Journal of proteome research

Article

pubs.acs.org/jpr

Characterization of Domain−Peptide Interaction Interface: Prediction of SH3 Domain-Mediated Protein−Protein Interaction Network in Yeast by Generic Structure-Based Models

Tingjun Hou,*[,†,‡] Nan Li,[§] Youyong Li,[†] and Wei Wang*[,§]

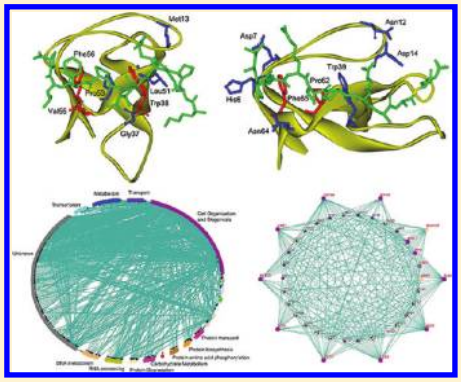
[†]Institute of Functional Nano & Soft Materials (FUNSOM) and Jiangsu Key Laboratory for Carbon-Based Functional Materials & Devices, Soochow University, Suzhou, Jiangsu 215123, China

[‡]College of Pharmaceutical Science, Soochow University, Suzhou, Jiangsu 215123, China

[§]Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, California 92093, United States

Supporting Information

ABSTRACT: Determination of the binding specificity of SH3 domain, a peptide recognition module (PRM), is important to understand their biological functions and reconstruct the SH3-mediated protein−protein interaction network. In the present study, the SH3-peptide interactions for both class I and II SH3 domains were characterized by the intermolecular residue−residue interaction network. We developed generic MIEC-SVM models to infer SH3 domain-peptide recognition specificity that achieved satisfactory prediction accuracy. By investigating the domain−peptide recognition mechanisms at the residue level, we found that the class-I and class-II binding peptides have different binding modes even though they occupy the same binding site of SH3. Furthermore, we predicted the potential binding partners of SH3 domains in the yeast proteome and constructed the SH3-mediated protein−protein interaction network. Comparison with the experimentally determined interactions confirmed the effectiveness of our approach. This study showed that our sophisticated computational approach not only provides a powerful platform to decipher protein recognition code at the molecular level but also allows identification of peptide-mediated protein interactions at a proteomic scale. We believe that such an approach is general to be applicable to other domain−peptide interactions.



KEYWORDS: *protein recognition code, molecular interaction energy component (MIEC), support vector machine (SVM), molecular dynamics, MM/GBSA, binding interface*

INTRODUCTION

Protein−protein interactions (PPI) are essential for cellular functions and often mediated by interactions between modular domains and peptides.[1] The SH3 domain is abundant in the proteomes[2] and plays crucial functional roles in many proteins, especially signaling and cytoskeletal proteins.[3] SH3 domains are about 60 residues long and with a structure characterized by a five-stranded antiparallel β-barrel.[4−6] Clustered aromatic residues form a hydrophobic site on its surface recognizing ∼10-residue-long peptides with a PXXP core motif (P represents proline and X represents any amino acid) that has a left-handed polyproline type II (PPII) helical conformation.[7] Peptide ligands can bind to SH3 domains in two opposite orientations, conforming to either class-I ([R/K]XXPXXP) or class-II (PXXPX[R/K]) consensus motif, respectively.[6,8]

Identification of the binding peptides of SH3 domains is a crucial step toward understanding the SH3-mediated protein−protein interaction network. Peptide library or peptide array are often exploited to identify SH3-binding peptides.[9−13] Computational approaches have also been developed to identify the peptide segments in proteins that are potentially bound by SH3 domains.[14−24] For example, the SH3-SPOT method builds a position-specific contact frequency matrix (PSCFM) based on the protein−peptide contacts in numerous SH3/peptide crystal structures and the probability that a peptide would bind to the given SH3 domain is then calculated based on the PSCFM.[14] Improved versions of this approach have also been developed by using machine-learning algorithms, such as artificial neural network (ANN) and support vector machine (SVM), to analyze the contact matrix.[15,19,20,22] The contact-based approach does not quantitatively characterize the residue−residue interactions between peptide and SH3. It is also limited by the huge number of possible combinations of contact residue pairs that cannot be well sampled by the relatively small available SH3-peptide interaction pairs. Alternatively, molecular modeling techniques, such as molecular docking, molecular dynamics (MD) simulations and free energy calculation, can quantitatively predict the SH3-peptide binding.[17,19−21] These structure-based approaches do not train models for a specific

Received: January 20, 2012
Published: April 2, 2012

system, but the performance of such calculations often relies on long simulations required for accurately modeling the domain-peptide complex structures and calculating the binding affinities.

To overcome these hurdles, we have developed a method called MIEC-SVM[19,20] that uses molecular interaction energy components (MIECs) to characterize the residue−residue interaction pattern between peptide and SH3. SVM is then trained on the MIECs to classify peptides into binder or nonbinder class. Previously, we conducted a proof-of-concept study on 18 SH3 domains binding to the class-I peptides and illustrated the usefulness of the MIEC-SVM technique in deciphering protein recognition code. In the present study, we further improved this method and developed a generic MIEC-SVM model to infer the binding specificities of the SH3 domains recognizing the class-II peptides. We then updated the model for the class-I peptide binding SH3 domains by considering additional binding data and energetic components to be consistent with the class-II peptide model. The two MIEC-SVM models clearly illustrated that the SH3 recognition codes for the two classes of binding peptides are different because the important residues for these two types of binding peptides are not completely identical.

To further illustrate the power of our approach, we tackled an important but challenging problem of constructing protein interaction networks mediated by domain-peptide binding. We predicted the interacting partners of the yeast SH3 domains using the generic MIEC-SH3 models in the yeast proteome and assembled these interactions into a SH3-mediated protein−protein interaction network. Our predictions correlated well with the experimental measurements using yeast two-hybrid, peptide array and phase display. The energetic analysis for each interacting domain−peptide pair provides molecular insights into the formation of protein interaction network, which is complementary to the high throughput approach that only determines a wired diagram. We expect such a structure-based approach will become increasingly useful in understanding protein recognition and constructing protein interaction network with the fast advancement of structure genomics and proteomics.

MATERIALS AND METHODS

1. Data Set for the Class-I Binding Peptides

We studied 23 SH3 domains recognizing class-I peptides: Abl, Boi1, Bzz1_1, Bzz1_2, Fyn, Grb2, Hck, Hse1, Itk, Lsb3, Lyn, Myo3, Myo5, Nbp2, P85a, Pex13, Rvs167, Sla1_3, Spta2, c-Src, Sho1, Yes and Ysc84 (see Table S1 for details, Supporting Information). The class-I binding peptides for these SH3 domains were collected from literature.[9−12,25] All but five of these domains were included in our previous study:[20] Bzz1_1, Bzz1_2, Hse1, Pex13 and Sho1. All peptides were ten-residue-long. If a binding peptide only had nine residues, for example, PTYPPPPPP for the Abl SH3 domain, we randomly generated 5 peptides by adding 1 amino acid to make it 10-residues long. We assumed that the added residues would not drastically change the binding specificity of these peptides. We ignored the known binding peptides less than 9 residues.

The experimentally determined nonbinders for Boi2, Lsb3, Myo5, Rvs167 and Ysc84 were included.[10] Because the ratio between nonbinders and binders of a SH3 domain is about 20 in a given proteome,[10] we had to include additional nonbinder peptides to mimic this scenario in the data set, same as in our previous work.[19,20] For SH3 domains without experimental nonbinders, because the percentage of a random peptide being binder is small, we randomly selected 10-residue-long peptides as nonbinders from the Swiss-Prot database[26] using two motifs: half nonbinders with the PXXP motif and half nonbinders with a more specific motif ((Y/W/F/M)XXPXXP for Abl, (Y/W/F/M)XXPXXP for Bzz1_1 and Bzz1_2, and (R/K)XXPXXP for the rest). In total, there were 491 binders and 9820 nonbinders for the class-I peptides (Table S1 in the Supporting Information).

2. Data Set for the Class-II Binding Peptides

In this study, 16 SH3 domains recognizing the class-II peptides were considered: Amph, Asp2, Bbc1, Boi1, Boi2, Crk, Grb2, Lsb1, Lsb3, Lsb4, Pig1, Pin3, Rvs167, Sh3g2, Src8 and c-Src (see Table S2 for details, Supporting Information). The 10-residue-long binding peptides for these domains were collected from literature.[9,10,25,27] For a binding peptide with 9 residues, similar to class-I peptides, 5 peptides were randomly generated by adding 1 amino acid to make it 10-residues long. The binding peptides less than 9 residues were ignored.

Same as the class-I nonbinding peptides, the ratio of nonbinders versus binders for the class-II SH3 domains was also set to 20. For Amph, Boi1, Boi2, Lsb3, Lsb4, Rvs167, Rvs167 and Sh3g2, the nonbinders given by experiments were included in the data set.[10] For each SH3 domain without experimental nonbinder, ten-residue-long peptides randomly selected from the Swiss-Prot database were used as nonbinders: half nonbinders with the PXXP motif and half nonbinders with the PXXPX(R/K) motif. In total, there were 599 binders and 11980 nonbinders for the class-II peptides (Table S2 in the Supporting Information).

3. Modeling the Class-I SH3-Peptide Complexes

Among the 23 class-I SH3 domains, five of them, including Abl (PDB entry: 1bbz),[28] c-Src (1qwf),[29] Hck (2oi3),[30] Fyn (1fyn)[5] and Sho1 (2vkn),[31] had crystal structures complexed with class-I peptides; two of them, including Grb2 (1gbq)[32] and Spta2 (2pqh),[33] had crystal complex structures with class-II peptides; 13 of them had protein crystal structures only (without binding peptides): Itk (1awj),[34] Lyn (1w1f),[35] Myo3 (1ruw),[36] Myo5 (1yp5),[37] Nbp2 (1yn8),[38] P85a (1pht),[39] Sla1_3 (2jt4),[40] Yes (2hda),[41] Lsb3 (1oot),[42] Ysc84 (2a08),[31] Pex13 (1jqq),[43] Bzz1_1 (1zuu),[31] and Bzz1_2 (2a28).[31] For the rest SH3 domains, including Hse1, Rvs167 and Boi1, no crystal structure was available and thus homology modeling was used to model their structures from the scratch. The homology models for Rvs167 and Boi1 reported in our previous work were used here.[20] The templates in PDB which had the high sequence similarities with Hse1 were searched by *Blast protocol* in Discovery Studio.[44] Because we could not find a single template with high sequence similarity with Hse1, 13 templates were used in comparative homology modeling. The sequence and structure alignment were performed using *Align sequence with structure protocol* in Discovery Studio, and the homology model was constructed using *Build homology models protocol* in Discovery Studio. Next, the modeled Hse1 structure was immersed in a sphere of water molecules with harmonic restraint and minimized with the CHARMM force field[45] using *Minimization protocol* in Discovery Studio. The quality of the modeled Hse1 structure was verified by Profile-3D[46] (score = 22.6, expected high score = 29.2 and expect low score = 13.1) in Discovery Studio.

dx.doi.org/10.1021/pr3000688 | J. Proteome Res. 2012, 11, 2982−2995

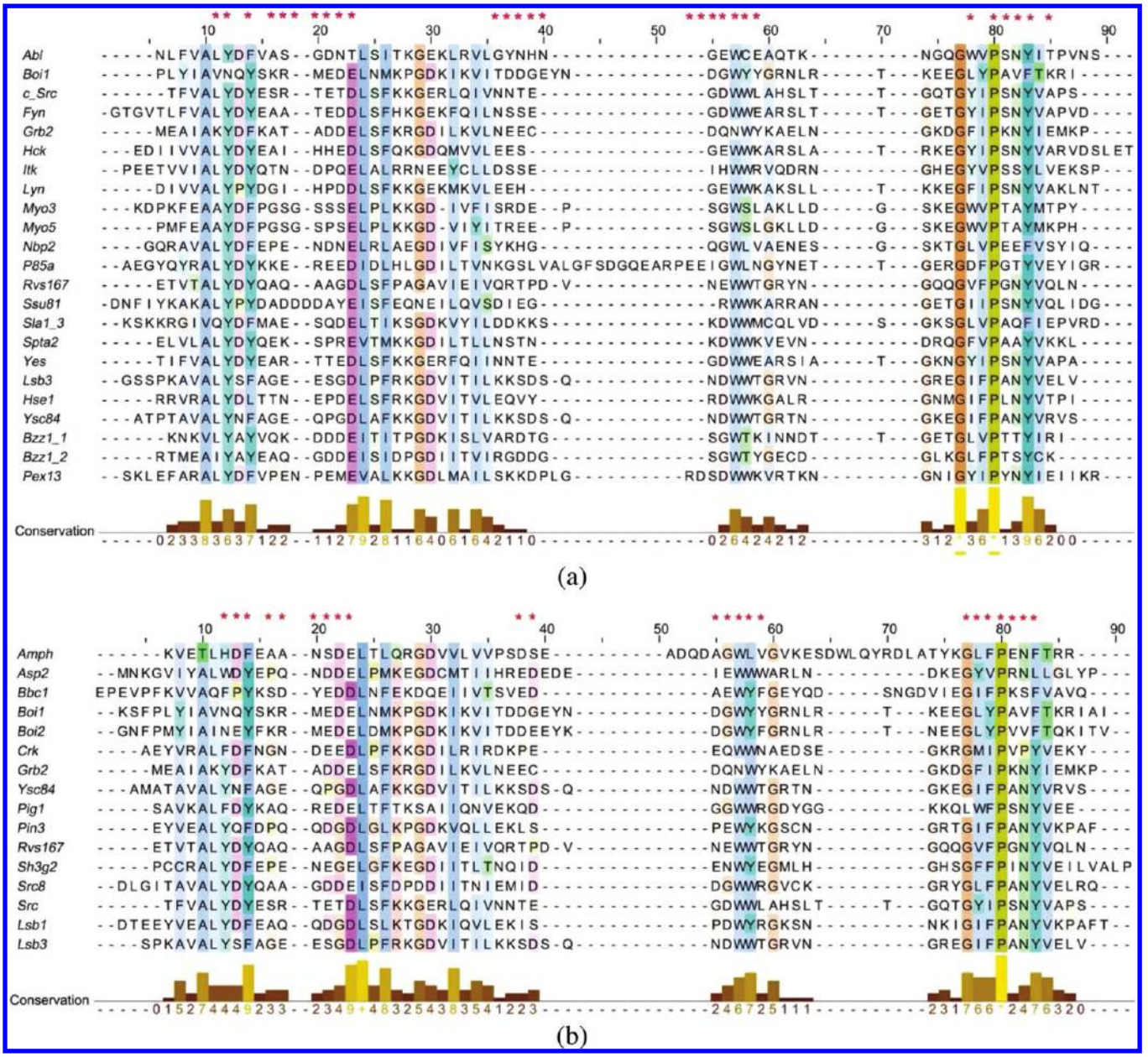


Figure 1. Important positions used to calculate the SH3-peptide MIECs. Asterisks in the first line of the multiple sequence alignments show the important positions for (a) the class-I SH3 domains and (b) the class-II SH3 domains. The alignments are colored according to the consensus sequence conservation (conservation larger than 20%) using the ClustalX coloring scheme.[69] The figures were generated using Jalview.[70]

For the 18 SH3 domains having no complex structures with the class-I binding peptides, each modeled or crystal unbound SH3 domain was aligned to the 5 SH3 crystal structures complexed with the class-I binding peptides. Then 1 crystal complex was chosen as a template based on structural similarity. The binding peptide in the template complex was merged into the unbound SH3 and the peptide was mutated to that of the unbound SH3 using the *scap* program.[47] The modeled complexes were optimized by 5000 steps of molecular mechanics (MM) minimizations followed by molecular dynamics (MD) simulations. The MM minimization and MD simulations were performed using the AMBER10.0 software package[48] and the AMBER03 force field.[49] The complex was solvated in a rectangular box that extended 9 Å away from any solute atom. Counter ions of Na^+ were placed near the SH3 domain on a grid based on the Columbic potential to keep the entire system neutral. Particle Mesh Ewald (PME) was employed to calculate the long-range electrostatic interactions.[50] The SHAKE procedure was employed to constrain all bonds involving hydrogen atoms[51] and the time step was 2.0 fs. In the MD simulations, temperature was gradually increased from 10 to 300 K during the first 20 ps, and the following 2 ns simulation was for equilibration and data collection. The final snapshot of the MD simulation was optimized by 5000 steps of MM minimization and the minimized conformation was used as the template structure for modeling the other peptides in the data set interacting with the same SH3 domain. After MD simulations and MM minimizations, the PPII helical conformation of the binding peptide and the important contacts between SH3 and peptide were retained.

dx.doi.org/10.1021/pr3000688 | *J. Proteome Res.* 2012, 11, 2982−2995

4. Modeling the Class-II SH3-Peptide Complexes

In the 16 class-II SH3 domains, four of them, including Crk (1cka),[52] Grb2 (1gbq),[32] Pig1 (1ywo),[53] and c-Src (1qwe),[29] had the crystal structures complexed with the class-II binding peptides; two of them, including Src8 (2d1x)[54] and Bbc1 (1zuk),[31] had crystal complexes with the class-I binding peptides. Five SH3 domains, Asp2 (1ycs),[55] Lsb3 (1oot),[42] Ysc84 (2a08),[31] Pin3 (1zx6)[31] and Sh3g2 (2dbm),[56] had unbound crystal structures. For the other five SH3 domains, including Amph, Boi1, Boi2, Lsb1 and Rvs167, no crystal structure was available. The homology models for Amph, Rvs167 and Boi1 were obtained from our previous study.[20,57] The homology model of Boi2 was constructed based on 2cuc[58] as the template (sequence similarity = 50.7%) by *Build homology models protocol* in Discovery Studio, and that of Lsb1 was constructed based on 1zx6[31] as the template (sequence similarity = 70.8%). The modeled structures for Boi2 and Lsb1 showed good quality evaluated by Profile-3D (data not shown here). For the 12 SH3 domains without the complex structures, we used the 4 crystal complex structures as the template and employed the protocol described above to model their complex structures.

In total, we obtained 23 complexes for the class-I SH3 domains and 16 complexes for the class-II SH3 domains. These complexes were used as the initial templates to construct the complexes for all 25620 peptides in the data set. For each SH3, the peptide in the template was mutated to the target peptide using the *scap* program.[47] Then each modeled complex was minimized by the *sander* program in AMBER10.0[48] using the AMBER03 force field.[49] Because of the large number of peptides under consideration, the generalized Born (GB) model (*igb* = 2)[59] implemented in sander was used to consider the solvent effect. The maximum number of minimization steps was set to 4000 and the convergence criterion for the root-mean-square (rms) of the Cartesian elements of the energy gradient was 0.05 kcal/mol/Å. The first 500 steps were performed with the steepest descent algorithm and the rest of the steps with the conjugate gradient algorithm.

5. Calculating the Molecular Interaction Energy Components (MIECs)

For each complex, the minimized conformation was used to calculate MIECs. First, we identified the important residues located close to the binding peptide in any of the template complexes. It is possible that residues important for one SH3 domain may not be important for another and/or insertion/deletion may occur at this position in another SH3 domain. To build a generic model for SH3-peptide interactions, we took a union of the important interacting pairs identified from all SH3 domains (Figure 1). The distance cutoff to identify residue contacts was determined based on the prediction capability of the final model (6 and 5 Å were used for class-I and class-II SH3 domains, respectively). SH3 residues in 28 and 23 positions were identified this way that may form significant interactions with the class-I and class-II peptides, respectively. As an example, the spatial distribution of the important residues for Boi1 is shown in Figure S1 in the Supporting Information.

Next, based on the multiple sequence alignment shown in Figure 1, 75 important SH3-peptide interacting pairs were determined in both class-I and class-II complexes. An example of these interacting pairs for the Lsb3 SH3 domain is shown in Table S3 in the Supporting Information. Note that Lsb3 has gaps at three residues (represented as 0 in Table S3) interacting with the class-I peptides and correspondingly the MIECs of these gap positions were set to 0.

The binding free energy between each peptide and SH3 was decomposed into residue−residue pairs by the MM/GBSA free energy decomposition protocol[19,20,60−63] using the *mm_pbsa* program in AMBER10.[48] Then the MIECs for each interacting pair shown in Table S3 in the Supporting Information were extracted from the decomposition results. The MIECs included: (a) electrostatic (Columbic) interaction ΔE_{ele}, (b) van der Waals interaction ΔE_{vdw}, (c) polar contribution to desolvation free energy ΔG_{GB}, (d) nonpolar contribution to desolvation free energy ΔG_{SA}. The cutoff for calculating ΔE_{vdw} and ΔE_{ele} was set to 18.0 Å. A distance-independent interior dielectric constant of 1 was used to calculate ΔE_{ele}. In the GB calculations, the charges were taken from the AMBER03 force field and the GB parameters developed by Onufriev and co-workers were used.[59] The values of interior dielectric and exterior dielectric constants in the GB calculations were set to 1 and 80, respectively. The nonpolar contribution to desolvation was computed based on solvent-accessible surface area (SASA) using the LCPO method:[64] $\Delta G_{SA} = 0.0072 \times \Delta SASA$.

We also calculated the four MIECs for the 9 adjacent residue pairs of the 10-residue long peptides and internal energies of each peptide residue to reflect the conformational preference of the peptide. In total, for each peptide MIECs of 84 (= 75 + 9) residue−residue pairs and 10 peptide residues were represented by a MIEC vector $\boldsymbol{X}$. The dimension of $\boldsymbol{X}$ depends on which energy terms were included in the model. For example, when only ΔE_{vdw} was considered, the dimension of $\boldsymbol{X}$ was 94; when all four energy terms were considered, the dimension of $\boldsymbol{X}$ was 376 (= 94 × 4).

6. Constructing the MIEC-SVM Models

The value of the response variable $\boldsymbol{Y}$ was 1 for a binder or −1 for a nonbinder. The MIEC matrix was then normalized and trained by support vector machine (SVM)[65,66] implemented in the *libsvm* program.[67] The entire data set was randomly partitioned into three groups with equal sizes. Two groups were used for training and the third group for validation. This procedure was run for 500 times to evaluate the performance of the SVM classifiers. For each SVM, TP (true positive), FP (false positive), TN (true negative), and FN (false negative) of the 500 test sets were counted. The prediction performance was evaluated by calculating the average values of: sensitivity, SE = TP/(TP + FN); specificity, SP = TN/(TN + FP); prediction accuracy for binders, Q_+ = TP/(TP + FP); prediction accuracy for nonbinders, Q_- = TN/(TN + FN); and Matthews correlation coefficient, $C = (TP \times TN - FN \times FP)/(((TP + FN)(TP + FP)(TN + FN)(TN + FP))^{1/2})$. Because the numbers of positives and negatives were quite unbalanced, a higher weight (k_+) was applied to the positive class. k_+ was initially set to 12, and then various k_+ values were tested to achieve the best performance (see discussions below).

7. Screening the Yeast Proteome to Identify Putative Binders of SH3 Domains

We used the generic MIEC-SVM models to identify potential binding peptides of these SH3 domains in the yeast proteome, which would help construct the SH3-mediated protein−protein interaction network. First, we defined several relaxed sequence motifs based on the available experimental data to scan the yeast proteome. The motif search was efficient and reduced the searching space significantly (the relaxed motifs are shown in Table S4 in the Supporting Information). Second, for each SH3

dx.doi.org/10.1021/pr3000688 | J. Proteome Res. 2012, 11, 2982−2995

Table 1. Performance of Various MIEC-SVM Models for the SH3 Class-II Binding Peptides[a]

model	MIECs[b]	SE_{train} (%)	SP_{train} (%)	SE_{test} (%)	SP_{test} (%)	Q_+ (%)	Q_+ (%)	C
				SH3-peptide MIECs				
1	ΔE_{vdw}, ΔE_{ele}	90.2	87.0	84.9	86.7	24.2	99.1	0.409
2	ΔE_{polar}[c], $\Delta E_{nonpolar}$[c]	86.4	85.0	82.7	84.7	21.3	99.0	0.370
3	ΔE_{vdw}, ΔE_{ele}, ΔG_{GB}	90.2	87.3	85.5	86.9	24.6	99.2	0.415
4	ΔE_{vdw}, ΔE_{ele}, ΔG_{GB}, ΔG_{sa}	90.5	87.7	86.2	87.3	25.3	99.2	0.425
				SH3-peptide MIECs and peptide adjacent residue MIECs				
5	ΔE_{vdw}, ΔE_{ele}	90.7	90.2	85.0	89.9	29.7	99.2	0.465
6	ΔE_{polar}, $\Delta E_{nonpolar}$	90.3	90.3	85.1	90.0	29.9	99.2	0.467
7	ΔE_{vdw}, ΔE_{ele}, ΔG_{GB}	90.4	90.5	84.4	90.2	30.1	99.1	0.467
8	ΔE_{vdw}, ΔE_{ele}, ΔG_{GB}, ΔG_{sa}	92.8	91.0	87.3	90.7	32.0	99.3	0.494
				SH3-peptide MIECs, peptide adjacent residue MIECs and peptide residue internal energies				
9	ΔE_{vdw}, ΔE_{ele}	93.6	90.9	88.6	90.6	32.0	99.4	0.499
10	ΔE_{polar}, $\Delta E_{nonpolar}$	93.1	91.8	87.9	91.4	33.9	99.3	0.514
11	ΔE_{vdw}, ΔE_{ele}, ΔG_{GB}	95.2	91.6	90.2	91.2	34.0	99.5	0.523
12	ΔE_{vdw}, ΔE_{ele}, ΔG_{GB}, ΔG_{sa}	95.6	92.4	90.1	92.1	36.3	99.5	0.542

[a]RBF kernel was used in SVM. [b]ΔE_{vdw}, ΔE_{ele}, ΔG_{GB}, ΔG_{SA} are van der Waals, electrostatic, polar contribution to desolvation, and nonpolar contribution to desolvation, respectively. [c]$\Delta G_{polar} = \Delta E_{ele} + \Delta G_{GB}$, $\Delta G_{nonpolar} = \Delta E_{vdw} + \Delta G_{SA}$

Table 2. Performance of Various MIEC-SVM Models for the SH3 Class-I Binding Peptides[a]

model	MIECs[b]	SE_{train} (%)	SP_{train} (%)	SE_{test} (%)	SP_{test} (%)	Q_+ (%)	Q_+ (%)	C
				SH3-peptide MIECs				
1	ΔE_{vdw}, ΔE_{ele}	78.2	89.5	71.7	89.1	30.3	98.0	0.415
2	ΔE_{polar}[c], $\Delta E_{nonpolar}$[c]	72.2	89.5	66.6	89.2	28.9	97.6	0.385
3	ΔE_{vdw},ΔE_{ele},ΔG_{GB}	79.3	89.1	73.4	88.8	30.1	98.1	0.419
4	ΔE_{vdw},ΔE_{ele},ΔG_{GB}, ΔG_{SA}	80.3	89.7	74.2	89.4	31.6	98.1	0.435
				SH3-peptide MIECs and SH3 residue−residue MIECs				
5	ΔE_{vdw}, ΔE_{ele}	85.0	91.2	79.7	90.9	36.6	98.5	0.498
6	ΔE_{polar}, $\Delta E_{nonpolar}$	85.7	90.5	80.2	90.2	35.1	98.6	0.487
7	ΔE_{vdw}, ΔE_{ele}, ΔG_{GB}	90.8	91.2	86.3	90.9	38.3	99.0	0.537
8	ΔE_{vdw}, ΔE_{ele}, ΔG_{GB}, ΔG_{SA}	92.1	91.7	87.9	91.3	40.1	99.1	0.558
				SH3-peptide MIECs, SH3 residue−residue MIECs and peptide residue internal energies				
9	ΔE_{vdw}, ΔE_{ele}	90.5	93.6	85.7	93.3	45.7	99.0	0.595
10	ΔE_{polar}, $\Delta E_{nonpolar}$	89.8	93.7	84.7	93.3	45.6	98.9	0.589
11	ΔE_{vdw}, ΔE_{ele}, ΔG_{GB}	93.2	94.1	89.2	93.8	48.6	99.2	0.630
12	ΔE_{vdw}, ΔE_{ele}, ΔG_{GB}, ΔG_{SA}	94.1	94.8	90.4	94.5	51.9	99.3	0.659

[a]RBF kernel was used in SVM. [b]ΔE_{vdw}, ΔE_{ele}, ΔG_{GB}, ΔG_{SA} are van der Waals, electrostatic, polar contribution to desolvation, and nonpolar contribution to desolvation, respectively. [c]$\Delta G_{polar} = \Delta E_{ele} + \Delta G_{GB}$, $\Delta G_{nonpolar} = \Delta E_{vdw} + \Delta G_{SA}$.

complex template, the binding peptide was mutated to the peptides found by the motif search using the *scap* program.[47] Each modeled complex was minimized by the *sander* program in AMBER10 using the AMBER03 force field.[49] The minimization procedure was same as training the SVM models. Third, MIECs were calculated by MM/GBSA free energy decomposition for each peptide, same as described above. Fourth, the MIEC vector for each peptide was normalized using the scaling parameters determined from the scaling process in training the SVM models. Finally, based on the normalized MIEC vector each peptide was classified by the MIEC-SVM models as binder or nonbinder. If any 10-residues-long peptide segment in a yeast protein was a predicted binder of a given SH3 domain, this protein was assumed to potentially interact with this SH3 domain.

RESULTS AND DISCUSSIONS

1. Unified MIEC-SVM Model for the Class-II SH3 Binding Peptides

To have a comprehensive understanding of the SH3 binding specificity, we also built a unified MIEC-SVM model for the SH3 domains recognizing class-II peptides. In the initial training process, a large weight ($k_+ = 12$) was given to the binder class and small weight ($k_- = 1$) to the nonbinder class. First, the performance of the SVM models based on the combinations of various MIECs was evaluated (models 1−4 in Table 1). When using four individual energy terms, the MIEC-SVM model (model 4 in Table 1) performed best in the 500 runs of cross-validations ($C = 0.425$, $SE_{test} = 86.3\%$ and $SP_{test} = 87.3\%$). It was unsurprising because different energy terms characterize different aspects of the peptide-SH3 interactions. In our previous work[19,20] nonpolar contribution to desolvation (ΔG_{sa}) was not considered but it did contribute to improve the prediction accuracy in this study. Inclusion of the adjacent peptide residue MIECs further improved the performance of SVM (models 5 to 8 in Table 1), which is consistent with our previous observations,[20] because these MIECs reflect the conformational preferences of the binding peptides. Moreover, the internal energy of each peptide residue was included to characterize the conformational preference of that residue. For example, larger residue should have larger G_{SA}, and strong hydrophilic residue should have significant polar contribution

dx.doi.org/10.1021/pr3000688 | J. Proteome Res. 2012, 11, 2982−2995

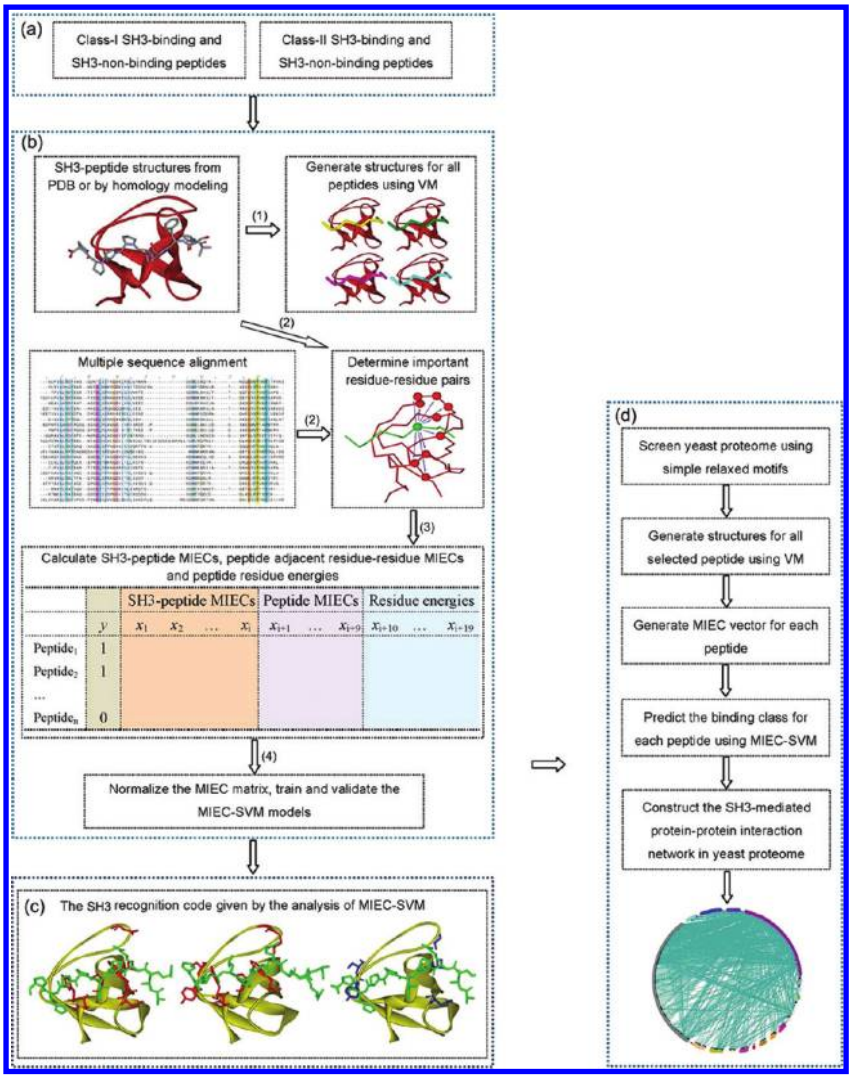


Figure 2. Determination of the SH3 recognition codes and the predictions of the SH3-mediated protein−protein interaction network using the MIEC-SVM models. (a) Peptide binders for two classes of SH3 were collected from literature; some peptide nonbinders were collected from literature and the others were randomly selected from the Swissprot sequence database with the predefined motifs. (b) Procedure to construct the MIEC-SVM models: (1). Model the SH3-peptide complexes using Virtual Mutagenesis (VM) and GB-based molecular mechanics minimization; (2) Identify the important SH3 residues that form effective interactions with the peptides based on the complex structures and the multiple sequence alignment (the residue of peptide is shown as the green CPK model and the SH3 residues which can form effective interactions with the peptide residue are shown as the red CPK models); (3) Calculate the SH3-peptide MIECs, the peptide adjacent residue−residue MIECs and the peptide residue energies using the MM/GBSA free energy decomposition analysis; the calculation results are saved as a MIEC matrix; In the MIEC matrix, column y is the binding class for each peptide, 1 for binder and −1 for nonbinder; columns x_1 to x_i are the MIECs for the SH3-peptide interaction pairs; columns x_{i+1} to x_{i+9} are the MIECs for the nine pairs between the adjacent peptide residues; columns x_{i+10} to x_{i+19} are the energies for the ten residues in a peptide; it should be noted in this figure only one energy term is used; (4) Normalize the MIEC matrix, train and validate the universal MIEC-SVM models. (c) Determination of the SH3 recognition codes by analysis of the MIEC-SVM models; as an example, the important residues of the Lsb3 SH3 for recognizing the class-I peptide binding, the important residues of the Lsb3 SH3 for recognizing the class-II peptide binding, and the different residues of the Lsb3 SH3 for these two different classes of peptides are shown in three figures from left to right. (d). Construction of the SH3-mediated protein−protein interaction network in the yeast proteome by the two universal MIEC-SVM models.

to desolvation. As shown in Table 1 (models 9 to 12), addition of the peptide residue internal energies improved the performance of the MIEC-SVM model considerably. In summary, the best MIEC-SVM model performed quite well as evaluated by the 500 runs of cross-validations (C = 0.542, SE_{test} = 90.1% and SP_{test} = 92.0%).

Because binders and nonbinders of a SH3 domain in the proteome are quite unbalanced (the ratio of binders to nonbinders is about 1:20),[10] it is crucial to choose the weights of positives and negatives in SVM. We set k_- to 1 and systematically evaluated the performance of MIEC-SVM using k_+ in the range from 2 to 14 (Figure S2 in the Supporting Information), which showed that k_+ = 3 or 4 was an optimal choice to achieve balanced sensitivity, specificity, MCC and prediction accuracy for the test set.

2. Unified MIEC-SVM Prediction Model for the Class-I SH3 Binding Peptides

In our previous work, we already developed a unified MIEC-SVM prediction model based on 18 class-I SH3 domains.[20] Our previous work showed that binders and nonbinders exhibit different peptide-SH3 residue−residue interaction patterns and such patterns can be captured by MIECs.[20] Based on the MIEC

dx.doi.org/10.1021/pr3000688 | J. Proteome Res. 2012, 11, 2982−2995

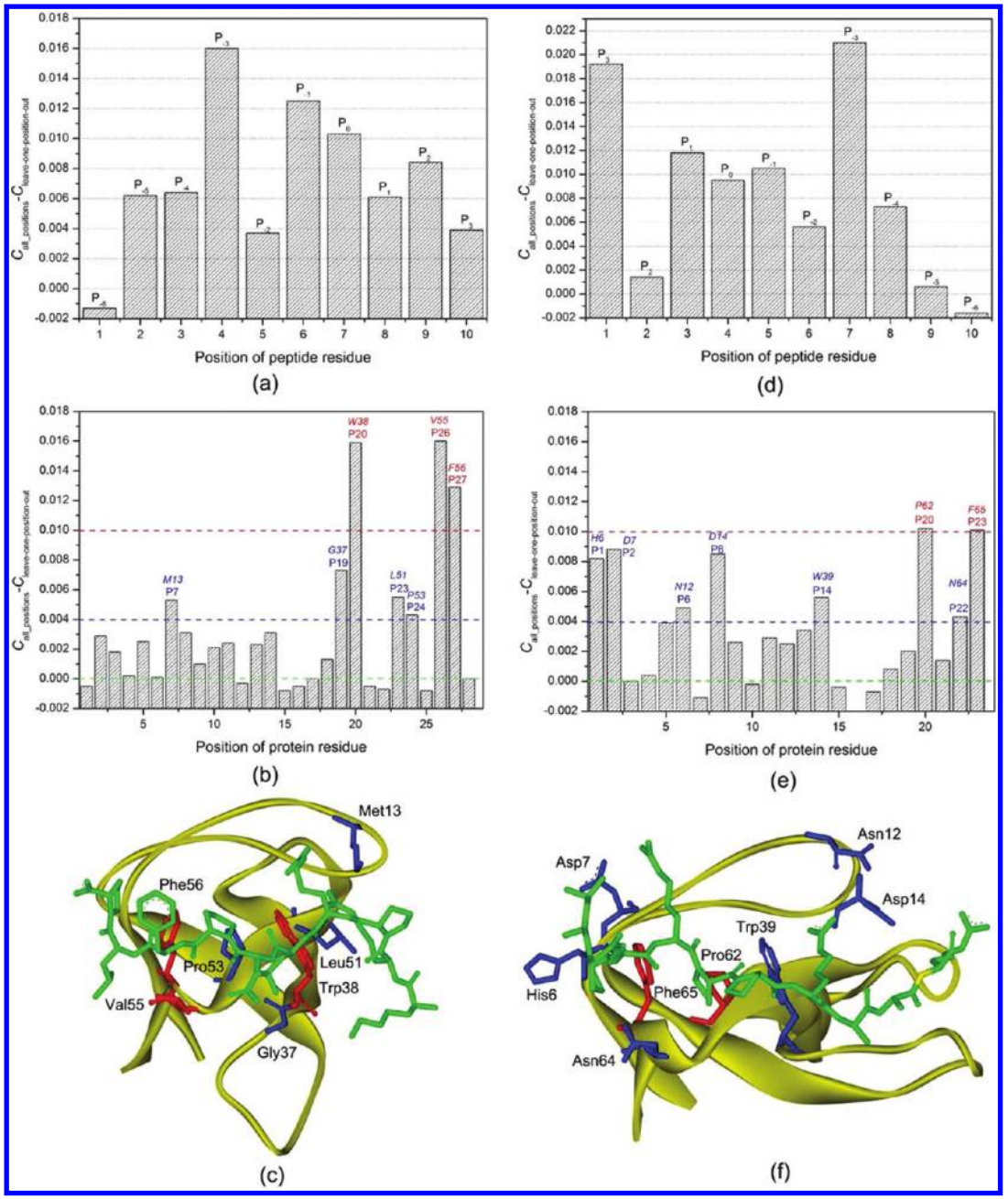


Figure 3. Contributions of the domain-peptide residues to SH3 binding specificity. (a) Changes of the Matthews correlation coefficients (*C*) in the leave-one-position-out cross-validation for the class-I peptides. (b) Changes of the Matthews correlation coefficient in the leave-one-position-out cross-validation for the 28 important class-I SH3 domain positions. (c) Spatial locations of the seven SH3 domain positions that have a change of *C* larger than 0.004 in Boi1. The SH3 domain is shown in strand. The peptide and the domain residues at the important positions are shown in stick. The three residues with a change of *C* larger than 0.01 are colored in red and the other four important residues in blue. (d) Changes of the Matthews correlation coefficients (*C*) in the leave-one-position-out cross-validation for the class-II peptides. (e) Changes of the Matthews correlation coefficient in the leave-one-position-out cross-validation for the 24 important class-II SH3 domain positions. (f) Spatial locations of the seven SH3 domain positions that have a change of *C* larger than 0.004 in Amph. The SH3 domain is shown in strand. The peptide and the domain residues at the important positions are shown in stick. The two residues with a change of *C* larger than 0.01 are colored in red and the other six important residues in blue.

matrix, support vector machine (SVM) can be used to train a unified model to distinguish binders and nonbinders. Here we refined the previous model based on 23 class-I SH3 domains including six additional SH3 domains, Bzz1_1, Bzz1_2, Hck, Hse1, Pex13 and Sho1, and removed Yha2 whose binding data was unreliable. All the new but Hck SH3 domains are from yeast. We included as many yeast SH3 domains as we could because we aimed to predict the SH3-mediated protein−protein interaction network in the yeast proteome. We used radial basis function (RBF) kernel in SVM because it performed better than the other three kernels (linear, polynomial and sigmoid) in our previous work.[20]

The performance of the MIEC-SVM models for the class-I SH3 binding peptides based on different combinations of various MIECs is shown in Table 2. The influence of different k_+ values on predictions was investigated systematically (Figure

dx.doi.org/10.1021/pr3000688 | J. Proteome Res. 2012, 11, 2982−2995

S3 in the Supporting Information) and $k_+ = 4$ was found to be a balanced choice for achieving satisfactory sensitivity, specificity, prediction accuracy and correlation. When all MIEC terms (four types of domain-peptide MIECs, adjacent peptide residue MIECs and peptide residue internal energies) were included, the SVM model (model 12 in Table 2) achieved the best prediction accuracies as shown by the 500 runs of cross-validations (C = 0.659, SE_{test} = 90.4% and SP_{test} = 94.5%), which was significantly superior to our previous model based on 18 SH3 domains (C = 0.532, SE_{test} = 84.2% and SP_{test} = 93.0%). The improvement can be explained by the following reasons: First, additional MIEC terms including the nonpolar contribution to desolvation upon peptide binding and peptide residue internal energies that were not explicitly considered in our previous work were included. Second, some low-quality experimental data were excluded in the present study, for example, the binding peptides are less than nine-residue-long.

Performances of the best unified MIEC-SVM models for the class-I and class-II binding peptides were evaluated further by the area under the ROC curves (Figure S4 in the Supporting Information). The best MIEC-SVM model for the class-I and class-II SH3 domains achieved 0.967 and 0.959 of averaged accuracy from 500 runs of cross-validations.

3. Generalization Capability of the MIEC-SVM Models

To assess the generalization capability of the prediction models, we used a leave-one-domain-out (LODO) cross-validation: the data for one SH3 domain were completely left out, and a MIEC-SVM model was trained using the remaining data of the other domains and tested on the left-out domain. This test provided a stringent assessment because the interaction data of the left-out domain was not used in the training.

First, the generalization capability of the MIEC models for the class-I and class-II binding peptides was assessed separately. A series of k_+ were used in the models and the average sensitivity (SE), specificity (SP), prediction accuracy (Q_+) of the binder class and Matthews correlation coefficient (C) for the left-out SH3 were calculated. The change of the average sensitivity (SE), specificity (SP), prediction accuracy (Q_+) and correlation coefficient versus k_+ is shown in Figure S5 in the Supporting Information. When k_+ was 2, the SVM model had the best prediction accuracy for the binders of the left-out domains but the average sensitivity was not satisfactory. It is clear that when k_+ = 3 or 4, the MIEC-SVM model reached an optimal generalization capability while maintaining a good coverage of the binders (the LODO test at k_+ = 4 is shown in Tables S5 and S6, Supporting Information).

The two universal MIEC-SVM models for the class-I and class-II binding peptides showed satisfactory generalization capabilities (40−50% prediction accuracy for the binding peptides), which is crucial for expanding our predictions to SH3 domains not included in the training data.

4. Structural Insights into the SH3-Peptide Interactions for the Class-I Binding Peptides

To evaluate the contribution of each position to the performance of the SVM model, we exploited a leave-one-position-out cross-validation: first, the MIECs of a peptide position were removed; then, 3-fold cross validations were conducted 500 times to evaluate performance of the trained SVM models; the importance of the peptide position was assessed by the change of the Matthews correlation coefficient C of the model on the test sets (Figure 3a). Note that the MIEC matrices after removing the MIEC terms for different positions had different dimensions. To compare the performance of the MIEC-SVM with different dimensions, we had to use a very large k_+ value to get stable predictions (k_+ = 500 was used here). As seen in Figure 3a, all but P_{-6} peptide positions have positive contributions to the binding specificity. The change of MCC of position P_{-6} is close to zero, suggesting its negligible contribution to SH3-peptide binding specificity. P_{-6} is located at the N-terminal of the peptide, and the residue at this position is conformationally dynamic in the MD simulations (the RMSF (root-mean-square-fluctuation) values for the 10 residues in the boi1 complex are shown as an example in Figure S6 in the Supporting Information). Among the nine positions with positive contributions, P_{-3} is the most important, which is consistent with the experimental observation that P_{-3} is the key position to determine the binding specificity of SH3.[2] Position P_{-1} is the second important position. Structural analysis showed that the side chain of P_{-1} orients toward the binding surface of SH3 and forms close contacts with several important residues in SH3. P_0 is the third important position and only proline was found at this position for all peptides in this study. The understanding of the relatively large change of C of P_0 is not straightforward because the residue at P_0 is conserved and then its contribution should be constant. However, our simulations show that the mutations at other positions still can influence the interactions between SH3 and the residue at P_0. Compared with P_0, the other Pro in the PXXP motif at P_3 position seems less important indicated by a smaller change of C. Pro at P_3 is also conserved in all peptides; however, the interactions between this Pro and SH3 may not be affected significantly by the mutations of other residues because Pro at P_3 is a C-terminal residue. In all positions, P_2 is the fourth important. The residue at P_2 is not completely buried upon binding to SH3, but it still forms effective interactions with several important residues of SH3.

Then we conducted leave-one-position-out cross-validations for the 28 positions in SH3 used to calculate the MIEC matrix. As shown in Figure 3b, the contributions of these SH3 positions are quite different. In all these positions, P20, P26 and P27 are more important than the others. The corresponding residues at these three positions in Boi1 are shown in Figure 3c, and they are Trp38, Val55 and Phe56. Among these three residues, Trp38 and Phe56 are almost conserved across all SH3 domains, and can form extensive hydrophobic and van der Waals interactions with several key residue in peptide, including Pro at P_3, Pro at P_0, residues at P_{-3} and P_{-1}. Val55 is not well conserved across all SH3 domains; however, this residue can form effective interactions with residues at P_{-1} and P_2, and possibly contribute a lot to the specificity of SH3. As shown in Figure 3b, four positions (P7, P19, P23 and P24) are less important than P20, P26 and P27, while more important than the others. The corresponding residues in Boi1 are Met13, Gly37, Leu51 and Pro53. Structural analysis shown in Figure 3c indicates that these four residues are quite important for the peptide binding. Met13 forms strong interactions with the residue at P_{-3}, obviously important for the specificity of SH3; Leu51 usually forms strong van der Wals interactions with the residue at P_{-5}, which can partially determine the specificity of SH3; Pro53 forms effective van der Waals interactions with Pro at P_0, which is essential for the peptide binding; Gly37 in Boi1 cannot form strong interactions with peptide, but in many SH3, the residue (Asp, His, Glu, et al) at this position usually can

dx.doi.org/10.1021/pr3000688 | J. Proteome Res. 2012, 11, 2982−2995

form effective electrostatic interactions with the residue at P_{-1} and partially determine the binding specificity of SH3.

As shown in Figure 3a, in these 28 SH3 positions, 11 positions nearly do not influence the prediction of the SVM model, and seven of them even show a little negative contribution to the model. These seven unimportant positions include P1, P12, P15, P16, P21, P22 and P25, and the corresponding residues at these positions in Boi1 are Val6, Asp30, Glu33, gap, Tyr39, Tyr40 and Ala54 (Figure S7a in the Supporting Information). As shown in Figure S7a, these six residues do not form close contacts with the peptide in the Boi1 complex. Although they can be identified as the important residues using the predefined cutoff in few complexes, they are far from the binding interface in most cases and do not contribute to peptide binding effectively.

5. Structural Insights into the SH3-Peptide Interactions for the Class-II Binding Peptides

The leave-one-position-out cross-validations for the ten positions in class-II binding peptides are shown in Figure 3d. As shown in Figure 3d, for one position, P_{-6}, the change of the Matthews correlation coefficient *C* is close to zero. Therefore, the contribution of this position is ignorable. The position P_{-6} is located at the C-terminal of the peptide. Similar to the residue at position P_{-6} in the N-terminal of the class-I binding peptide, the residue at P_{-6} is very dynamic and not important to the peptide binding.[57] For the other nine positions, two positions, P_3 and P_{-3}, are more important than the others. Position P_3 is at the N-terminal. The structural analysis shows that the residue at P_3 occupies most space taken by Pro at P_3 in class-I binding peptides. The residue at P_3 usually forms strong interactions with SH3, such as His6, Asp7 and Glu9 in Amph SH3, and is very important for the binding specificity of SH3. The position P_{-3}, which usually represents R/K in the PXXPX+ motif (where + refers to a positively charged residue), is well-known as a crucial position to determine the binding specificity of SH3. As shown in Figure 3d, position P_3 is nearly as important as position P_{-3}. This finding is very interesting, because according to the traditional knowledge, position P_{-3} is much more important than P_3 for the specificity of SH3. However, P_3 is a crucial position judged by our predictions.

The leave-one-position-out cross-validations for the important positions in SH3 which recognizes the class-II binding peptides were conducted and shown in Figure 3e. As shown in Figure 3e, two positions, P20 and *P*23, are more important than the others. As an example, the residues at positions P20 (Pro62) and *P*23 (Phe65) in Amph are shown in Figure 3f. Pro62 at P20 forms strong van der Waals interactions with the conserved Pro at P_{-1} and the side chain of the residue at P_0; Phe65 at *P*23 forms strong van der Waals interactions with the conserved Pro at P_2 and the side chain of the residue at P_0. The other important positions with the change of *C* between 0.004 and 0.01 include P1, P2, P6, P8, P14 and P22, and the corresponding residues at these six positions in Amph are His6, Asp7, Asn12, Asp14, Trp39 and Asn64, respectively. In the residues at these six positions, His6 and Asp7 at positions P1 and P2 form close contacts with the residue at P_3 of the peptide, and should be quite important to the binding specificity of SH3; Asn64 at P22 forms effective interactions with the two conserved Pro residues at P_2 and P_0 of the peptide; Trp39 at P14 forms strong van der Waals interactions with Pro at P_0 and the residue at P_{-3} of the peptide; Asn12 and Asp14 at positions P6 and P8 have strong interactions with the side chain of Arg at P_{-3} of the peptide, which are essential to determine the binding specificity of SH3. In all the predefined 23 positions in protein, five of them show negative contributions to the predictions, including P7, P10, P15, P16 and P17. As shown in Figure S7b in the Supporting Information, the corresponding residues in Amph SH3 are relatively far from the binding interface.

6. Do Class-I and Class-II Binding Peptides Have the Same Recognition Code?

By comparing the important positions shown in Figure 3b and e, we can identify the difference of the interaction patterns between the class-I and class-II binding peptides. Here, the structure of Lsb3 was used as an example because it can recognize two types of binding peptides. Guided by Figure 3a and d, for Lsb3 the important residues for recognizing the class-I binding peptides are Glu16, Asp39, Trp40, Ile51, Pro53, Asn55 and Tyr56 (the left figure of Figure 2c), and those for recognizing the class-II binding peptides are Tyr10, Ser11, Glu16, Gly18, Trp40, Pro53, Asn55 and Tyr56 (the middle figure of Figure 2c). It is easy to observe that the important residues for these two types of binding peptides are not completely identical, and five residues, including Tyr10, Ser11, Gly18, Asp39 and Ile51, are different (the right figure of Figure 2c). So it is obvious that the SH3 recognition codes for these two types of binding peptides are not quite identical even they almost occupy the same binding pocket.

7. SH3-Mediated Protein−Protein Interaction Network in the Yeast Proteome

In our data set, 13 yeast SH3 domains can recognize the class-I binding peptides and 8 yeast SH3 domains can recognize the class-II binding peptides. In the yeast proteome, 29 SH3 domains can be found; however, for eight SH3 domains, including Bem1_2, Cdc25, Sla1_1, Sla1_2, Bud14, Sdc25, Cyk3 and Hof1, no binding data are available according to Tong's data.[25] Moreover, according to Tong's data,[25] the binding peptides for Bem1_1, Fus1 and Abp1 do not have usual PXXP binding motifs. Therefore, in our calculations, only 18 SH3 domains in yeast were finally included.

Since we already got two universal MIEC-SVM models for SH3, it is straightforward to use these two models to screen the yeast proteome, find the possible binding peptide segments of these SH3 domains, and finally construct the SH3-mediated protein−protein interaction network in yeast. First, we screened the yeast proteome using relaxed consensus motifs, such as (R/K)XXPXXP and PXXPXX(R/K), to select all yeast peptides with these motifs; then, these peptides were submitted to a VM and MM optimization; then MM/GBSA free energy decomposition was used to generate the MIEC vector for each peptide; finally the MIEC vector for each peptide was normalized and used as the input for the MIEC-SVM model, and the potential binding peptides could be identified. The number of potential binders and nonbinders predicted by the SVM models is shown in Table S7 and S8 in the Supporting Information. Generally, when the weight parameter, k_+, was set to 3, 4.85% of peptides which have the binding motifs of the class-I binding peptides and 4.44% of peptides which have the binding motifs of the class-II binding peptides were predicted as true binders. When k_+ is equal to 4, more peptides were predicted as true binders; however, the rate of false positives should increase obviously at the mean time.

Then based on the predictions of the MIEC-SVM models, we constructed the SH3-mediated protein−protein interaction

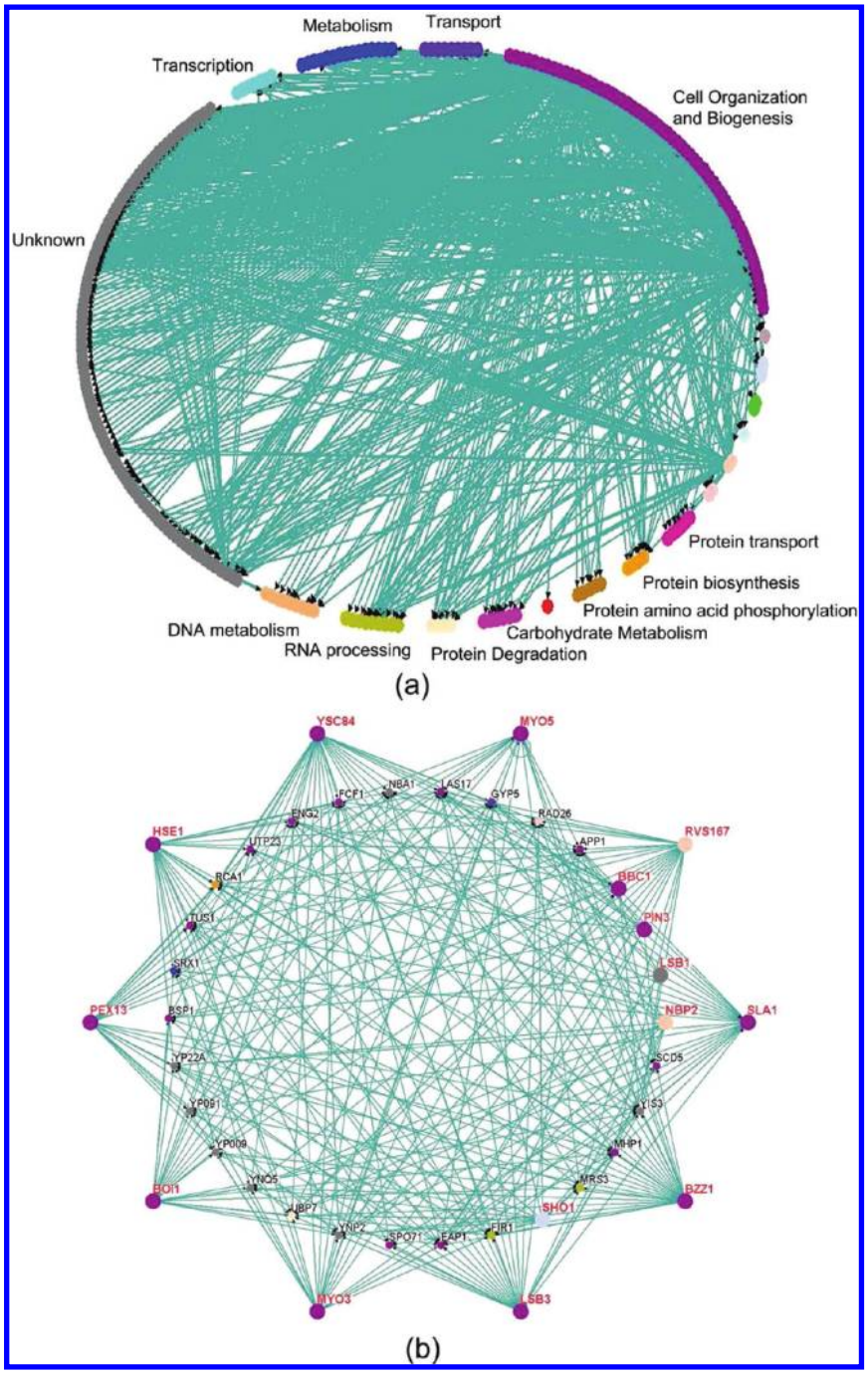


Figure 4. (a) SH3-mediated protein−protein interaction network mediated by the MIEC-SVM models; the biological process annotation associated with the SH3 domain ligands were defined by Gene Ontology (GO). (b) Subnetwork with the nodes that can form interactions with at least seven SH3 domains; the studied SH3 domains are shown as big dots. The figures were generated using the Osprey network visualization program (ver 1.2.0).[71]

network based on the following assumption: if one ten-residue-long peptide segment in one protein is binder of a given SH3, then this protein was assumed to interact with this SH3 domain. The SH3-medidated protein−protein interaction network is shown in Figure 4a. In total, the network has 249 nodes (proteins) and 745 edges (interactions), with an average of 43.8 interactions per SH3 domain and an average of 2.99 interactions per node. The number of interactions per SH3 domain is larger than that reported by the ORFeome or gDNA screening reported by Tonikian (11 interactions per SH3 in the ORFeome screening and 22.6 interactions per SH3 domain in the gDNA screening).[13] It is reasonable because the existence of a binding segment in one protein of a SH3 domain cannot guarantee the recognition between this protein and the corresponding SH3.

As one approach to assessing the biological relevance of the protein−protein interaction network identified by the MIEC-SVM models, we examined the biological process annotation associated with the SH3 ligands defined by Gene Ontology (GO).[68] It can be found that the most abundant biological processes are cell organization and biogenesis (64/249 = 25.7%). Actually these abundant biological processes are known

dx.doi.org/10.1021/pr3000688 | J. Proteome Res. 2012, 11, 2982−2995

to be associated with the yeast SH3 domain biology. Then, we extracted the subnetwork with the nodes which can form interactions with at least seven SH3 domains (Figure 4b). In total, 25 proteins can recognize at least seven SH3 domains. We believe that those proteins shown in Figure 4b may have more probability to be the ligands of the SH3 domains. In these 25 proteins, Las17 is known as the ligand for many yeast SH3 domains.[25] Moreover, in these 25 proteins, two of them, Scd5 and Bsp1, are important for normal cortical actin organization and endocytosis and may interact with the yeast SH3 domains.[13] Certainly, the protein−protein interaction network predicted by us may contain a lot of false positives and false negatives and should be validated by experiments. However, we believe that our predictions give a good starting point for experiments.

8. Validating the Prediction Accuracies of the MIEC-SVM Models

a. Comparison with the Experimental Data from WISE. The MIEC-SVM models were developed based on noisy and limited data, and therefore it is necessary to evaluate the prediction performance of the MIEC-SVM mode by experiments. Recently, Tonikian et al. used the WISE approach to identify possible binding peptides of 26 SH3 domains.[13] In total, 295 possible peptides showed a positive signal with at least one SH3 domain. The experimental results given by WISE can be used as a validation set for the MIEC-SVM models. However, in Tonikian's work, all peptides are 16-residue-long and have the X6-PXXP-X6 consensus motif; in this study, all peptides in the data set are only 10-residue-long. Moreover, we found that many peptides identified by WISE have multiple 10-residue peptides which have a (R/K)XXPXXP or PXXPXX(R/K) motif. Here, in order to compare our predictions with the WISE's data directly, we made the following assumption: if one ten-residue-long segment in a given 16-residue-long peptide used by WISE was predicted as a binding peptide of one SH3 domain, this 16-residue-long peptide was then assumed to be a true binder of this SH3 domain. Moreover, it should be noted that the WISE approach is semiquantitative, and we need to define an arbitrary cutoff for the intensity of the WISE signal to distinguish binder or nonbinder. Here, when a cutoff of 200 was used 644 domain-peptide interactions could be identified. Among 644 domain-peptide interactions, 505 were mediated by the SH3 domains used in our data set. The comparison between our predictions and the experimental data by WISE was shown in Tables S9 and S10 in the Supporting Information. As shown in Table S9, when k_+ was set to 3, 33.1% (167/505) of the true interactions and 94.7% (4273/4510) of the true noninteractions could be correctly predicted by the MIEC-SVM models. The good performance of the SVM model was further validated by the good prediction accuracy (41.3%) for the true binders. Considering the unbalanced nature (binders/nonbinders = 8.93%) of the data given by WISE, our predictions are really exciting. If we used the same binder to nonbinder ratio of 0.0893, our prediction accuracy for the binders can increase to $167/(167+(0.0893 \times 20) \times 237) = 55.7\%$. When k_+ was set to 4 (Table S10 in the Supporting Information), more true interactions could be correctly predicted (189/505 = 37.4%); however, more true noninteractions (364) were predicted as true interactions, and then the prediction accuracy of the true interactions decreased (34.2%).

More encouraging, among these 167 true interactions predicted by the MIEC-SVM model, only 68 interactions were mediated by the binding peptides used in our data set for training the MIEC-SVM model. That is to say, the other 99 true binding peptides which were not used in our data set could be identified as the true binders by the SVM models. Therefore, even the model was constructed based on the limited data, it still has good prediction capability to find new possible binders.

As a more stringent test, we removed all 10-residue-long peptides found in the WISE data from the data set for training the MIEC-SVM models, and then retrained the MIEC-SVM models and made predictions for all peptides selected from the yeast proteome. The predictions for the WISE data based on the new models were shown in Tables S11 and S12 in the Supporting Information. According to our predictions, when $k_+ = 3$, 24.2% of the true binders determined by WISE could be identified as the true binders and 95.9% of the true nonbinders determined by WISE could be correctly predicted. Moreover, the prediction accuracy for the binder class is about 40%. Our predictions for the WISE data are really encouraging because the MIEC-SVM can successfully identify 24.2% of new SH3-peptide interactions which were not included in the data set. Therefore, we believe that our models can be used to screen proteome of any species and found new binding partners of the SH3 domain family.

b. Predictions of the MIEC-SVM Model for the Gold-Standard Set. To evaluate the prediction capability of the MIEC-SVM model further, we checked the performance of the MIEC-SVM model for a gold-standard set. The gold-standard set manually compiled by Tonikian et al. includes a total of 41 nonredundant interactions known to be mediated by some SH3 domains.[13] Each interaction in the gold-standard set was supported by multiple experiments reported in one or more focused studies. Among these 41 interactions, 35 are mediated by the SH3 domains used in our data set. The predictions of the MIEC-SVM model for these 35 interactions are shown in Table S13 in the Supporting Information. Among these 35 interactions, the binding partners of the SH3 domains for 22 interactions do have the 10-residue-long peptides segments which are the true binders of SH3 predicted by the MIEC-SVM model. Therefore, according to our predictions, most proteins known to be the binding partners of SH3 (62.9% = 22/35) could be successfully identified by the MIEC-SVM models.

CONCLUSION

In this study, we applied molecular interaction energy components (MIECs) derived from free energy decomposition analysis to describe the SH3-peptide binding interface quantitatively. On the basis of the MIEC matrix and machine learning technique, we have developed two generic theoretical models to interpret the binding specificity of the SH3 domain family, which can recognize two classes of peptides. Both of these two generic models show very satisfactory prediction accuracies.

It is well-known that a lot of theoretical prediction models have been reported to investigate the binding specificity of modular domains. However, most of them do not or only slightly rely on domain-peptide complex structures, and they built prediction models based on the domain-peptide residue−residue contacts. Obviously, domain-peptide interaction pattern cannot be well characterized by the physiochemical properties of contact pairs. Furthermore, the contact matrix used for training the model is sparse while the MIEC matrix is a fully

filled matrix because the interactions between residue pairs are represented by energy terms, regardless of amino acid type. For training classifiers, this MIEC matrix is more informative and less prone to noise or error than the contact matrix.

In this study, the contribution of each peptide or domain position to the prediction accuracy of the MIEC-SVM models was analyzed, and then the contributions were mapped to the complex structures and the recognition codes for these two classes of SH3-binding peptides were determined. An interesting finding is that the class-I and class-II binding peptides have different recognition codes even they occupy the same binding site of SH3. For example, we mentioned that for Lsb3, the important residues for these two types of binding peptides are not completely identical, and five residues, including Tyr10, Ser11, Gly18, Asp39 and Ile51, are different. In these five residues, only Tyr10 interacts with one Pro residue in PXXP core of the peptide, and the other four residues interact with these residues outside the PXXP core. Actually Tyr10 can form effective interactions with the residue at P_3 position of the class-II binding peptides. Therefore, it is obvious that although the peptide-domain interactions are very similar in both peptide orientations and the interactions between SH3 and those residues outside the proline-rich core of the peptide determine the orientation of the binding peptide.

On the basis of the two generic MIEC-SVM models, we virtually screened the yeast proteome and predicted the potential binding partners of the studied SH3 domains. We compared our predictions with the recent experimental results given by WISE. If all 10-residue-long peptides found in the WISE data were removed from the data set for training the MIEC-SVM models, the prediction accuracies for the binder and nonbinder classes are 39.6 and 91.9%, respectively. So it is obvious that the MIEC-SVM models are really efficient to find new binding partners in the yeast proteome. But one thing we need to mention is that only 24.2% of true binders were identified in both of the MIEC-SVM screening and the WISE screening. Three major reasons could be used to explain the difference between our predictions and WISE. First, the peptides used in the WISE screening are 16-residue-long and those used in our predictions are only 10-residue-long. Second, 35.3% (104/295) of the peptides used in WISE do not have the PXXP binding motifs used in the data set for training the MIEC-SVM models, and therefore all these peptides which do not have the PXXP binding motifs were predicted as nonbinders by the MIEC-SVM models. Third, the data for training the MIEC-SVM models are not reliable enough. For example, for Sho1 and Hse1, only five and eight true binders with nine residues could be found, and for each binder five peptides were randomly generated by attaching amino acids at the N terminal. Certainly the randomly generated binders from nine-residue-long peptides are not reliable enough. As shown in Table S11 in the Supporting Information, the predictions of the WISE data for the Sho1 and Hse1 SH3 domains are poor. Although only 24.2% true binders from WISE were successfully identified, it is undoubted that our predictions are still very promising because no experimental or theoretical method can identify all domain-peptide or protein−protein interactions with high accuracies considering that the domain-peptide or protein−protein interactions are usually weak and transit. We believe that the different methods may have complementary features, and our predictions are the reliable supplement to experiments.

■ ASSOCIATED CONTENT

Supporting Information

Supplemental Tables S1−S13 and Figures S1−S7. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tingjun Hou, Institute of Functional Nano & Soft Materials (FUNSOM) and Jiangsu Key Laboratory for Carbon-Based Functional Materials & Devices, Soochow University, Suzhou, Jiangsu 215123, P. R. China. Phone: 8610-65882039. E-mail: tjhou@suda.edu.cn or tingjunhou@hotmail.com. Wei Wang, Department of Chemistry and Biochemistry, 9500 Gilman Drive, University of California at San Diego, La Jolla, CA 92093-0359. E-mail: wei-wang@ucsd.edu. Phone: +1(0)-858-822-4240.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was supported by the National Science Foundation of China (21173156 to T.H.), the National Basic Research Program of China (973 program, 2012CB932600 to T.H.), the National Institutes of Health Grant (R01GM085188 to W.W.) and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

■ REFERENCES

(1) Pawson, T.; Nash, P. Assembly of cell regulatory systems through protein interaction domains. *Science* **2003**, *300* (5618), 445−452.
(2) Mayer, B. J. SH3 domains: complexity in moderation. *J. Cell Sci.* **2001**, *114* (7), 1253−1263.
(3) Musacchio, A.; Gibson, T.; Lehto, V. P.; Saraste, M. Sh3 - an abundant protein domain in search of a function. *Febs Lett.* **1992**, *307* (1), 55−61.
(4) Yu, H. T.; Rosen, M. K.; Shin, T. B.; Seideldugan, C.; Brugge, J. S.; Schreiber, S. L. Solution structure of the Sh3 domain of Src and identification of its ligand-binding site. *Science* **1992**, *258* (5088), 1665−1668.
(5) Musacchio, A.; Saraste, M.; Wilmanns, M. High-resolution crystal-structures of tyrosine kinase Sh3 domains complexed with proline-rich peptides. *Nat. Struct. Biol.* **1994**, *1* (8), 546−551.
(6) Lim, W. A.; Richards, F. M.; Fox, R. O. Structural determinants of peptide-binding orientation and of sequence specificity in Sh3 domains. *Nature* **1994**, *372* (6504), 375−379.
(7) Sudol, M. From Src Homology domains to other signaling modules: proposal of the 'protein recognition code'. *Oncogene* **1998**, *17* (11), 1469−1474.
(8) Feng, S. B.; Chen, J. K.; Yu, H. T.; Simon, J. A.; Schreiber, S. L. 2 Binding orientations for peptides to the Src Sh3 domain - development of a general-model for Sh3-ligand interactions. *Science* **1994**, *266* (5188), 1241−1247.
(9) Sparks, A. B.; Rider, J. E.; Hoffman, N. G.; Fowlkes, D. M.; Quilliam, L. A.; Kay, B. K. Distinct ligand preferences of Src homology 3 domains from Src, Yes, Abl, Cortactin, p53bp2, PLC gamma, Crk, and Grb2. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93* (4), 1540−1544.
(10) Landgraf, C.; Panni, S.; Montecchi-Palazzi, L.; Castagnoli, L.; Schneider-Mergener, J.; Volkmer-Engert, R.; Cesareni, G. Protein interaction networks by proteome peptide scanning. *PLoS Biol.* **2004**, *2* (1), 94−103.
(11) Rickles, R. J.; Botfield, M. C.; Zhou, X. M.; Henry, P. A.; Brugge, J. S.; Zoller, M. J. Phage display selection of ligand residues important for Src-homology-3 domain binding-specificity. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92* (24), 10909−10913.

(12) Rickles, R. J.; Botfield, M. C.; Weng, Z. G.; Taylor, J. A.; Green, O. M.; Brugge, J. S.; Zoller, M. J. Identification of Src, Fyn, Lyn, Pi3k and Abl Sh3 domain ligands using phage display libraries. *EMBO J.* **1994**, *13* (23), 5598−5604.
(13) Tonikian, R.; Xin, X. F.; Toret, C. P.; Gfeller, D.; Landgraf, C.; Panni, S.; Paoluzi, S.; Castagnoli, L.; Currell, B.; Seshagiri, S.; Yu, H. Y.; Winsor, B.; Vidal, M.; Gerstein, M. B.; Bader, G. D.; Volkmer, R.; Cesareni, G.; Drubin, D. G.; Kim, P. M.; Sidhu, S. S.; Boone, C. Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. *PLoS Biol.* **2009**, *7* (10), 1.
(14) Brannetti, B.; Via, A.; Cestra, G.; Cesareni, G.; Citterich, M. H. SH3-SPOT: An algorithm to predict preferred ligands to different members of the SH3 gene family. *J. Mol. Biol.* **2000**, *298* (2), 313−328.
(15) Ferraro, E.; Via, A.; Ausiello, G.; Helmer-Citterich, M. A novel structure-based encoding for machine-learning applied to the inference of SH3 domain specificity. *Bioinformatics* **2006**, *22* (19), 2333−2339.
(16) Lehrach, W. P.; Husmeier, D.; Williams, C. K. I. A regularized discriminative model for the prediction of protein-peptide interactions. *Bioinformatics* **2006**, *22* (5), 532−540.
(17) McLaughlin, W. A.; Hou, T. J.; Wang, W. Prediction of binding sites of peptide recognition domains: An application on Grb2 and SAP SH2 domains. *J. Mol. Biol.* **2006**, *357* (4), 1322−1334.
(18) Obenauer, J. C.; Cantley, L. C.; Yaffe, M. B. Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.* **2003**, *31* (13), 3635−3641.
(19) Hou, T. J.; Zhang, W.; Case, D. A.; Wang, W. Characterization of domain-peptide interaction interface: A case study on the amphiphysin-1 SH3 domain. *J. Mol. Biol.* **2008**, *376* (4), 1201−1214.
(20) Hou, T. J.; Xu, Z.; Zhang, W.; McLaughlin, W. A.; Case, D. A.; Xu, Y.; Wang, W. Characterization of domain-peptide interaction interface: a generic structure-based model to decipher the binding specificity of SH3 domains. *Mol. Cell. Proteomics* **2009**, *8* (4), 639−649.
(21) Hou, T. J.; Chen, K.; McLaughlin, W. A.; Lu, B. Z.; Wang, W. Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain. *PLoS Comput. Biol.* **2006**, *2* (1), 46−55.
(22) Zhang, L.; Shao, C.; Zheng, D. X.; Gao, Y. H. An integrated machine learning system to computationally screen protein databases for protein binding peptide ligands. *Mol. Cell. Proteomics* **2006**, *5* (7), 1224−1232.
(23) Wunderlich, Z.; Mirny, L. A. Using genome-wide measurements for computational prediction of SH2-peptide interactions. *Nucleic Acids Res.* **2009**, *37* (14), 4629−41.
(24) Xu, Z.; Hou, T. J.; Li, N.; Xu, Y.; Wang, W. Proteome-wide detection of Abl1 SH3 binding peptides by integrating computational prediction and peptide microarray. *Mol. Cell. Proteomics* **2012**, *11* (1), O111.010389.
(25) Tong, A. H. Y.; Drees, B.; Nardelli, G.; Bader, G. D.; Brannetti, B.; Castagnoli, L.; Evangelista, M.; Ferracuti, S.; Nelson, B.; Paoluzi, S.; Quondam, M.; Zucconi, A.; Hogue, C. W. V.; Fields, S.; Boone, C.; Cesareni, G. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **2002**, *295* (5553), 321−324.
(26) Bairoch, A.; Consortium, U.; Bougueleret, L.; Altairac, S.; Amendolia, V.; Auchincloss, A.; Argoud-Puy, G.; Axelsen, K.; Baratin, D.; Blatter, M. C.; Boeckmann, B.; Bolleman, J.; Bollondi, L.; Boutet, E.; Quintaje, S. B.; Breuza, L.; Bridge, A.; Decastro, E.; Ciapina, L.; Coral, D.; Coudert, E.; Cusin, I.; Delbard, G.; Dornevil, D.; Roggli, P. D.; Duvaud, S.; Estreicher, A.; Famiglietti, L.; Feuermann, M.; Gehant, S.; Farriol-Mathis, N.; Ferro, S.; Gasteiger, E.; Gateau, A.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Hulo, N.; James, J.; Jimenez, S.; Jungo, F.; Junker, V.; Kappler, T.; Keller, G.; Lachaize, C.; Lane-Guermonprez, L.; Langendijk-Genevaux, P.; Lara, V.; Lemercier, P.; Le Saux, V.; Lieberherr, D.; Lima, T. D.; Mangold, V.; Martin, X.; Masson, P.; Michoud, K.; Moinat, M.; Morgat, A.; Mottaz, A.; Paesano, S.; Pedruzzi, I.; Phan, I.; Pilbout, S.; Pillet, V.; Poux, S.; Pozzato, M.; Redaschi, N.; Reynaud, S.; Rivoire, C.; Roechert, B.; Schneider, M.; Sigrist, C.; Sonesson, K.; Staehli, S.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Veuthey, A. L.; Yip, L.; Zuletta, L.; Apweiler, R.; Alam-Faruque, Y.; Antunes, R.; Barrell, D.; Binns, D.; Bower, L.; Browne, P.; Chan, W. M.; Dimmer, E.; Eberhardt, R.; Fedotov, A.; Foulger, R.; Garavelli, J.; Golin, R.; Horne, A.; Huntley, R.; Jacobsen, J.; Kleen, M.; Kersey, P.; Laiho, K.; Leinonen, R.; Legge, D.; Lin, Q.; Magrane, M.; Martin, M. J.; O'Donovan, C.; Orchard, S.; O'Rourke, J.; Patient, S.; Pruess, M.; Sitnov, A.; Stanley, E.; Corbett, M.; di Martino, G.; Donnelly, M.; Luo, J.; van Rensburg, P.; Wu, C.; Arighi, C.; Arminski, L.; Barker, W.; Chen, Y. X.; Hu, Z. Z.; Hua, H. K.; Huang, H. Z.; Mazumder, R.; McGarvey, P.; Natale, D. A.; Nikolskaya, A.; Petrova, N.; Suzek, B. E.; Vasudevan, S.; Vinayaka, C. R.; Yeh, L. S.; Zhang, J. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.* **2009**, *37*, D169−D174.
(27) Cestra, G.; Castagnoli, L.; Dente, L.; Minenkova, O.; Petrelli, A.; Migone, N.; Hoffmuller, U.; Schneider-Mergener, J.; Cesareni, G. The SH3 domains of endophilin and amphiphysin bind to the proline-rich region of synaptojanin 1 at distinct sites that display an unconventional binding specificity. *J. Biol. Chem.* **1999**, *274* (45), 32001−32007.
(28) Pisabarro, M. T.; Serrano, L.; Wilmanns, M. Crystal structure of the Abl-SH3 domain complexed with a designed high-affinity peptide ligand: Implications for SH3-ligand interactions. *J. Mol. Biol.* **1998**, *281* (3), 513−521.
(29) Feng, S. B.; Kasahara, C.; Rickles, R. J.; Schreiber, S. L. Specific interactions outside the proline-rich core of two classes of Src homology 3 ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92* (26), 12408−12415.
(30) Schmidt, H.; Hoffmann, S.; Tran, T.; Stoldt, M.; Stangler, T.; Wiesehan, K.; Willbold, D. Solution structure of a Hck SH3 domain ligand complex reveals novel interaction modes. *J. Mol. Biol.* **2007**, *365* (5), 1517−1532.
(31) Kursula, P.; Kursula, I.; Pinotsis, N.; Song, Y. H.; Lehmann, F.; Zou, P.; Wilmanns, M., Structural Genomics of Yeast SH3 Domains. *To be published.*
(32) Wittekind, M.; Mapelli, C.; Lee, V.; Goldfarb, V.; Friedrichs, M. S.; Meyers, C. A.; Mueller, L. Solution structure of the Grb2 N-terminal SH3 domain complexed with a ten-residue peptide derived from SOS: Direct refinement against NOEs, J-couplings and H-1 and C-13 chemical shifts. *J. Mol. Biol.* **1997**, *267* (4), 933−952.
(33) Gushchina, L. V.; Gabdoulkhakov, A. G.; Nikulin, A. D.; Nikonov, S. V.; Filimonov, V. V., Structure of SH3 chimera with a type II ligand linked to the chain C-terminal. *To be published.*
(34) Andreotti, A. H.; Bunnell, S. C.; Feng, S.; Berg, L. J.; Schreiber, S. L. Regulatory intramolecular association in a tyrosine kinase of the Tec family. *Nature* **1997**, *385* (6611), 93−97.
(35) Bauer, F.; Schweimer, K.; Meiselbach, H.; Hoffmann, S.; Rosch, P.; Sticht, H. Structural characterization of Lyn-SH3 domain in complex with a herpesviral protein reveals an extended recognition motif that enhances binding affinity. *Protein Sci.* **2005**, *14* (10), 2487−2498.
(36) Kursula, P.; Kursula, I.; Lehmann, F.; Song, Y. H.; Wilmanns, M. Crystal structure of the SH3 domain from S. cerevisiae Myo3. To be published.
(37) Gonfoni, S.; Kursula, I.; Sacco, R.; Cesareni, G.; Wilmanns, M. Yeast Myo5 SH3 domain, tetragonal crystal form. To be published.
(38) Kursula, P.; Kursula, I.; Zou, P.; Lehmann, F.; Song, Y. H.; Wilmanns, M. Structural analysis of the yeast SH3 domain proteome. To be published.
(39) Liang, J.; Chen, J. K.; Schreiber, S. L.; Clardy, J. Crystal structure of PI3K SH3 domain at 2.0 angstrom resolution. *J. Mol. Biol.* **1996**, *257* (3), 632−643.
(40) He, Y.; Hicke, L.; Radhakrishnan, I. Structural basis for ubiquitin recognition by SH3 domains. *J. Mol. Biol.* **2007**, *373* (1), 190−196.
(41) Martin-Garcia, J. M.; Luque, I.; Mateo, P. L.; Ruiz-Sanz, J.; Camara-Artigas, A. Crystallographic structure of the SH3 domain of the human c-Yes tyrosine kinase: Loop flexibility and amyloid aggregation. *Febs Lett.* **2007**, *581* (9), 1701−1706.

(42) Kursula, P.; Lehmann, F.; Song, Y. H.; Wilmanns, M., Crystal structure of the SH3 domain from a S. cerevisiae hypothetical 40.4 kDa protein at 1.39 A resolution. To be published.
(43) Douangamath, A.; Filipp, F. V.; Klein, A. T. J.; Barnett, P.; Zou, P. J.; Voorn-Brouwer, T.; Vega, M. C.; Mayans, O. M.; Sattler, M.; Distel, B.; Wilmanns, M. Topography for independent binding of alpha-helical and PPII-helical ligands to a peroxisomal SH3 domain. *Mol. Cell* **2002**, *10* (5), 1007−1017.
(44) *Discovery Studio 2.5 Guide*; Accelrys Inc.: San Diego, 2009; http://www.accelrys.com.
(45) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4* (2), 187−217.
(46) Eisenberg, D.; Luthy, R.; Bowie, J. U. VERIFY3D: Assessment of protein models with three-dimensional profiles. *Macromol. Crystallogr., Part B* **1997**, *277*, 396−404.
(47) Xiang, Z. X.; Honig, B. Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.* **2001**, *311* (2), 421−430.
(48) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26* (16), 1668−1688.
(49) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24* (16), 1999−2012.
(50) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald - an N.Log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98* (12), 10089−10092.
(51) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical-integration of cartesian equations of motion of a system with constraints - molecular-dynamics of N-alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327−341.
(52) Wu, X. D.; Knudsen, B.; Feller, S. M.; Zheng, J.; Sali, A.; Cowburn, D.; Hanafusa, H.; Kuriyan, J. Structural basis for the specific interaction of lysine-containing proline-rich peptides with the N-terminal Sh3 domain of C-Crk. *Structure* **1995**, *3* (2), 215−226.
(53) Deng, L.; Velikovsky, C. A.; Swaminathan, C. P.; Cho, S. W.; Mariuzza, R. A. Structural basis for recognition of the T cell adaptor protein SLP-76 by the SH3 domain of phospholipase C gamma 1. *J. Mol. Biol.* **2005**, *352* (1), 1−10.
(54) Hashimoto, S.; Hiroso, M.; Hashimoto, A.; Morishige, M.; Yamada, A.; Hosaka, H.; Akagi, K. I.; Ogawa, E.; Oneyama, C.; Agatsuma, T.; Okada, M.; Kobayashi, H.; Wada, H.; Nakano, H.; Ikegami, T.; Nakagawa, A.; Sabe, H. Targeting AMAP1 and cortactin binding bearing an atypical src horology 3/proline interface for prevention of breast cancer invasion and metastasis. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103* (18), 7036−7041.
(55) Gorina, S.; Pavletich, N. P. Structure of the p53 tumor suppressor bound to the ankyrin and SH3 domains of 53BP2. *Science* **1996**, *274* (5289), 1001−1005.
(56) Sato, M.; Koshiba, S.; Inoue, M.; Kigawa, T.; Yokoyama, S., Solution structures of the SH3 domain of human SH3-containing GRB2-like protein 2. To be published.
(57) Hou, T. J.; McLaughlin, W.; Lu, B.; Chen, K.; Wang, W. Prediction of binding affinities between the human amphiphysin-1 SH3 domain and its peptide ligands using homology modeling, molecular dynamics and molecular field analysis. *J. Proteome Res.* **2006**, *5* (1), 32−43.
(58) Ohnishi, S.; Kigawa, T.; Koshiba, S.; Inoue, M.; Yokoyama, S., Solution structure of the SH3 domain of the mouse hypothetical protein SH3RF2. To be published.
(59) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins-Struct. Function Bioinform.* **2004**, *55* (2), 383−394.
(60) Gohlke, H.; Kiel, C.; Case, D. A. Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RaIGDS complexes. *J. Mol. Biol.* **2003**, *330* (4), 891−913.
(61) Hou, T. J.; Zhang, W.; Wang, J.; Wang, W. Predicting drug resistance of the HIV-1 protease using molecular interaction energy components. *Proteins-Struct. Function Bioinform.* **2009**, *74* (4), 837−846.
(62) Hou, T. J.; Wang, J.; Li, Y. Y.; Wang, W. Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. *J. Comput. Chem.* **2011**, *32* (5), 866−877.
(63) Hou, T. J.; Wang, J.; Li, Y. Y.; Wang, W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* **2011**, *51* (1), 69−82.
(64) Weiser, J.; Shenkin, P. S.; Still, W. C. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J. Comput. Chem.* **1999**, *20* (2), 217−230.
(65) Vapnik, V.; Chervonenkis, A. *Theory of Pattern Recognition*; Nauka: Moscow, 1971.
(66) Ivanciuc, O. Applications of support vector machines in chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Cundari, T. R., Boyd, D. B., Eds.; VCH: New York, 2007; Vol. *23*, pp 291−400.
(67) Fan, R. E.; Chen, P. H.; Lin, C. J. Working set selection using the second order information for training SVM. *J. Machine Learning Res.* **2005**, *6*, 1889−1918.
(68) Harris, M. A.; Clark, J.; Ireland, A.; Lomax, J.; Ashburner, M.; Foulger, R.; Eilbeck, K.; Lewis, S.; Marshall, B.; Mungall, C.; Richter, J.; Rubin, G. M.; Blake, J. A.; Bult, C.; Dolan, M.; Drabkin, H.; Eppig, J. T.; Hill, D. P.; Ni, L.; Ringwald, M.; Balakrishnan, R.; Cherry, J. M.; Christie, K. R.; Costanzo, M. C.; Dwight, S. S.; Engel, S.; Fisk, D. G.; Hirschman, J. E.; Hong, E. L.; Nash, R. S.; Sethuraman, A.; Theesfeld, C. L.; Botstein, D.; Dolinski, K.; Feierbach, B.; Berardini, T.; Mundodi, S.; Rhee, S. Y.; Apweiler, R.; Barrell, D.; Camon, E.; Dimmer, E.; Lee, V.; Chisholm, R.; Gaudet, P.; Kibbe, W.; Kishore, R.; Schwarz, E. M.; Sternberg, P.; Gwinn, M.; Hannick, L.; Wortman, J.; Berriman, M.; Wood, V.; de la Cruz, N.; Tonellato, P.; Jaiswal, P.; Seigfried, T.; White, R.; Consortium, G. O. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **2004**, *32*, D258−D261.
(69) Thompson, J. D.; Gibson, T. J.; Plewniak, F.; Jeanmougin, F.; Higgins, D. G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **1997**, *25* (24), 4876−4882.
(70) Clamp, M.; Cuff, J.; Searle, S. M.; Barton, G. J. The Jalview Java alignment editor. *Bioinformatics* **2004**, *20* (3), 426−427.
(71) Breitkreutz, B. J.; Stark, C.; Tyers, M. Osprey: a network visualization system. *Genome Biol.* **2003**, *4* (3), R22.