolutionary Properties. As with most peptide binding domains, only few (approximately one-third) hot-spot residues in peptides are required for binding to SH3 domains (45, 46). Basically, these residues identified as non-X sites (henceforth, “consensus residues”) have specific amino acid identities, e.g., P and [RK] in [RK]XXPXXP and PXXPX[RK], and mutation of these residues disrupts binding, whereas wildcard positions identified as X sites (henceforth, “nonconsensus residues”) can be mutated without altering binding (47). A key test for our strategy was to determine whether it is capable of identifying the consensus residues and

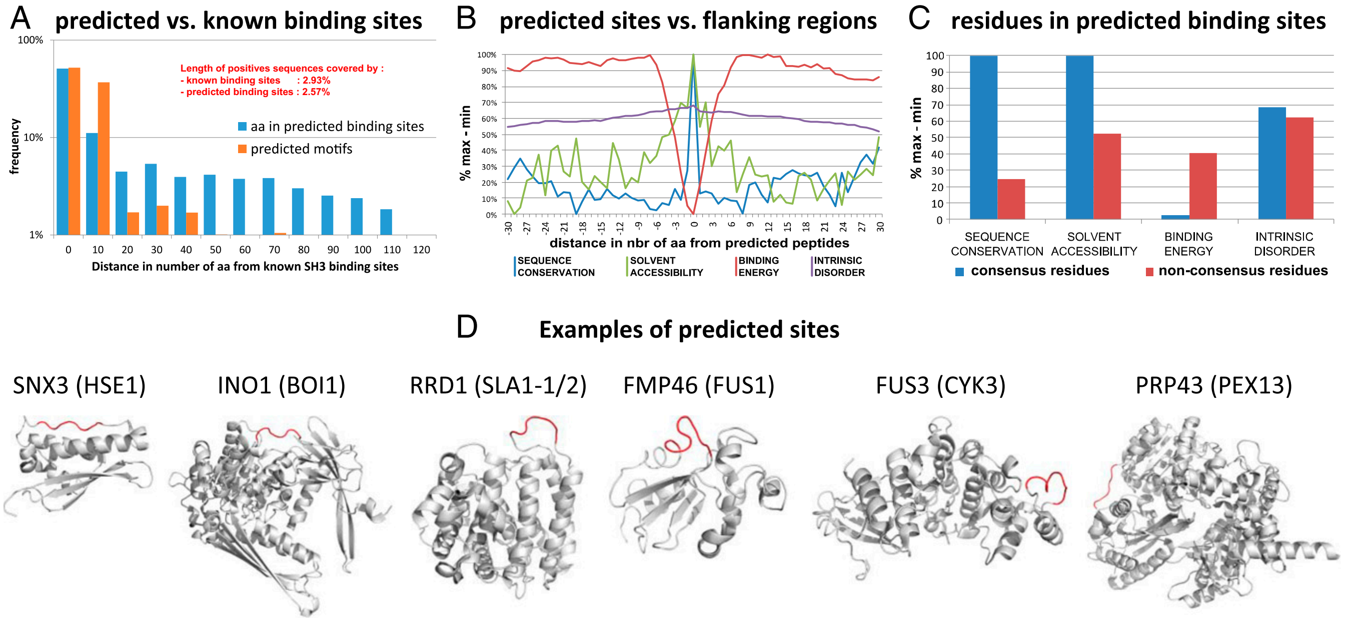


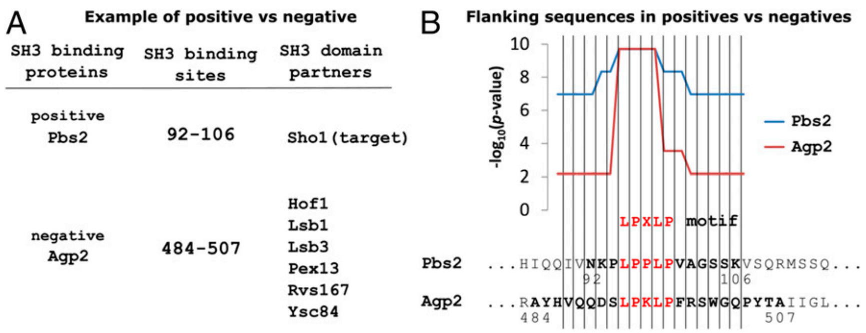
Fig. 2. Properties of predicted SH3 binding sites. (*A*) The majority of predicted SH3 binding sites overlap with known SH3 binding sites. The overlap between predicted and known SH3 binding sites was measured by the distribution of their distances. The figure summarizes the frequency of amino acids in predicted binding sites (blue bars) and predicted motifs (orange bars) at different distances from known binding sites; the *x* axis represents the distance in number of amino acids from the nearest known binding site, zero meaning inside; the *y* axis represents the frequency of predicted motifs (orange bars) and the frequency of amino acids belonging to predicted binding sites (blue bars); the length (in percentage) of positive sequences covered by predicted and known binding sites are indicated; results for individual SH3 domains are provided in *Supporting Information*. (*B*) Structural properties and conservation of predicted SH3 binding sites are distinct from their flanking sequences: The *x* axis shows the distance in number of amino acids from predicted binding sites. All amino acids in predicted binding sites have a distance equal to zero. The positive and negative distances represent C- and N-terminal sides of the predicted binding sites respectively. The *y* axis shows the mean value of each property at different distances from predicted binding sites relative to minimum and maximum values. (*C*) Structural properties and conservation of consensus residues are different from those of nonconsensus residues in predicted SH3 binding sites: the difference in the properties of consensus residues, e.g., P sites in PXXP, and nonconsensus residues, e.g., X sites in PXXP, in predicted SH3 binding sites. (*D*) Flanking sequences and predicted SH3 binding sites are both intrinsically disordered: visualization of predicted motifs on available structures. Partners of specific SH3 domains are shown, and their standard name is in parentheses. The binding site is highlighted in red. The PDB codes for the structures illustrated here are from left to right, 1OCU, 1JKI, 2IXP, 1WPI, 2B9I, and 3KX2.

their identities of amino acids that determine the binding specificity to individual SH3 domains. To this end, we compared the three key structural properties and positional amino acid conservation in predicted SH3 binding sites for both consensus and nonconsensus residues (Fig. 2*C*). Except intrinsic disorder, we observed differences; with conservation, solvent accessibility, and energetic contribution showing higher values for consensus residues. This result highlights that residues with defined amino acid identities in the predicted SH3 binding sites are important for binding, and their properties are consistent with the key sites that form favorable interactions with SH3 domain binding grooves (13, 48).

Flanking Sequences Have Determinant Role in Positive/Negative Binding Selectivity. We compared the instances of enriched motifs in the positives and negatives. As expected, instances of the motifs in the positives were significantly more conserved, more intrinsically disordered, and showed higher solvent accessibilities than instances in the negatives (Fig. S3). We found that there was a sharp distinction between the properties of the flanking sequences of the motifs in the positives and the negatives. We found that the binding specificity of the flanking sequences in the positives were higher than those in the negatives, which means that, even if the flanking sequences contain motifs less enriched than those we selected, they are nevertheless significantly enriched compared with the flanking sequences in the negatives (Fig. S3). This contrast between the flanking sequences in the positives and negatives suggests an unexpected role of the flanking sequences in binding selectivity (Fig. S3).

To understand this role, we compared two experimentally determined SH3 binding sites, the peptide “NKPLPPLPVAGSSKV” in Pbs2 (residues 92–106) that binds to the Sho1 SH3 domain, and the peptide “AYHVQQDSLPKLPFRSWGQPYTA” in Agp2 (residues 484–507) that does not bind to the Sho1 SH3 domain but does bind to other SH3 domains (Fig. 3*A*). Both peptides encode the motif “LPXLP” that we predicted to have high binding specificity (*P*-value of 10^{-10}) for the Sho1 SH3 domain (Fig. 3*B*). Thus, if this motif is mediating binding to Sho1 SH3 domain, why does its presence in Agp2 not result in its binding to the Sho1 SH3 domain? To answer this question, we compared the flanking sequences of LPXLP motifs in both peptides (Fig. 3*B*). We found, as expected, that the motif LPXLP in Pbs2 has higher binding specificity than its flanking sequences. However, the flanking sequences of LPXLP in Pbs2 have significantly higher binding specificity than the flanking sequences of LPXLP in Agp2 (Fig. 3*B*). This result illustrates how the flanking sequences can play a role in the definition of the positive/negative binding selection of peptides in the proteome for the Sho1 SH3 binding domain. These results are also consistent with experimental evidence that amino acid substitution both within or flanking the motif could decrease the specificity without disrupting and in some cases enhancing the binding of Pbs2 to Sho1 (10).

Specificity of Discovered Motifs Correlates with Structural Properties and Conservation. In the sections above, we analyzed a tiny fraction of the motifs we discovered, those with the best *P*-values that cover comparable length in the positives to known SH3 binding sites, and we showed the distinct structural and evolutionary



A Example of positive vs negative

SH3 binding proteins	SH3 binding sites	SH3 domain partners
positive Pbs2	92-106	Sho1 (target)
negative Agp2	484-507	Hof1 Lsb1 Lsb3 Pex13 Rvs167 Ysc84

Fig. 3. Binding specificity for the Sho1 SH3 domain of flanking sequences in SH3 binding sites of Pbs2 and Agp2. Comparison of the binding specificity for the Sho1 SH3 domain of the flanking sequences of the motif LPXLP in the experimentally determined SH3 binding sites in Pbs2 and Agp2. (*A*) Both Pbs2 and Agp2 encode a SH3 binding peptide. The binding peptide in Pbs2 (residues 92–106) binds to the Sho1 SH3 domain. The binding peptide in Agp2 (residues 484–507) does not bind to the Sho1 SH3 domain but does bind to other SH3 domains. As defined in our strategy, Pbs2 is a positive for the Sho1 SH3 domain and Agp2 is a negative. (*B*) The motif LPXLP is among those we predicted with the highest binding specificity for the Sho1 SH3 domain (Dataset S3). The presence of LPXLP in the binding peptide of Pbs2 is expected, because it binds to the Sho1 SH3 domain. However, the presence of LPXLP in the binding peptide of Agp2 is not expected, because it does not bind to the Sho1 SH3 domain. The high contrast between the binding specificity in the flanking sequences of the motif LPXLP in both proteins suggests that the flanking sequences are playing an important role to promote binding of Pbs2 to the Sho1 SH3 domain and inhibit binding of Agp2 to the Sho1 SH3 domain.

properties of their instances in the positives. In this section, we describe analyses of all of the motifs we discovered from highest to lowest binding specificity. To this end, we compared the structural and evolutionary properties for the instances within the positives of discovered motifs selected at different binding specificities, i.e., *P*-value cutoffs, to values calculated for known SH3 binding sites (Fig. 4). We found a strong correlation between each property and binding specificity for our discovered motifs, both on average (Fig. 4) and for individual SH3 domains (Figs. S4–S7). All of the properties of stringently selected motifs (i.e., with strongest *P*-values) were similar to those of known SH3 binding sites (48). The correlations observed are consistent with chemical intuition in a manner that has not, to our knowledge, been previously described: the binding specificity of the instance of discovered motifs for individual SH3 domains is correlated with their structural properties and sequence conservation (Fig. 4). This suggests that the binding specificity of SH3 binding sites is a continuous function of their structural and evolutionary properties. The binding specificity captures the degree to which an SH3 domain will bind to proteins and not to the rest of the proteome or all observable negative binders. This implies that an SH3 domain could bind to any potentially compatible peptide depending on such parameters as protein abundance and localization (49, 50). An optimally specific binding site would be one evolved to bind with

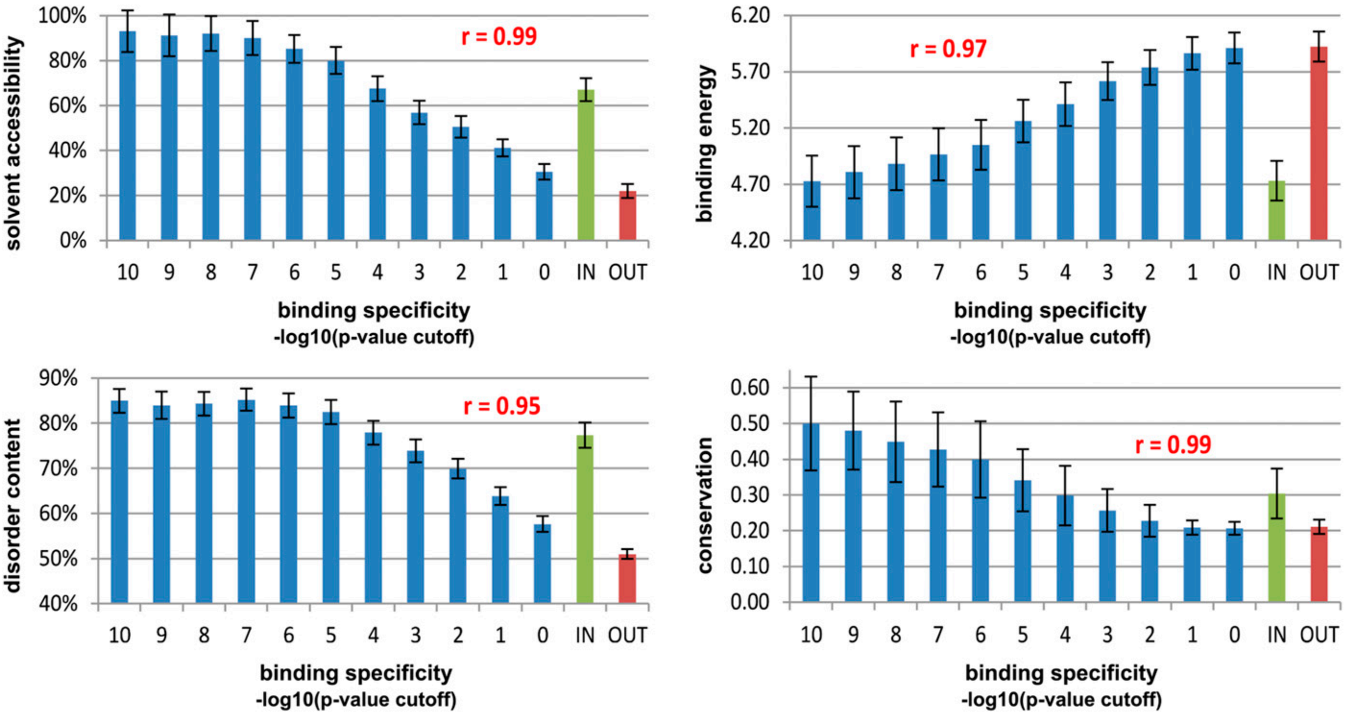


Fig. 4. Specificity vs. structural and evolutionary properties of discovered motifs compared with known SH3 binding sites. The figure summarizes four key properties (solvent accessibility, binding energy, intrinsic disorder, and sequence conservation) of known SH3 binding sites (green bar), of discovered motifs ranked according to their *P*-values (blue bars), and of sequences of all the amino acids outside the known SH3 binding sites (red bar). The correlation (*r* value) between each property and *P*-value cutoffs is indicated. The error bars correspond to the SEM.

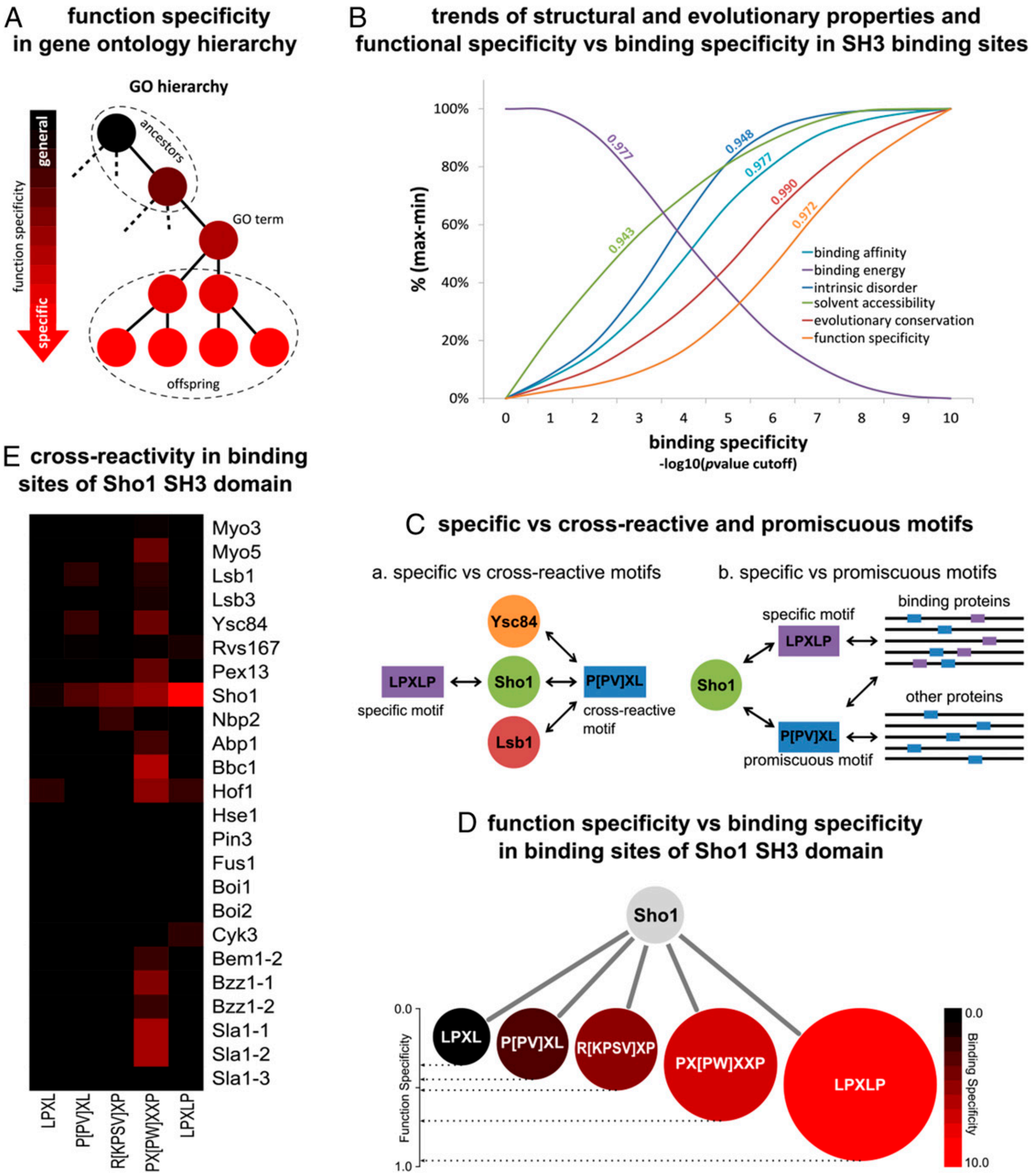


Fig. 5. Evolution of SH3 binding sites to coadapt specificity and affinity to functional specificity of proteins. (*A*) Functional specificity in the GO hierarchy. The GO hierarchy, represented by a tree in which nodes correspond to ontology terms and edges correspond to parent–child term relations, is organized so that ontology terms close to the top of the tree represent general functions, whereas those close to the bottom represent specific functions. (*B*) Binding specificity is correlated with structural and evolutionary properties, affinity, and functional specificity. For different *P*-value cutoffs (*x* axis), different properties were calculated for each of the 749 SH3 binding sites that we curated from the literature (*y* axis): evolutionary conservation, binding affinity, intrinsic disorder, solvent accessibility, binding energy, and functional specificity (*Materials and Methods*). For each known SH3 binding site, its *P*-value was obtained from its matching discovered motif with the best *P*-value. The trend of each property was obtained by regression with multiorder (max order, 3) polynomial functions. (*C*) Distinctions between specific vs. cross-reactive and promiscuous motifs. (*a*) The motif P[PV]XL (blue) cross-reacts with the SH3 domains of three proteins: Sho1, Ysc84, and Lsb1. In contrast, the motif LPXLP (purple) is interacting only with Sho1; the purple motif is specific, and the blue motif is cross-reactive. (*b*) Both motifs bind to the Sho1 SH3 domain and are enriched in proteins binding to the Sho1 SH3 domain, but only the blue motif is enriched in proteins that do not bind to the Sho1 SH3 domain; the purple motif is thus specific, and the blue one is promiscuous. (*D*) Functional specificity vs. binding specificity among binding sites for the Sho1 SH3 domain. For each experimentally determined binding site for the Sho1 SH3 domain, we determined a matching motif and computed a functional specificity score for proteins that encode the motif (*Materials and Methods*). The figure illustrates the relationship between functional specificity (size of round forms) vs. binding specificity (black to red gradient) for binding sites of the Sho1 SH3 domain. The Sho1 SH3 domain recognizes five motifs with different levels of binding specificity, and these motifs exhibit different levels of functional specificity. The level of binding specificity is increasing together with the level of functional specificity. (*E*) Cross-reactivity of binding sites for the Sho1 domain with other SH3 domain. Heat map shows binding specificity for specific pairs of SH3 domain and binding motifs. The motifs LPXL, P[PV]XL, and R[KPSV]XP cross-react with different SH3 domains with equivalent low-to-moderate specificity. The motif PX[PW]XXP has high affinity for Sho1 but cross-reacts with other SH3 domains. Only LPXLP exhibits both high specificity and high affinity to Sho1.

maximum possible affinity to one SH3 domain while minimizing interactions with all other SH3 domains (10).

SH3 Binding Sites Evolved to Coadapt Affinity and Cross-Reactivity to Functional Diversity. Over two decades of research on SH3 domains have resulted in detailed understanding of how they recognize linear peptides with distinct binding specificities, and also decrypt structural and sequence properties of these peptides (48). However, at the cellular level, we still do not grasp why the properties of certain peptides were optimized to be highly specific, whereas others exhibit high cross-reactivity. To answer this question, we explored the relationship between binding specificity, binding affinity, and structural and evolutionary properties of 749 experimentally characterized SH3 binding sites that we curated from the literature (Dataset S1). We also related these properties to the functional specificity of the proteins in which the sites were identified (Fig. 5). The functional specificity of each protein was measured by the relative depth of its associated GO terms in the GO hierarchy. The GO hierarchy is indeed organized so that GO terms close to the root represent broad functions such as "catalysis," whereas GO terms close to the leaves represent specific functions such as "cAMP-dependent kinase" (51) (Fig. 5*A* and *Materials and Methods*).

As with the predicted motifs, we observed significant correlations among known SH3 binding sites, between their binding specificity and both structural properties and evolution (Fig. 5*B*). Surprisingly, both binding affinity and functional specificity also correlated with binding specificity. These trends suggested that highly conserved SH3 binding sites have been optimized and retained throughout evolution to bind to SH3 domains with high specificity and affinity, and they are also involved in highly specific functions. These binding sites correspond to motifs enriched in the positives and rare in the negatives and in the rest of the proteome. We thus find that unwanted (i.e., nonfunctional) cross-reactive and promiscuous interactions for the proteins involved in the most specific functions are minimized by directing them to their relevant biological targets and preventing aberrant interactions with nonphysiological targets. Other SH3 binding sites have emerged with less constrained structural properties and binding specificity, allowing them to bind to a broader range of partners involved in more general functions, and easier exchange between them because of their relatively lower affinity to any individual SH3 domain. We call these cases "cross-reactive" motifs and their signatures are weak, i.e., strong P_{NEG} but weak P_{BAK} values. The least conserved binding sites correspond to motifs with weak *P*-values (i.e., both P_{NEG} and P_{BAK}), which means they are not enriched in the positives relative to the negatives and the background; in fact, they are also frequent in the negatives and the rest of the proteome; we refer to these motifs as "promiscuous" (Fig. 5*C*). These correlations suggest that the structural properties and sequences of SH3 binding sites have coevolved to achieve the levels of binding specificity and binding affinity that are required for the different levels of functional specificity/diversity of the proteins where they are located.

Example of the Peptide in MAPKK Pbs2 Binding to the SH3 Domain of Osmosensor Protein Sho1. An example of a highly specific SH3 binding site is found in the Pbs2 protein that acts both as a scaffold and a MAP kinase kinase; Pbs2 is an essential component of the osmotic stress signal transduction response pathway in yeast. Pbs2 encodes the peptide NKPLPPLPVAGSSKV (residues 92–106) that binds with absolute specificity to the osmosensor protein Sho1 SH3 domain (10), which is known to interact with peptides in 36 different proteins (Dataset S4). The Pbs2 peptide matches the LPXLP motif for which we infer a high binding specificity for the Sho1 SH3 domain, reflected in the *P*-value of ~10^{-10}. We found 14 other proteins among the binding partners of the Sho1 SH3 domain that contain peptides matching the motif LPXLP, and these peptides also bind to the Sho1 SH3 domain with absolute (or near) specificity (Fig. 5*D* and Dataset S3). Among the peptides known to bind to the Sho1 SH3 domain, a number of them correspond to other motifs (Dataset S3). We examined the relationship between binding specificity and both functional specificity and cross-reactivity for all of these cases (Fig. 5 *D* and *E*). The relationship between binding specificity and both structural and evolutionary properties, including binding affinity and functional specificity, are available in Fig. S8. The results revealed that binding sites of the Sho1 SH3 domain fall into the motifs "LPXL" (*P*-value ~ 10^{-1}), "P[PV]XL" (*P*-value ~ 10^{-2}), "R[KPSV]XP" (*P*-value ~ 10^{-3}), and "PX[PW]XXP" (*P*-value ~ 10^{-4}), and bind to the Sho1 SH3 domain with increasing specificity, but all have lower specificity than LPXLP (*P*-value ~ 10^{-10}) (Fig. 5*D*). Importantly, the binding specificity of these peptides for the Sho1 SH3 domain increases with the functional specificity of the proteins in which they were identified (Fig. 5*D*). In addition, these peptides cross-react with other SH3 domains with low-to-moderate specificity and are involved in diverse functions. The cross-reactivity is most visible for motif PX[PW]XXP, which binds to the Sho1 SH3 domain with moderate specificity and to other SH3 domains with higher specificity; for instance, those of proteins Bbc1, Sla1-1, Sla1-2, and Bzz1-1 involved in actin cytoskeletal dynamics, a process that affects or is affected by many other cellular processes, in part, through many alternative cross-reactive interactions (52). The results revealed also that peptides belonging to the LPXLP motif are involved in specific functions and exhibit minimal cross talk with other SH3 domains (Fig. 5*E*). A typical example is the high-osmolarity glycerol (HOG) pathway in which the interaction of Sho1 via its SH3 domain to Pbs2 is critical for pathway activation (53). This result illustrates the ability of our strategy to distinguish between the different levels of binding specificity of linear peptides for their binding domains.

Marles et al. (53) have demonstrated a strong correlation between the binding affinity of the interaction of the Sho1 SH3 domain with its binding site in Pbs2 and the quantitative in vivo outputs from the HOG high-osmolarity response pathway controlled by Sho1. In addition, they found that reduction in binding affinity in Sho1–Pbs2 interaction within this pathway causes aberrant cross-talk activation of the mating response. Moreover, they found that reducing binding affinity causes proportional increase in misactivation of the mating pheromone response pathway. These findings confirm the importance of the relationship we have established between the level of binding specificity, binding affinity, and functional specificity of the Sho1–Pbs2 SH3 domain–peptide interaction. In contrast, the results obtained by Zarrinpar et al. (10) showed that increasing binding affinity in the interaction of the Sho1–Pbs2 SH3 domain–peptide interaction increases cross-reactivity of Pbs2 with other SH3 domains, causing a fitness defect in strains expressing the higher affinity mutants. Thus, to assure maximum functional specificity, a peptide sequence may evolve only to a maximum affinity that also assures maximum specificity.

We found other cases of SH3 domain–peptide interactions with absolute specificity. For instance, the Lsb1 SH3 domain interacts with 95 distinct binding sites spread over 83 different proteins (Dataset S1), among which the binding site in Las17 protein binds to the Lsb1 SH3 domain with absolute specificity, whereas the other binding sites exhibit variable cross-reactivity with other SH3 domains.

Discussion

The strategy we presented here provides a general way to identify binding sites for any protein domain based solely on protein–protein interaction data where the baits are individual domains screened against the proteome. When applied to the network of SH3 domain–ligand interactions in yeast, we showed that our strategy could be used to predict known and uncharacterized motifs. The latter may bind directly to SH3 domains, may be extensions of a known binding peptide in flanking regions, or be in a structurally distinct region of the protein that modulates the peptide–SH3 domain interaction (54). Thus, our approach paves the way to expanding the known repertoires of protein domain–peptide interactions and their regulation, in an unbiased way and based solely on protein–protein interaction data and amino acid sequences.

Our study revealed remarkably simple relationships among the structural properties of binding sites, thermodynamics of domain–peptide interactions, and binding and functional specificity. First, we saw that binding affinity correlates to binding and functional specificity. This result implies that binding of domains to motifs follow a continuum in which proteins involved in general functions have lower affinity and therefore most readily exchange with their domain binding partners. In contrast, at high affinity, motifs exhibit specific structural properties and high binding specificity, allowing proteins involved in the most specific functions to bind with high affinity to their cognate binding partners. This enables minimizing unwanted cross-reactive and promiscuous interactions, by directing proteins to their relevant biological targets and preventing aberrant interactions with nonphysiological targets.

The fact that binding affinity and specificity, and structural and evolutionary properties correlate with GO hierarchy suggests a deep relationship between the thermodynamics of protein binding and functional specificity that is strikingly reflected in a human conception of the organization of biological processes (55). Recently, Dutkowski et al. (56) demonstrated that existing and unforeseen GO hierarchies can be derived based on analyses of protein–protein and genetic interaction networks. Our results raise the mirror idea that GO hierarchies reflect biophysical properties of protein interaction networks.

Materials and Methods

Benchmark. For the purpose of this study, we integrated a benchmark of a number of experimental results from different studies. To this end, we manually curated the literature for 890 domain–protein interactions in budding yeast, *Saccharomyces cerevisiae*, between 24 SH3 domains and 361 proteins, including 749 binding sites, each of which was identified to be recognized by one or multiple SH3 domains, and supported by multiple experiments. Among selected studies, the one presented by Tong et al. (11), pioneering in combining experimental and computational methods at large scale, is among those with the most impact. We also integrated the high-confidence SH3 domain interaction network obtained by Tonikian et al. (27), which is to date the largest contribution to the SH3 domain interactions network in *S. cerevisiae*. The complete set of positives is available in Dataset S1.

A key issue in this work was the choice of the negatives to be considered with each SH3 domain. The challenge was to avoid negatives (nonbinding proteins) that are actually positives (binding proteins) but had been misidentified in experimental studies, i.e., false negatives. This misidentification risks distorting motifs overrepresentation in positives relative to negatives. To reduce this risk, we compiled for each SH3 domain all negatives obtained from experimental studies we used to obtain the positives, and only intersection of all these sources was retained in our benchmark. The complete set of negatives is available in Dataset S2.

Combinatorial Space of Variations in Motifs. Our strategy involved searching all possible variations of each motif by substitution of wildcards with all possible combinations of amino acids, e.g., [IVL] or [DE]. The combinatorial space of variations in motifs can be calculated using the binomial coefficient, which allows computing the number of ways of picking k unordered outcomes from n possibilities, as follows:

$$C_n^k = \frac{k!}{n!(k-n)!}.$$

Above, n represents the number of different amino acids (=20), whereas k represents the number of different amino acids picked to substitute a wildcard. By considering w as the total number of wildcards in a motif, the combinatorial space of variations of the motif is calculated as follows:

$$S_w = \prod_w \sum_{k=1}^{20} C_{20}^k = \prod_w \sum_{k=1}^{20} \frac{k!}{20!(k-20)!}.$$

As an example, we calculate below the combinatorial space of variations of the motif PXXP:

$$S_2 = \prod_2 \sum_{k=1}^{20} C_{20}^k = \left(\sum_{k=1}^{20} \frac{k!}{20!(k-20)!} \right)^2 \approx 10^{12}.$$

Normalization of *P*-Values. For each motif, two *P*-values were calculated, P_{NEG} and P_{BAK}, to score motif enrichment in the positives in comparison with the negatives and the background. The distribution of P_{BAK} was rescaled to have the same min and max as the distribution of P_{NEG}. Therefore, the same cutoff could be used as threshold for both *P*-values. We describe below the successive operations that were performed on the distribution of P_{BAK}. Below, the terms P_{BAK} and P_{NEG} are used as vector variables encompassing their respective distributions.

1. Logarithmic transformation of *P*-values (i.e., order of magnitude scale):

$$P_{BAK} = -\log_{10} P_{BAK},$$

$$P_{NEG} = -\log_{10} P_{NEG}.$$

2. Shift minimum of P_{BAK} to 0:

$$P_{BAK} = P_{BAK} - \min(P_{BAK}).$$

3. Scale P_{BAK} from 0 to 1:

$$P_{BAK} = \frac{P_{BAK}}{\max(P_{BAK})}.$$

4. Scale P_{BAK} from 0 to (max(P_{NEG}) − min(P_{NEG})):

$$P_{BAK} = P_{BAK}(\max(P_{NEG}) - \min(P_{NEG})).$$

5. Scale P_{BAK} from min(P_{NEG}) to max(P_{NEG}):

$$P_{BAK} = P_{BAK} + \min(P_{NEG}).$$

6. Exponential transformation of P_{BAK} and P_{NEG} (i.e., back to *P*-value scale):

$$P_{BAK} = 10^{-P_{BAK}},$$

$$P_{NEG} = 10^{-P_{NEG}}.$$

Probability to Find the Overlap Between Predicted and Known SH3 Binding Sites by Chance. The SH3 domain interactions we manually curated from the literature involve 361 proteins with a total length of 272,979 aa, encoding 749 experimentally known SH3 binding sites with a total length of 8,003 aa. The SH3 binding sites we predicted cover a length of 7,015 aa, among which 5,612 overlap with known SH3 binding sites. We calculate the probability to obtain such overlap by chance using the cumulative hypergeometric distribution, which scores the probability to see by chance at least k successes in a sample of size n picked from a finite population of size N containing m successes (Fig. 1*D*). Therefore, we consider the population as the total length of the 361 proteins (N = 272,979), the successes in the population as the amino acids within known SH3 binding sites (m = 8,003), the sample as the amino acids in predicted SH3 binding sites (n = 7,015), and successes in the sample as the amino acids overlapping between known and predicted SH3 binding sites (k = 5,612). By applying the formula described in Fig. 1*D*, i.e., using the high accuracy calculator located at keisan.casio.com, we obtained a probability inferior to 10^{-100}.

Binding Energy. Binding energy derived from physical energy terms, such as van der Waals, electrostatic, and desolvation energies, were obtained using the collection of high-confidence position-specific scoring matrices developed by Fernandez-Ballester et al. (57), and available in the ADAN database (58). For each SH3 domain in yeast, the ADAN database includes a series of positional matrices describing the contribution of each amino acid in terms of binding and stability energy between an SH3 domain and a target binding peptide.

Solvent Accessibility. Protein solvent accessibility was obtained with SABLE, version 2 (59) (using default input parameters), a program used for predicting real valued relative solvent accessibilities of amino acid residues in proteins. In our experiments, only residues with highest confidence level of solvent accessibility were considered in the analysis.

Sequence Conservation. For an input protein sequence, highly homologous sequences were collected from a proteome reference (i.e., here fungi proteome) using PSI-BLAST (60) (input: *e* value, 10^{-5}; comp-based stats, 1; number of iterations, 5) with 35% minimum homology. After that, highly similar sequences among collected homologues were filtered using CD-HIT with 95% maximum homology. After filtering, remaining homologous sequences including the input protein sequence were aligned using the MUSCLE algorithm (61) (using default input parameters). Finally, the Rate4Site program (62) (using default input parameters) was applied to the multiple sequence alignment to compute position-specific conservation scores of the input protein sequence across diverse species.

Intrinsic Disorder. Protein disorder was determined using DISOPRED 2 (63) with default input parameters. This software is designed to predict residues in protein sequences that are likely to be natively disordered. In our experimentation, only residues with the highest confidence level of disorder were considered as disordered.

Binding Affinity. The binding affinity was obtained from the work of Tonikian et al. (27), in which SH3 binding peptides were identified by SPOT peptide arrays, and then their binding specificity was scored based on signal intensity. In total, 295 peptides showed positive signal with at least one SH3 domain.

Functional Specificity. Given an SH3 binding site, we found its matching motif (among those we discovered) with the best *P*-value, which reflects its binding specificity. Then, we scanned the proteome to find proteins that matched that motif. We then found enriched GO terms for these proteins. Only manually curated GO terms were used. Specifically, among all of the evidence codes available for GO terms, we did not consider those with the code IEA (Inferred from Electronic Annotation) because they have not been manually assigned by a curator (described in *Guide to GO Evidence Codes* available at geneontology.org/page/guide-go-evidence-codes). The version of the GO annotation used in this work was downloaded in April 2015.

The GO enrichment was performed using the hyperGTest function from the GOstats R package (i.e., details in Table S1). The obtained *P*-values were corrected for multiple hypothesis testing using the “Bonferroni” method, after which we picked GO terms that corresponded to the corrected *P*-values that were lower than 10^{-3}. Selected GO terms were then used to calculate the functional specificity of the motif. In GO hierarchy, we considered both ancestors and offspring of each GO term, and then we calculated functional specificity that incorporates this information, by measuring the proportion of ancestors of each GO term over the total number of “reachable” terms, i.e., ancestors plus offspring, as follows:

$$\text{function specificity} = \frac{\text{number of ancestors}}{\text{number of ancestors} + \text{number of offspring}}.$$

When, for a given motif, multiple GO terms were found to be enriched, the functional specificity was calculated separately for each GO term, and then they were simply averaged to obtain the functional specificity for that motif. During our experiments, we used each branch separately: MF, Molecular Function, and BP, Biological Process. The correlations we obtained were significant for both branches. However, we decided to present only the correlation obtained for the Molecular Function branch, for the reason that it is related to “functional diversity.”

Because linear motifs with high binding specificity are found in fewer proteins than motifs with lower binding specificity, we wanted to be certain that GO enrichment was not biased by the number of proteins that we had in each set. To this end, we conducted a randomization-based analysis to determine whether there is any relationship between the number of proteins in a set and functional specificity. Thus, we randomly generated 1,000,000 sets of proteins with different sizes (i.e., size 10, 20, 30, . . ., 100, 200, 300, . . ., 1,000, equally represented) from the *Saccharomyces cerevisiae* proteome. Then we calculated the functional specificity for each set of proteins as described above (Fig. S9). We found that functional specificity is not related to the number of proteins in a set (correlation of −0.14); the average and SD of the functional specificity obtained for the different sizes were similar. Moreover, for all sizes, we observed a large variability of functional specificity, which suggests that, for a set of proteins of any number, we might obtain either high or low functional specificity.

ACKNOWLEDGMENTS. This work is dedicated to Stuart L. Schreiber on the occasion of his 60th birthday. This work was supported by the Canadian Institutes for Health Research Grants MOP-GMX-152556 and MOP-GMX-231013 (to S.W.M.), Natural Sciences and Engineering Research Council Grant 05707-2015 (to S.W.M.), the Fonds Québécois de la Recherche sur la Nature et les Technologies (A.K.), the Human Frontier Science Project Postdoctoral Fellowships Program (to E.D.L.), and the Weizmann Institute of Science (E.D.L.).

1. Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. *Science* 300(5626):1701–1703.
2. Pawson T (1995) Protein modules and signalling networks. *Nature* 373(6515):573–580.
3. Pawson T, Nash P (2003) Assembly of cell regulatory systems through protein interaction domains. *Science* 300(5618):445–452.
4. Pawson T, Raina M, Nash P (2002) Interaction domains: From simple binding events to complex cellular behavior. *FEBS Lett* 513(1):2–10.
5. Vogel C, Chothia C (2006) Protein family expansions and biological complexity. *PLoS Comput Biol* 2(5):e48.
6. Davey NE, Travé G, Gibson TJ (2011) How viruses hijack cell regulation. *Trends Biochem Sci* 36(3):159–169.
7. Neduva V, Russell RB (2005) Linear motifs: Evolutionary interaction switches. *FEBS Lett* 579(15):3342–3345.
8. McNiven MA (1998) Dynamin: A molecular motor with pinchase action. *Cell* 94(2): 151–154.
9. Cohen GB, Ren R, Baltimore D (1995) Modular binding domains in signal transduction proteins. *Cell* 80(2):237–248.
10. Zarrinpar A, Park SH, Lim WA (2003) Optimization of specificity in a cellular protein interaction network by negative selection. *Nature* 426(6967):676–680.
11. Tong AH, et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 295(5553):321–324.
12. Feng S, Chen JK, Yu H, Simon JA, Schreiber SL (1994) Two binding orientations for peptides to the Src SH3 domain: Development of a general model for SH3-ligand interactions. *Science* 266(5188):1241–1247.
13. Zarrinpar A, Bhattacharyya RP, Lim WA (2003) The structure and function of proline recognition domains. *Sci STKE* 2003(179):RE8.
14. Kaneko T, et al. (2003) Structural insight into modest binding of a non-PXXP ligand to the signal transducing adaptor molecule-2 Src homology 3 domain. *J Biol Chem* 278(48):48162–48168.
15. Kim J, Lee CD, Rath A, Davidson AR (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(3):889–901.
16. Berry DM, Nash P, Liu SKW, Pawson T, McGlade CJ (2002) A high-affinity Arg-X-X-Lys SH3 binding motif confers specificity for the interaction between Gads and SLP-76 in T cell signaling. *Curr Biol* 12(15):1336–1341.
17. Fazi B, et al. (2002) Unusual binding properties of the SH3 domain of the yeast actin-binding protein Abp1: Structural and functional analysis. *J Biol Chem* 277(7): 5290–5298.
18. Harkiolaki M, et al. (2003) Structural basis for SH3 domain-mediated high-affinity binding between Mona/Gads and SLP-76. *EMBO J* 22(11):2571–2582.
19. Kang H, et al. (2000) SH3 domain recognition of a proline-independent tyrosine-based RKxxYxxY motif in immune cell adaptor SKAP55. *EMBO J* 19(12):2889–2899.
20. Hoelz A, et al. (2006) Crystal structure of the SH3 domain of betaPIX in complex with a high affinity peptide from PAK2. *J Mol Biol* 358(2):509–522.
21. Seet BT, et al. (2007) Efficient T-cell receptor signaling requires a high-affinity interaction between the Gads C-SH3 domain and the SLP-76 RxxK motif. *EMBO J* 26(3): 678–689.
22. Liu Q, et al. (2003) Structural basis for specific binding of the Gads SH3 domain to an RxxK motif-containing SLP-76 peptide: A novel mode of peptide recognition. *Mol Cell* 11(2):471–481.
23. Mongioví AM, et al. (1999) A novel peptide-SH3 interaction. *EMBO J* 18(19): 5300–5309.
24. Lewitzky M, Harkiolaki M, Domart M-C, Jones EY, Feller SM (2004) Mona/Gads SH3C binding to hematopoietic progenitor kinase 1 (HPK1) combines an atypical SH3 binding motif, R/KXXK, with a classical PXXP motif embedded in a polyproline type II (PPII) helix. *J Biol Chem* 279(27):28724–28732.
25. Stamenova SD, et al. (2007) Ubiquitin binds to and regulates a subset of SH3 domains. *Mol Cell* 25(2):273–284.
26. Asbach B, Kolb M, Liss M, Wagner R, Schäferling M (2010) Protein microarray assay for the screening of SH3 domain interactions. *Anal Bioanal Chem* 398(5):1937–1946.

27. Tonikian R, et al. (2009) Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. *PLoS Biol* 7(10):e1000218.
28. Rodi DJ, Makowski L (1999) Phage-display technology—finding a needle in a vast molecular haystack. *Curr Opin Biotechnol* 10(1):87–93.
29. Mooney C, Pollastri G, Shields DC, Haslam NJ (2012) Prediction of short linear protein binding regions. *J Mol Biol* 415(1):193–204.
30. Fuxreiter M, Tompa P, Simon I (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* 23(8):950–956.
31. Stein A, Aloy P (2008) Contextual specificity in peptide-mediated protein interactions. *PLoS One* 3(7):e2524.
32. Dinkel H, et al. (2012) ELM–the database of eukaryotic linear motifs. *Nucleic Acids Res* 40(Database issue):D242–D251.
33. Stein A, Aloy P (2010) Novel peptide-mediated interactions derived from high-resolution 3-dimensional structures. *PLoS Comput Biol* 6(5):e1000789.
34. Davey NE, et al. (2012) SLiMPrints: Conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. *Nucleic Acids Res* 40(21): 10628–10641.
35. Edwards RJ, Davey NE, Shields DC (2007) SLiMFinder: A probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS One* 2(10):e967.
36. Neduva V, et al. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol* 3(12):e405.
37. Neduva V, Russell RB (2006) DILIMOT: Discovery of linear motifs in proteins. *Nucleic Acids Res* 34(Web Server issue):W350–W355.
38. Lieber DS, Elemento O, Tavazoie S (2010) Large-scale discovery and characterization of protein regulatory motifs in eukaryotes. *PLoS One* 5(12):e14444.
39. Nguyen Ba AN, et al. (2012) Proteome-wide discovery of evolutionary conserved sequences in disordered regions. *Sci Signal* 5(215):rs1.
40. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* (AAAI Press, Menlo Park, CA), Vol 2, pp 28–36.
41. Doğruel M, Down TA, Hubbard TJ (2008) NestedMICA as an ab initio protein motif discovery tool. *BMC Bioinformatics* 9:19.
42. Dinh H, Rajasekaran S, Davila J (2012) qPMS7: A fast algorithm for finding (ℓ, d)-motifs in DNA and protein sequences. *PLoS One* 7(7):e41425.
43. Gfeller D, et al. (2011) The multiple-specificity landscape of modular peptide recognition domains. *Mol Syst Biol* 7:484.
44. Bellay J, et al. (2011) Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol* 12(2):R14.
45. London N, Movshovitz-Attias D, Schueler-Furman O (2010) The structural basis of peptide-protein binding strategies. *Structure* 18(2):188–199.
46. Davey NE, et al. (2012) Attributes of short linear motifs. *Mol Biosyst* 8(1):268–281.
47. Jia CY, Nie J, Wu C, Li C, Li SS (2005) Novel Src homology 3 domain-binding motifs identified from proteomic screen of a Pro-rich region. *Mol Cell Proteomics* 4(8): 1155–1166.
48. Musacchio A (2002) How SH3 domains recognize proline. *Adv Protein Chem* 61: 211–268.
49. Kuriyan J, Eisenberg D (2007) The origin of protein interactions and allostery in colocalization. *Nature* 450(7172):983–990.
50. Levy ED, Kowarzyk J, Michnick SW (2014) High-resolution mapping of protein concentration reveals principles of proteome architecture and adaptation. *Cell Rep* 7(4): 1333–1340.
51. Ashburner M, et al.; The Gene Ontology Consortium (2000) Gene ontology: Tool for the unification of biology. *Nat Genet* 25(1):25–29.
52. Moseley JB, Goode BL (2006) The yeast actin cytoskeleton: From cellular function to biochemical mechanism. *Microbiol Mol Biol Rev* 70(3):605–645.
53. Marles JA, Dahesh S, Haynes J, Andrews BJ, Davidson AR (2004) Protein-protein interaction affinity plays a crucial role in controlling the Sho1p-mediated signal transduction pathway in yeast. *Mol Cell* 14(6):813–823.
54. Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286(5438):295–299.
55. Landry CR, Levy ED, Abd Rabbo D, Tarassov K, Michnick SW (2013) Extracting insight from noisy cellular networks. *Cell* 155(5):983–989.
56. Dutkowski J, et al. (2013) A gene ontology inferred from molecular networks. *Nat Biotechnol* 31(1):38–45.
57. Fernandez-Ballester G, et al. (2009) Structure-based prediction of the *Saccharomyces cerevisiae* SH3-ligand interactions. *J Mol Biol* 388(4):902–916.
58. Encinar JA, et al. (2009) ADAN: A database for prediction of protein-protein interaction of modular domains mediated by linear motifs. *Bioinformatics* 25(18): 2418–2424.
59. Wagner M, Adamczak R, Porollo A, Meller J (2005) Linear regression models for solvent accessibility prediction in proteins. *J Comput Biol* 12(3):355–369.
60. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
61. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
62. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18(Suppl 1):S71–S77.
63. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337(3):635–645.