doi:10.1016/j.jmb.2006.01.005 J. Mol. Biol. (2006) 357, 1322–1334

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Prediction of Binding Sites of Peptide Recognition Domains: An Application on Grb2 and SAP SH2 Domains

William A. McLaughlin, Tingjun Hou and Wei Wang*

Department of Chemistry and Biochemistry, Center for Theoretical Biological Physics University of California, San Diego, 9500 Gilman Drive La Jolla, CA 92093-0359, USA

*Corresponding author

Determination of the binding motif and identification of interaction partners of the modular domains such as SH2 domains can enhance our understanding of the regulatory mechanism of protein–protein interactions. We propose here a new computational method to achieve this goal by integrating the orthogonal information obtained from binding free energy estimation and peptide sequence analysis. We performed a proof-of-concept study on the SH2 domains of SAP and Grb2 proteins. The method involves the following steps: (1) estimating the binding free energy of a set of randomly selected peptides along with a sample of known binders; (2) clustering all these peptides using sequence and energy characteristics; (3) extracting a sequence motif, which is represented by a hidden Markov model (HMM), from the cluster of peptides containing the sample of known binders; and (4) scanning the human proteome to identify binding sites of the domain. The binding motifs of the SAP and Grb2 SH2 domains derived by the method agree well with those determined through experimental studies. Using the derived binding motifs, we have predicted new possible interaction partners for the Grb2 and SAP SH2 domains as well as possible interaction sites for interaction partners already known. We also suggested novel roles for the proteins by reviewing their top interaction candidates.

Keywords: binding site motif; MM/PBSA; protein–protein interactions; SH2 domain

Introduction

The Src homology 2 (SH2) domain functions as a protein-binding module that is used in the control of cellular signal transduction.[1–3] It can serve as an adapter molecule that coordinates the assembly of intracellular signaling proteins in response to an extracellular signal.[2,4–6] Signals mediated by SH2 domains ultimately lead to alterations of the cellular processes such as growth, differentiation, and metabolism.[7] Malfunctions in SH2 domains can lead to a host of human diseases.[8]

Two examples of the SH2 domains are found in the Grb2 and the SLAM-associated proteins (SAP). The Grb2 protein is composed of an SH2 domain and two flanking SH3 domains.[9] A primary function of the Grb2 protein is to bind to protein receptors at the cell surface *via* its SH2 domain and to bind to the SOS protein through its SH3 domains, thereby coupling SOS protein to the membrane where it can activate the Ras protein to initiate a kinase signaling cascade that ultimately leads to modifications in transcription (reviewed by Schlessinger).[10] The SAP protein consists solely of the SH2 domain and is a regulator of signaling events induced by members of the SLAM-related protein receptors found on the surface of T and NK cells.[11–14]

A common feature of all SH2 domains is that binding to an interaction partner is regulated, in part, by the phosphorylation state of a tyrosine residue within that partner.[7] While phosphorylation is required for binding for most SH2 domains,[15] there are SH2 domains that bind to their partners in the absence of phosphorylation, albeit with lower affinity.[3,11,16] Using the method of

Abbreviations used: SH2, Src homology 2; SAP, SLAM-associated protein; HMM, hidden Markov model.

E-mail address of the corresponding author: wei-wang@ucsd.edu

peptide library screening (finding the sequence binding preference based on affinity measurements of oriented degenerate set of random peptide sequences), it has been demonstrated that the sequence determinants of binding or binding specificity of an SH2 domain depend partly on the sequences flanking the tyrosine phosphorylation site.[17,18] For the majority of SH2 domains characterized to date, that specificity is dictated by residues C-terminal to the phosphorylation site, but in some cases, such as the SAP SH2 and the EAT2 SH2 domain, amino acid residues at positions C and N-terminal to the phosphorylation site have been demonstrated to play a role in binding.[19,20]

Peptide library screening has been applied to a number of SH2 domains and there has been accumulation of binding sequence motifs corresponding to each of the domains that have been studied. Many of these binding motifs have been compiled into a web resource called SCANSITE.[21,22] A searching script at that website can be used to predict interaction partners of particular protein domains and thus provides a starting point for identifying candidate interaction partners. Since SCANSITE searching can be conducted only for those domains that have been experimentally characterized, and the strong binding peptides present in the random library but not in the human genome may bias the binding motif determined by the peptide library experiment, alternative techniques are currently needed.

With the aim of addressing these limitations, we have developed a computational method for identifying binding candidates of the modular domains and applied it to the SH2 domains in the proteins of Grb2 and SAP. The method does not rely on peptide library experiments and combines information obtained from binding affinity estimation and sequence motif preference, which is different from the previous approaches using only either type of the information.[22,23] We first created three-dimensional models of known binding peptides as well as randomly selected peptides from the human proteome in complex with the SH2 domain. We next estimated their binding free energies using the molecular mechanics/Poisson–Boltzmann solvent-accessible surface area (MM/PBSA) method.[24,25] These peptides were then clustered based on the binding energy and sequence characteristics and a binding motif was extracted from the cluster of peptides containing those known to interact. The resulting motifs were represented by hidden Markov models (HMMs)[26] and utilized to scan a representative set of human protein sequences in the SWISS-PROT database for likely interaction partners,[27] among which experimental documentation for an interaction with the associated SH2 domain was identified. Possible sites of interaction with the associated SH2 domain were also identified for each interaction candidate. Moreover, based on a literature review of the candidate proteins, new biological roles for the SAP and Grb2 SH2 domains were inferred.

Results

Energy measurements separate known binders from random peptides

The known binding peptides should, on average, have more negative or more favorable binding free energies than peptides selected at random from the background. Two energy measurement protocols were examined with respect to how well the known binding peptides could be separated from the background set of peptides using MM/PBSA. One protocol had the peptides in a phosphorylated state while the other had the peptides in an unphosphorylated state. Student's *t*-test was used to evaluate the significance of the difference between the means.

For peptides in the phosphorylated state, the *p*-value associated with the difference in the mean binding energies of the 15 known peptide binders *versus* the 1400 randomly selected peptides in the Grb2 SH2 domain dataset was 6.41×10^{-5}. For the peptides in the unphosphorylated state, the *p*-value associated with the separation of the two means was lower at 2.31×10^{-9}, which indicated a better separation. Similarly, the *p*-value associated with the separation of the mean binding energy for 11 known binders and the 1799 other peptides in the SAP SH2 domain dataset was lower for the unphosphorylated peptides (7.37×10^{-6}) than for the phosphorylated peptides (3.48×10^{-5}). Figure 1 illustrates the separation between the known binders and peptide candidates in an unphosphorylated state by a histogram plot.

To predict binding motifs and interacting partners of SH2 domains, we chose to use unphosphorylated peptides for the following reasons. First, in our method, rather than to calculate the binding free energy for each binding or non-binding peptide accurately, we only need to establish two distinctive distributions for binders and non-binders. We assume that excluding phosphate does not distort these two distributions, which seems reasonable based on the comparison between the distributions of energy calculations using phosphorylated and unphosphorylated peptides. Second, the binding energy contribution by the phosphate moiety was similar for the known binding peptides and the background set of peptides, and there was a relatively high error associated with its calculation. The binding energy contributed by the phosphate moiety was estimated by subtracting the binding energy of the phosphorylated peptides from the binding energy of the unphosphorylated peptides. For the known binding peptide the average energy contribution due to phosphate binding was 64.78 ($\pm$13.01) kcal/mol for Grb2 and 75.28($\pm$8.83) kcal/mol for SAP. For the background set of peptides, the average energy contribution due to phosphate was estimated to be 71.41($\pm$17.80) kcal/mol for Grb2 and 75.77($\pm$24.34) kcal/mol for SAP. Therefore, the average contribution of phosphate to binding and

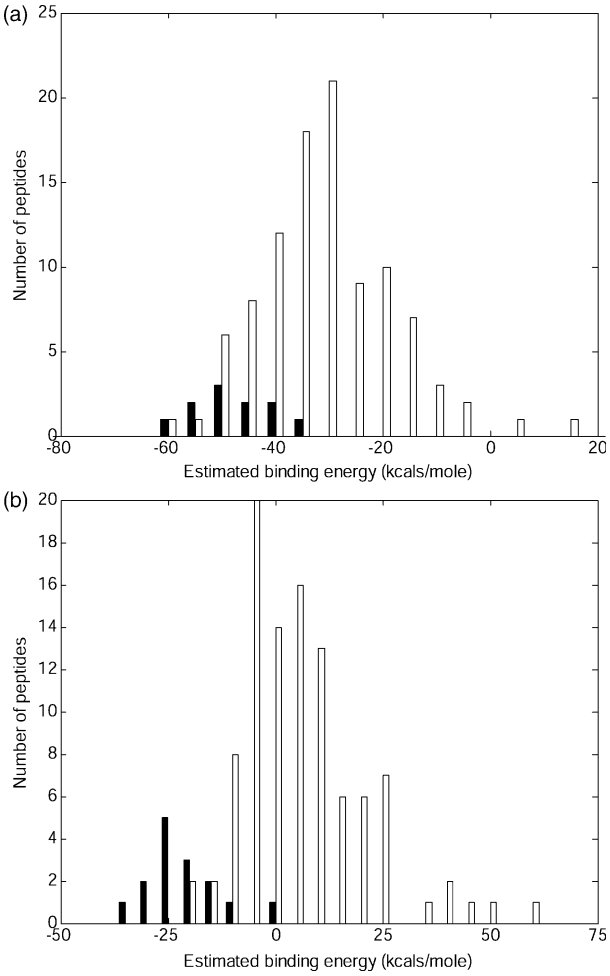

Figure 1. Histogram plot of the calculated binding free energy of known binding peptides and candidate peptides. (a) The energies for the 11 known binders (black) and 100 peptides picked at random from binding candidates in the SAP SH2 domain dataset (white). (b) A similar plot for the Grb2 SH2 domain, with 15 known binders and 100 peptides picked at random from the rest of the dataset.

non-binding peptides has about 10% difference for Grb2 and almost the same for SAP with a larger variance for phosphorylated peptides. Considering that the average difference between the known binding and the background set for the unphosphorylated peptides was 29.92 kcal/mol for Grb2 and 20.23 kcal/mol for SAP, we concluded that the contribution to binding by the phosphate moiety partly obscured the difference between the known binding peptides and the background set that was contributed by the flanking sequences.

By removing the phosphate contribution to binding and therefore reducing the associated noise, focus was placed on the discriminative contribution to binding by the amino acid residues flanking the conserved tyrosine site. A similar strategy was employed by Schueler-Furman *et al*. when computing the free energy of binding between a peptide and major histocompatibilty complex (MHC): contacts that were present in both the binding and non-binding peptides, i.e. the protein–peptide backbone contacts, were removed to improve the performance of discrimination between the binding and non-binding peptides.[28]

Binding motifs derived from the binding clusters were consistent with the experimental results

Peptides were clustered in an unsupervised manner using three schemes as described in Methods: (1) using binding energy only; (2) using sequence only; and (3) using sequence and binding energy together (Table 1).

For the SAP domain dataset, there were two clusters generated using sequence only, five clusters generated using energy only, and six clusters

Table 1. Unsupervised clustering of peptides in the Grb2 and SAP SH2 domain datasets

Cluster number	Candidates	Known binders	Average energy (kcal/mol)
A. Sequence and energy			
SAP SH2 domain			
1	442	1	−35.48±4.30
2	166	0	−5.25±11.64
3	401	0	−21.44±5.17
4	433	0	−30.39±4.01
5	300	10	−44.55±5.33
6	57	0	27.50±368.71
Grb2 SH2 domain			
1	357	1	4.13±3.61
2	13	0	302.15±184.96
3	262	0	13.82±6.03
4	118	14	−16.86±6.31
5	425	0	−4.24±3.44
6	225	0	30.99±8.83
B. Sequence only			
SAP SH2 domain			
1	134	0	N/A
2	1665	11	N/A
Grb2 SH2 domain			
1	1400	15	N/A
C. Energy only			
SAP SH2 domain			
1	123	0	−5.9754±12.9621
2	5	0	802.8107±719.68
3	771	0	−26.1573±7.3608
4	899	11	−36.9477±7.81
5	1	0	−1215.96±66.0503
Grb2 SH2 domain			
1	509	14	1.6075±13.2933
2	13	0	285.1342±192.0195
3	218	0	29.3596±10.1252
4	660	1	0.2332±6.4038

Clustering was done by combining sequence and energy (A), using sequence only (B), and energy only (C). Clustering done using sequence and energy together produced the highest overlap of the known binders in a given cluster for both datasets. For the SAP domain dataset, there were six clusters generated using sequence and energy. The fifth cluster contained the majority of the known binding peptides and was assigned as the binding cluster. The mean and standard deviation of the binding cluster was −44.55 kcal/mol and 5.33 kcal/mol, respectively. For the Grb2 dataset, there were six clusters generated using both sequence and energy. Cluster four was assigned as the binding cluster and had a mean and standard deviation of the energy estimate of −16.86 kcal/mol and 6.31 kcal/mol, respectively.

generated using both sequence and energy. Similarly, for the Grb2 domain dataset there was one cluster generated using sequence alone, four generated using energy alone, and six generated using sequence and energy together. The clustering schemes were evaluated in terms of the overlap score with the known binders, calculated by multiplying the fraction of the known binders relative to the total number of peptides in the cluster. Based on the overlap score, the best clustering scheme for both datasets was achieved by using both sequence and energy.

Using the sequence-energy clustering scheme for the SAP dataset (Table 1), cluster five had the highest overlap score and was assigned as the binding cluster while all other clusters were collectively labeled as non-binding. The binding cluster had an average and standard deviation of the energy estimate of −44.55 kcal/mol and 5.33 kcal/mol, respectively. For the Grb2 dataset, using the sequence-energy clustering scheme the fourth cluster was assigned as the binding cluster. The mean and standard deviation of the energy estimate for the binding cluster was −16.86 kcal/mol and 6.31 kcal/mol, respectively. Since the conformational entropy was not included, these values are not absolute binding free energies. We assumed however that the conformational entropy for each peptide binding to the same protein was similar and thus the MM/PBSA method accurately predicted the relative binding affinities of the peptides. From the clustering results, it was apparent that the binding clusters had the most favorable binding free energies relative to the other clusters.

The peptides in each binding cluster were used to create an HMM (binding cluster HMM). In addition, control HMMs were created as described in Methods. Common sequence characteristics, i.e. a sequence motif, represented by these HMMs are shown in Figure 2. The experimentally derived motifs as described in the literature are also shown for comparison.

Figure 2(a) shows representations of the binding motifs of the SAP SH2 domain. The experimentally derived motif for SAP has been described by Poy *et al*.[20] as TIpYXX(V/I), where T,I,pY,X, and V represent threonine, isoleucine, phosphorylated tyrosine, any amino acid, and valine, respectively. In another study, Hwang *et al*. have described the SAP SH2 binding motif as (T/S)XXXX(V/I).[29] The motif derived by the alignment of 11 known interaction sites of the SAP domain is shown at the center of Figure 2(a). The motif showed the information regarding the conserved features of the binding motif described by the experimental studies but features specific to the 11 known binding peptides are over-represented. Alternatively, if the peptide sequences in the binding cluster were used, the motif conveyed more of the features of the experimentally derived binding motif: (1) no conservation at positions Y−4, Y−3, Y+1, and Y+4; (2) threonine at position Y−2; (3) isoleucine at position Y−1; and (4) valine at position Y+3. The conserved leucine at position Y+2 was also consistent with the peptide library studies by Poy *et al*. and Hwang *et al*., although this position was not chosen to be represented in the reported motifs.[20,29] In addition, the conserved alanine at position Y+5 was found to be consistent with the Hwang *et al*. study but that position was not examined by Poy *et al*. Overall, the motif

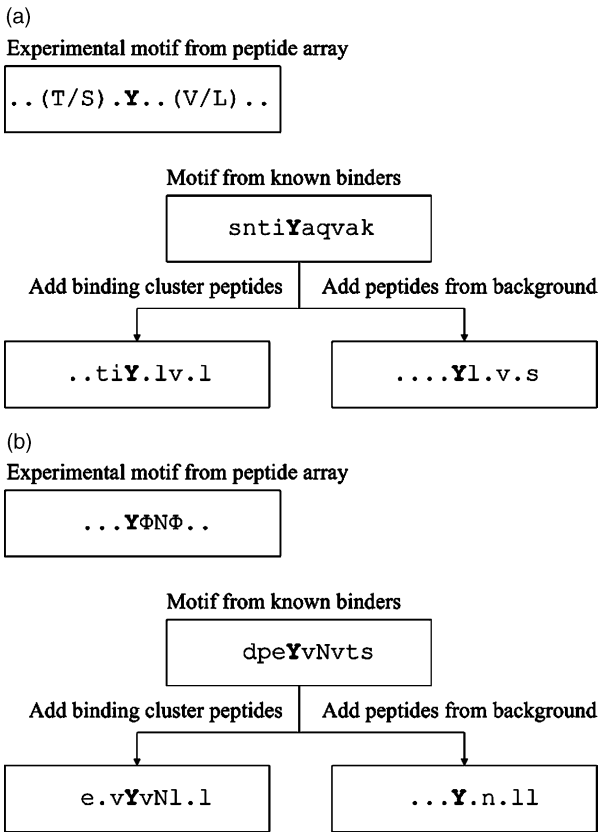

Figure 2. Reconstruction of the experimentally derived binding motifs for the Grb2 and SAP SH2 domains. The SAP SH2 domain binding motifs are shown in (a) while (b) shows the Grb2 binding motifs. The motifs on the top of each panel were taken from the literature. The rest of the motifs were constructed by finding the amino acid preference using the majority rule at each position of the corresponding HMM. The motifs in the middle of each panel correspond to HMMs generated using only the sequences of the known binding sequences. The motifs at the bottom left are from HMMs created using sequences of the binding cluster, which can be viewed as a set of sequences created by adding peptide sequences of the binding cluster to the known binding sequences. The motifs at the bottom right are from HMMs generated using sequences of the control cluster, which can be viewed as a set of sequences created by adding peptides from the non-binding cluster to the known binding sequences. Note the similarity between the experimentally derived motifs and the motifs derived from the binding cluster sequences. One letter codes follow the IUPAC convention except for the Greek letter Φ, which represents a hydrophobic residue.[59] Insertion points are represented as dots.

derived from the binding cluster represented a reconstruction of the experimentally derived motif. For the control HMM motif, one position (a valine residue conserved at position Y+3) was found to be the same for both the control and the experimental motif, but overall the motifs were not considered to be similar.

The binding motifs of the Grb2 SH2 domain are displayed in Figure 2(b). The experimentally derived motif for the domain was described by Songyang *et al.*[18] as pYΦNΦ, where Φ is a hydrophobic residue and N is an asparagine residue. In the motif derived from the 15 known binding peptides, i.e. the known only motif, the asparagine at position Y+2 was apparently highly conserved as it was given in upper case, which is a property indicating that it was present in greater than 50% of the cases. (The asparagine at position Y+2 was actually present in all of the known examples.) In examining the known only motif, the correspondence with the experimentally derived motif was difficult to discern, since the motif contained too many features idiomatic to the 15 known binding sequences. In contrast, the motif derived from the binding cluster emphasized the major features found by the experimental study. These features are a hydrophobic residue at position Y+1, an asparagine at Y+2, and a hydrophobic residue at Y+3. For the control, the majority of these characteristics were not exhibited, indicating that the experimental motif was not present for the sequences of the control cluster. We note that the partial enrichment of the asparagine residue at position Y+2 in the control motif reflected the fact that there is 100% conservation of that residue in the known binding sequences, which were part of the control cluster.

Results of database searches

We examined the ranks of the peptides known to interact with each domain upon a database search of 174,604 tyrosine-containing human peptide sequences in the SWISS-PROT database using the three different HMMs: (1) the binding cluster HMM created with sequences in the binding cluster; (2) the control cluster HMM created with the known binders and the background peptides randomly selected from the human proteome; and (3) the known only HMM created with only the sequences of the known binding peptides. Shown in Figure 3 are the log percentile ranks of the known binders after being retrieved by each one of these HMMs. Based on the distributions of the known binders, the performance in placing the known binders within the top scoring peptides was best for the known only HMM, followed by the binding cluster HMM, and then the control cluster HMM. Student's *t*-test was used to quantify the difference of the mean log percentile rank of the known binders when using the binding cluster HMM *versus* the control cluster HMM. The *p*-value of that *t*-test was 2.78×10^{-6} and 1.62×10^{-4} for the SAP and Grb2 binding site searches, respectively. These tests indicated that the binding clusters contained information with regard to binding that extended beyond that contained in the known binding sequences.

For the searches using the known only HMMs, it was not surprising that the known binding peptides were retrieved at the very top of the ranked lists. All of the 11 known binders for the SAP SH2 binding site search were being retrieved within the top 23 peptides of the ranked peptides and all of the 15 known binders of the Grb2 SH2 domain were within the top 54. The known only HMMs were therefore highly specific for the known binding sequences. These HMMs may be used to identify binding sites that are highly similar to the set of known sequences but could miss those with divergent sequences. We therefore focused on the binding cluster HMMs, as they exhibited more general motifs that were consistent with those derived experimentally.

For the binding cluster HMMs, the search results were analyzed in the following ways: (1) comparing the results to those of a similar search made using SCANSITE; (2) noting the possible interaction sites of known interaction partners documented in the BIND and MINT databases;[30,31] and (3) manually reviewing the top scoring candidates. The SCANSITE comparison could only be done for the Grb2, as the SAP SH2 domain binding motif was not yet available in SCANSITE, although the requisite peptide library experiments had been done.[20,29]

For the Grb2 SH2 domain binding site search, the top 2000 (the maximum number of) sites retrieved by SCANSITE search of human sequences in SWISS-PROT were compared to the top 2000 sites retrieved by the binding cluster HMM. The result was that there were a total of 650 overlapping sites, indicating that there was good agreement between the computationally derived motif and experimentally derived motif on which the SCANSITE search was based. To further check the performance of SCANSITE *versus* the binding cluster HMM, the ranks of known binders were examined. For both searches, 12 of the 15 known binding sites were found within the top 2000 ranked peptides.

The comparison results of the Grb2 SH2 binding site search showed that the binding cluster motif and the SCANSITE motif were similar. That is, although both motifs yielded a large number of false positives upon screening a large set of peptide sequences, the fact that the two motifs found overlapping groups of peptides indicated that they contained similar information. Further, since the two motifs were similar, we inferred that the computational screening method that was needed to generate the binding cluster HMM provided a viable alternative to the experimental peptide library screening that was required to derive the SCANSITE motif.

The second way to analyze the searches with the binding cluster HMMs was to note if the proteins containing any of the top 2000 ranked peptides

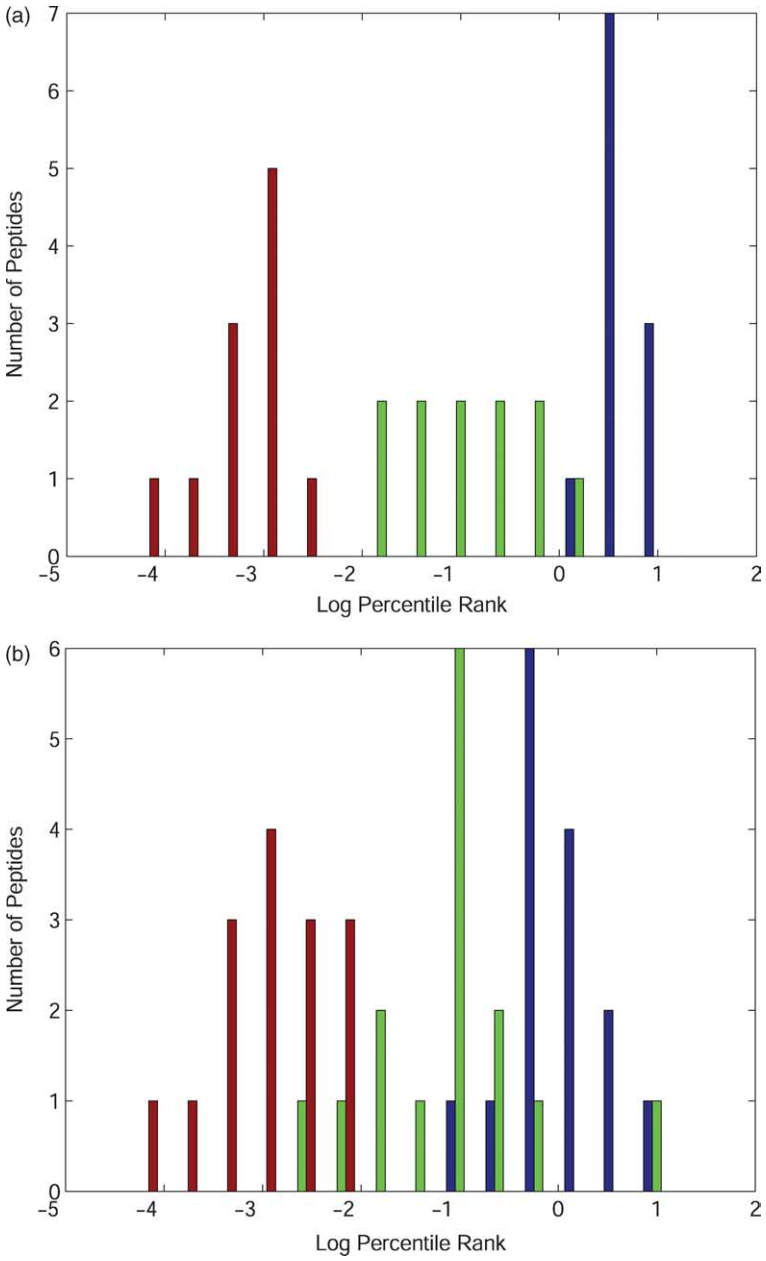

Figure 3. Plots of the log percentile ranks of the known binding peptides for database searches using three different HMMs: the known only HMM (red), the binding cluster HMM (green), and the control cluster HMM (blue). (a) Results of the SAP SH2 domain binding site search; (b) results for the Grb2 SH2 domain binding site search. Student's *t*-test was used to compare the mean log percentile ranks of known binders as retrieved using the binding cluster HMM and the control cluster HMM. The *p*-value for that *t*-test was 2.78×10^{-6} for the SAP SH2 domain binding site searches and was 1.62×10^{-4} for the Grb2 SH2 domain binding site searches.

were known to interact with the associated SH2 based on annotation in the protein interaction databases such as BIND and MINT. There were several of these overlapping proteins and they are listed in Table 2 for both the Grb2 and SAP SH2 domain binding site searches. For each interaction candidate listed, possible sites of interaction are suggested by our search.

The final way of analyzing the binding cluster HMM searches was to determine if the top ranking peptides were likely binding partners based on manual review. To increase the probability of finding true interaction partners within these top candidates, a conservation filter was first applied to remove less likely binding candidates. The steps involved in applying the conservation filter are described in Figure 4 and in Methods. The conservation filter removed approximately 20% of the top scoring peptides for the Grb2 search and approximately 5% of the top scoring peptides for the SAP search. The following describes some of the top 50 scoring conserved peptide candidates that had experimental evidence for the interaction based on the manual review.

For the Grb2 SH2 domain binding site search, the protein tyrosine-protein kinase receptor UFO (UniProt accession code P30530) contained a peptide that had an overall search score rank of 49. That site corresponds exactly to the Grb2 SH2 domain binding site reported by Braunger *et al.*[32] The binding site was apparently missed during the survey conducted by curators of the Phospho.ELM database[33] for documented peptide examples of Grb2 SH2 domain binding sites. Since the site was

Table 2. A list of possible interaction sites for the Grb2 and SAP SH2 domains

Protein name	Accession code	Search rank	Tyrosine positions
A.			
Beta-adaptin	P63010	74	276
GTPase-activating protein GAP	P20936	125, 632	472, 615
Spectrin alpha chain	Q13813	126	2430
Linker for activation of T cells	O43561	187	1156
60 S ribosomal protein L3	P39023	329	1118
Bullous pemphigoid antigen 1, isoforms 6/9/10	O94833	425, 1742	633, 739
Bullous pemphigoid antigen 1, isoforms 1/2/3/4/5/8	Q03001	426, 1708, 1743	1160, 460,1271
Receptor-type tyrosine-protein phosphatase alpha	P18433	446	588
Linker for activation of T cells	O43561	730, 740	220, 200
Polypyrimidine tract-binding protein 1	P26599	878	430
Myosin Ic	O00159	882	405
Actin-like protein 3	P61158	886	316
Receptor-type tyrosine-protein phosphatase alpha	P18433	893, 1076	295, 362
Putative RNA-binding protein Luc7-like 2	Q9Y383	912	173
Filamin B	O75369	967	181
Serine/threonine-protein kinase PAK 1	Q13153	1014	464
40 S ribosomal protein S11	P62280	1087	37
Epidermal growth factor receptor	P00533	1099	1110
USP6 N-terminal like protein	Q92738	1219	551
Alpha-actinin 1	P12814	1303	161
SHC transforming protein 3	Q92529	1342	341
Myosin Ia	Q9UBC5	1452	399
Glutathione *S*-transferase P	P09211	1462	198
Myosin IIIB	Q8WXR4	1464	1275
Alpha-actinin 4	O43707	1604	180
Protein phosphatase 1 regulatory subunit 12A	O14974	1768	68
Inositol 1,4,5-trisphosphate receptor type 3	Q14573	1813	1588
Ras GTPase-activating protein 1	P20936	1899	619
Tumor necrosis factor ligand superfamily member 6	P48023	18	258
Rap guanine nucleotide exchange factor 1	Q13905	994	485
B.			
TRK1 transforming tyrosine kinase protein	P04629	680	756
TrkB tyrosine kinase	Q16620	706	757
Proto-oncogene tyrosine-protein kinase LCK	P06239	469	841

The list was constructed by identifying proteins containing one of the top 2000 scoring peptides retrieved by a search with a binding cluster HMM that were documented to bind either the Grb2 protein or the SAP protein based on MINT or BIND annotation. The name of the protein, the UniProt accession code, the rank of the peptide according to the search, and the tyrosine position corresponding to the proposed binding site are listed. Group A shows possible binding sites of the Grb2 SH2 domain and group B shows possible binding sites for the SAP SH2 domain. The lists exclude those sites that were in the original set of known sites.

retrieved at rank 49 of 174,604 possible sites and was not part of the input known set, it suggests that the binding cluster HMM can identify true binding sites for the Grb2 SH2 domain. Also, the same site had a rank of 227 for the SCANSITE search, indicating that the two search techniques are of complementary utility, notwithstanding that the techniques were of similar utility as shown above. Finally, the site had a rank of 1600 using the known only HMM, indicating that the binding cluster HMM has the utility of finding true binding sites with a sequence somewhat dissimilar to the known binding sequences.

The UFO protein had been originally obtained from a person with myeloproliferative disorder and was named unknown functioning oncogene or UFO.[34] It is in the Axl protein family, which are characterized as having extracellular immunoglobulin domains and a fibronectin domain proposed to be involved in cell adhesion.[35] Ligands of the UFO protein receptor link it to hemostasis as Gas6, a sequence relative of protein S, has been demonstrated to be the interaction partner.[35] We infer that coagulation activities may be integrated to cellular responses, such as growth and repair, upon ligation of UFO with Gas6 through a signal transduction pathway that involves the Grb2 SH2 domain. That inference is based on the following facts: (1) the Grb2 SH2 domain binding site on UFO found here has been experimentally validated;[32] (2) ligation of a protein receptor related to UFO (tyro 3) with protein S protein causes specific tyrosine phosphorylation;[36] (3) the tyro 3/protein S system is analogous to the UFO–Gas6 system since the tyro 3 is related in sequence to UFO and protein S is related in sequence to Gas6;[36] and (4) the tyro 3/protein S system integrates coagulation activities and intracellular signaling.[36] The inference could have made based solely on literature but the binding site search reaffirmed the interaction between Grb2 and UFO and prompted an examination of the biological context of that interaction.

The second example is protein macrophage colony stimulating factor 1 receptor (P07333), which was ranked at number 31 in the Grb2 binding site search after application of the conservation

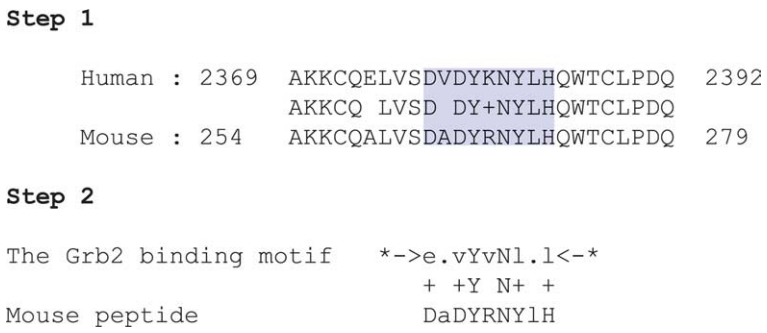

Figure 4. Diagram illustrating the method for testing whether a top scoring human peptide found during the database search had a conserved counterpart in mouse and should therefore be considered to be a more likely binding candidate. Step 1 was to perform a pairwise BLAST alignment of the protein from which the high scoring human peptide was derived against its mouse homolog in order to identify the portion of the mouse protein that aligned with that human peptide. That aligned portion between the high scoring human peptide with its mouse counterpart is shown highlighted in blue. Step 2 was to compare the aligned mouse peptide to the binding cluster HMM. If that comparison gave a high score, the human peptide was more likely to be a binding candidate and was kept for manual review.

filter (SCANSITE rank 1466). The protein is involved in regulating growth, survival, and differentiation in hematopoietic cells of the monocyte–macrophage lineage.[37] The tyrosine at position 697 was shown to be phosphorylated through autocatalysis and to interact with the Grb2 SH2 domain.[38] Based on the search results, the tyrosine at position 923 is also a binding site for the Grb2 SH2 domain. That site has been shown to be a tyrosine phosphorylation site of unknown function.[37] Neither the 923 site nor the 697 site were included in the known binding sequences that were documented in the Phospho.ELM database as binding to the Grb2 SH2 domain.

At rank number 40 (SCANSITE rank 45) was a peptide from the desmoglein-2 protein (Q14126). The desmoglein-2 protein is a transmembrane glycoprotein involved in calcium-dependent cell–cell adhesion.[39] Upon inhibition of the epidermal growth factor receptor (a kinase that creates a binding site for the Grb2 SH2 domain by autophosphorylation),[40] there is a decrease in desmoglein-2 tyrosine phosphorylation.[41] Based on the search results, we propose that the Grb2 SH2 domain binds to the phosphorylated desmoglein-2 protein at tyrosine position 511. That interaction would infer a role for the Grb2 protein in the regulation of cell adhesion *via* binding to the desmoglein-2 protein.

Searches using the binding cluster HMM created to identify the SAP SH2 binding sites also produced viable binding site candidates. One definitive example was for the protein NTB-A (Q96DU3), which contained a peptide at tyrosine position 309 that ranked number 4 in the list of 176,604 possible tyrosine binding sites being searched. The NTB-A has been shown to bind to the SAP SH2 domain,[42] but the specific binding site was not confirmed. Bottino *et al*. described the NTB-A protein as a novel binder of the SAP SH2 and showed that it had two tyrosine-based motifs with the consensus sequence TXYXX(V/I), one of which corresponded to tyrosine position 309 found during the search. We provide further support that position 309 is a binding site of the SAP SH2 domain. Note that the NTB-A protein has 22% sequence identity with the SLAM protein so it was retrieved when creating the initial dataset of candidate binding sites for the SAP SH2 domain. The procedure of including additional candidate binding sites by finding sequences related to a protein known to interact with the binding domain therefore appears to be validated for this case. (See Methods for details.)

An interesting candidate for binding to the SAP SH2 domain is the Vav proto-oncogene protein (P15498), whose position 826 was ranked 16 in the search. Evidence that the protein may bind to the SAP SH2 domain includes the following: the protein has been shown to be tyrosine phosphorylated, it is involved in signal transduction, and it plays a role in B-cell proliferation,[43] which is a function of the SAP SH2 domain.[11] Also, in NK cells of persons with the XLP syndrome, a disease caused by a deleterious mutation in SAP, Vav protein phosphorylation is deficient following the 2B4 receptor stimulation.[44] Moreover, it has recently been demonstrated that phosphorylation of the Vav protein is SAP-dependent and the mechanism is still under investigation.[44] We speculate that the mechanism involves an association of the SAP SH2 domain at position 826 of the Vav protein.

We anticipate that further manual review of more top scoring candidates would yield more viable candidates, which could suggest more novel functions of Grb2 and SAP proteins. The candidates outlined here show the utility of reviewing the search results motifs. For future analysis, the hidden Markov models can be retrieved†.

Discussion

We proposed here a computational method for identifying binding partners and binding sites of modular domains and have demonstrated its feasibility on SAP and Grb2 SH2 domains. Ideally, if the binding free energies between the domain of interest and all potential peptides in the genome were accurately calculated, binding sequences

† http://modem.ucsd.edu/billm/SH2_Supp/SH2_SupplementaryMaterial.htm

could have been easily identified. In reality, rigorous free energy calculations were too time-consuming to be applied at a genomic scale while rough free energy estimation methods would not have been accurate enough to reliably distinguish binding from non-binding peptides. It was reasonable to assume that binding peptides tend to contain the binding motif and have more favorable binding affinities. Our method thus combined information obtained from these two orthogonal approaches, binding free energy estimation and sequence analysis, to achieve a better performance than using either approach alone (see above). As we only needed distributions of binding affinities but not accurate ranking of peptides, we could estimate the binding free energies of thousands of peptides using efficient approaches.

Our method used a clustering scheme to combine sequence information and binding energy estimates and thus allowed those peptides containing sequence and binding energy patterns similar to the known binding sequences to be placed together in the same group, i.e. the so-called binding cluster. As a consequence, the sequence motif of the binding cluster, which was a synthesis of the predominant sequence features of the group, implicitly contained information about high affinity peptides that had an energy distribution similar to the known binders. To illustrate that point, shown in Figure 5 are the plots given in Figure 1 overlaid with the top 100 peptides retrieved from the human sequences in SWISS-PROT by each of the binding cluster HMMs.

The training set of peptide sequences used to create the binding cluster HMM was constructed by screening a set of randomly selected human peptide sequences for affinity similar to a group of known binding peptides: essentially new raw binding data were found from an *in silico* binding experiment. The test set was the set of all possible interaction candidates in a sequence database. The study is analogous to the peptide library experiment where a binding motif is derived from peptides with relatively high binding affinity and the test set is the set of all possible interaction candidates. Our method thus provides an alternative to suggesting candidates to experimentalists to narrow down the searching possibilities for novel binding sites of a modular domain.

The utility of the current method was that it extracted information regarding the contribution of binding by amino acid sequence flanking the conserved tyrosine position. That information was used to create an expanded motif that could identify likely interaction candidates for the Grb2 and SAP SH2 domains. The candidates were found to be useful in the following respects: they implicated the Grb2 and SAP SH2 domains in novel signaling pathways, aided in consolidating the known set of interaction partners, and provided possible sites of interaction for known interaction partners of the domains. The candidates provide starting points for further experimental studies.

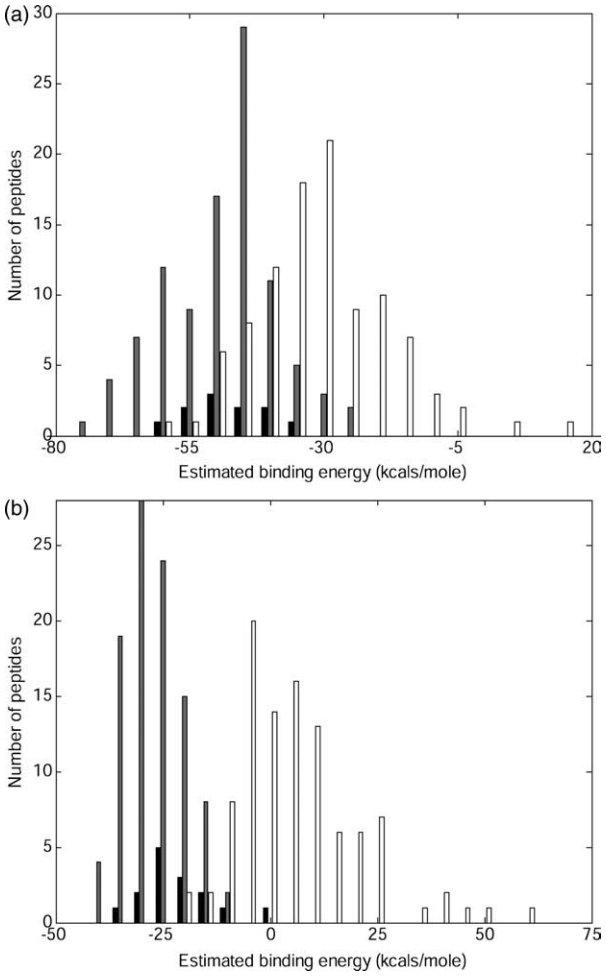

Figure 5. Histograms that were plotted originally in Figure 1 are shown again with the top 100 peptides that were found during the database search overlaid. The known binders are in black; 100 of the original candidate peptides are shown in white; and 100 of the top scoring peptides retrieved by a database search using the binding cluster HMM are shown in grey. The top scoring had a relatively high affinity that was close to that of the known binding peptides. (a) Plots for the SAP dataset; (b) the Grb2 dataset.

Since the method requires only a protein–peptide complex structure, it can be applied to any protein–peptide complex even when the results of a peptide library experiment are not available. Also, since it considers only peptides existing in the relevant proteome, the bias introduced by having strong non-physiological binding peptides present in a random peptide library can be avoided. Given the fast pace of advancement of structural genomics and homology modeling, we anticipate that the method will become more and more useful.

The method presented here is far from perfect and it can be improved in many aspects. For example, more accurate binding affinity calculation and more efficient sampling method are desired; additional information of protein interactions *in vivo* including the cellular compartmentalization

of proteins, the ability of the kinase to bind to the substrate (a step required prior to the high-affinity SH2 binding), and location of the peptide within the protein (either on the surface or buried) no doubt would help.

Methods

Datasets

Protein–peptide complexes of two SH2 domains, Grb2 and SAP, were found in the Protein Data Bank.[45] The complexes were solved by X-ray crystallography and had the best available resolution. The structure of SAP SH2 domain was that solved in complex with a peptide derived from the SLAM protein by Poy *et al*.[20] The structure of Grb2 SH2 domain was that solved in complex with Shc-derived peptide by Nioche *et al*.[46] For each domain, a set of possible binding peptides and a set of known binding peptides were generated, which had the same length and tyrosine position as the peptide in the solved peptide/domain complex. In the case of the Grb2 SH2 domain, the peptides were nine amino acid residues long with the sequence (XXXYXXXX), where X represents any amino acid and Y indicates tyrosine. A group of 1400 such peptides were selected at random from human protein sequences available in the SWISS-PROT database.[27] A total of 15 peptide sequences known to interact with the Grb2 domain within human proteins were found in the Phospho.ELM database.[33] For the SAP domain, the same 1400 candidate peptides were used again but had an extension of one residue to match the peptide length found in the solved complex. Additional peptides for the SAP dataset were selected from proteins related in sequence to the SLAM protein, a known interaction partner of the SAP SH2 domain. The peptides were found by performing a BLAST search with SLAM protein sequence against all human protein sequences in the UniProt database with an expectation value cutoff of 10.[47,48] Each tyrosine-containing peptide in the retrieved protein sequences related to the SLAM protein was then extracted and tested for conservation in the following manner: a BLAST search of the protein containing the peptide was made against the mouse protein sequences in UniProt; the highest scoring mouse sequence was taken to be the mouse homolog to the human protein; if there was a tyrosine residue at the aligned tyrosine position in the mouse homolog, the human peptide was kept. In total, there were 399 conserved peptides within SLAM-related protein sequences that were added to the SAP dataset. Eleven known interaction peptides of the SAP domain were documented by Li *et al*.[49] In summary, there were 1415 peptides in the Grb2 dataset (15 known binders and 1400 randomly chosen candidates) and there were 1810 peptides in the SAP dataset (11 known binders, 1400 randomly chosen candidates, and 399 candidates from SLAM related sequences).

Modeling of peptide/domain complexes and binding energy estimation

A three-dimensional model was generated for each SH2 in complex with each peptide sequence in its dataset using the backbone conformation of the peptide in the solved complex as a template and modeling side-chain conformations using the program SCWRL.[50] Each model complex was solvated in a box of TIP3P water molecules and optimized by 2500 steps of energy minimization using the AMBER 8 software package with the *parm*99 force field.[51,52] The minimization entailed 1250 steps using the steepest descent method followed by 1250 steps using the conjugate gradient method.

The binding energy of each peptide to its associated SH2 domain was calculated using the molecular mechanics/Possion–Boltzman solvent-accessible surface area (MM/PBSA) method.[24] The conformational entropy term was not considered because it was previously shown not to correlate with experimentally measured binding energy for a similar system[53] and such calculation was also time-consuming.

To determine an optimal set of energy measurement conditions, a comparison was made between the mean binding energy of the known binding peptides and the mean energy value calculated for all the candidate peptides using Student's *t*-test. Better separation of the means was indicated by a lower *p*-value of the *t*-test. Comparisons were made for the case where all the peptides were in a tyrosine-phosphorylated state and where all the peptides were in the unphosphorylated state.

Clustering of peptides based on binding free energy and sequence characteristics

The parameters used for clustering were the amino acid type at each position of the peptide, which could be any of the 20 natural amino acids, and the estimated binding energy of the peptide. For each domain, peptides were clustered using three schemes: sequence only, energy only, and sequence and binding energy together. Clustering was done in an unsupervised manner, using the *k*-means algorithm followed by optimization with expectation maximization (EM) algorithm as implemented in the Weka machine learning software.[54–56] Energy values were modeled as a normal distribution and peptide sequences were modeled as a position-specific frequency matrix (PSFM). The likelihood that was to be maximized using EM was calculated as:

$$\prod_k \sum_c \left(\prod f^c_{i(k)} \right) (\rho^c(E_k))(\lambda^c)$$

where k is the peptide instance, c is the cluster number, f is the frequency value for the amino acid type at the position i of the peptide for that cluster. The quantity $\rho^c(E_k)$ is the probability that a peptide takes on the energy value of the peptide k in cluster c. The quantity λ^c is the fraction of the peptides belonging to cluster c. To avoid multiplication steps, the log likelihood was calculated instead as the sum of the logs of the individual components.

The number of clusters was selected by a cross-validation-like procedure. For each of the ten divisions of the data, the likelihood measure shown above was calculated assuming that there was one cluster and averaged across the ten divisions. The process was then repeated assuming that there were two clusters. If the average likelihood calculated when assuming two clusters was greater than when assuming one cluster, the number of clusters was set to two. The number of clusters was successively increased in a similar way until the measured likelihood no longer increased.

Binding motif extraction

After clustering was complete for each domain dataset, cluster contents were examined with respect to the

number of known binders and candidate peptides. The cluster containing the majority of the known binding peptides was labeled as the binding cluster while all the other clusters were collectively labeled as the non-binding cluster. A hidden Markov model was generated using sequences in the binding cluster with the HMMER software†. Two additional control HMMs were also generated for each domain dataset. The first was created using the known binding peptides in the binding cluster plus peptides selected randomly from background to make the total number of the peptides in the control cluster the same as that in the binding cluster. (The set of background peptides was 174,604 tyrosine-containing peptides in human proteins available in SWISS-PROT.) The second control group contained only known binding peptides.

To visualize the major characteristics of each HMM, the program hmmemit in the HMMER software was utilized. The program uses a majority rule estimation of the amino acid type found at each position of the model in the case where the match score was greater than the insertion score. For the opposite case, an insertion point is shown represented as a dot at that position.

Database searches

A dataset of 11,935 human protein sequences were retrieved from the SWISS-PROT database on March 1, 2005. From these sequences, the program CD-HIT was used to remove the most highly redundant sequences so they did not obscure the database search results.[48,57] The representative list contained 11,426 sequences and had less than 90% sequence identity. For the Grb2 binding site scans, all peptides of nine amino acid residues length having tyrosine at position 4, i.e. with the sequence XXXYXXXXX, were extracted. Similarly, all peptides of ten amino acid residues length with tyrosine at position 5 were extracted for the SAP binding site scans. In total, there were 174,604 peptides available for both scans. All the peptides were scored with each of the Grb2 and SAP HMMs and the scores were converted to log percentile ranks.

To assess the utility of each binding cluster HMM, database searches were analyzed in different ways. The first way was to compare the mean log percentile rank of the known binders as they were retrieved using the binding cluster HMM with that using the control HMM built from the known binders plus peptides in the non-binding cluster. The second way was to compare the search results obtained by SCANSITE. Third, the overlap of the top 2000 candidates with interacting proteins documented in either the MINT or BIND database was found and possible site of interaction based on the position of the candidate peptide were proposed. Fourth, for the top 50 scoring peptides the literature was reviewed to identify evidence of the interaction with the associated SH2 domain.

Before undertaking the manual review process, putative false positives from the top scoring peptides were removed by applying a filter requiring the binding motif to be conserved in its mouse protein homolog. That was done by taking the human protein from which a top scoring peptide was derived and performing a BLAST search against all mouse protein sequences in UniProt. For the highest scoring mouse protein, the mouse peptide that aligned to the tyrosine site of the human peptide was scored using the binding cluster HMM. Only if the mouse peptide was scored high, i.e. a score that 80% of the known binders were above, the human peptide was kept. The top 50 human peptides found to be conserved in that way were assessed as to their viability of being true binding candidates by review of documentation sources that included Medline references,[58] SWISS-PROT annotation, the MINT database, and the BIND database.

† http://hmmer.wustl.edu/

Acknowledgements

W.A.M is supported by an NIH training grant (5 T32DK07233). T.H. is supported by a postdoctoral scholarship from the Center for Theoretical Biological Physics (CTBP) at UCSD. We thank CTBP for computational support and members of the J. Andrew McCammon group for helpful discussions. We thank Ray Luo for providing support for the PB solver in AMBER and Joseph Nachman for providing parameters of the phosphotyrosine residue.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.01.005

References

1. Pawson, T. (2004). Specificity in signal transduction: from phosphotyrosine–SH2 domain interactions to complex cellular systems. *Cell*, **116**, 191–203.
2. Schlessinger, J. & Lemmon, M. A. (2003). SH2 and PTB domains in tyrosine kinase signaling. *Sci. STKE*, RE12.
3. Yaffe, M. B. (2002). Phosphotyrosine-binding domains in signal transduction. *Nature Rev. Mol. Cell. Biol.* **3**, 177–186.
4. Schlessinger, J. & Ullrich, A. (1992). Growth factor signaling by receptor tyrosine kinases. *Neuron*, **9**, 383–391.
5. Shoelson, S. E. (1997). SH2 and PTB domain interactions in tyrosine kinase signal transduction. *Curr. Opin. Chem. Biol.* **1**, 227–234.
6. Kuriyan, J. & Cowburn, D. (1997). Modular peptide recognition domains in eukaryotic signaling. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 259–288.
7. Pawson, T., Gish, G. D. & Nash, P. (2001). SH2 domains, interaction modules and cellular wiring. *Trends Cell Biol.* **11**, 504–511.
8. Machida, K. & Mayer, B. J. (2005). The SH2 domain: versatile signaling module and pharmaceutical target. *Biochim. Biophys. Acta*, **1747**, 1–25.
9. Lowenstein, E. J., Daly, R. J., Batzer, A. G., Li, W., Margolis, B., Lammers, R. *et al*. (1992). The SH2 and SH3 domain-containing protein GRB2 links receptor tyrosine kinases to ras signaling. *Cell*, **70**, 431–442.
10. Schlessinger, J. (1993). How receptor tyrosine kinases activate Ras. *Trends Biochem. Sci.* **18**, 273–275.
11. Sayos, J., Wu, C., Morra, M., Wang, N., Zhang, X., Allen, D. *et al*. (1998). The X-linked lymphoproliferative-disease gene product SAP regulates signals induced through the co-receptor SLAM. *Nature*, **395**, 462–469.

12. Morra, M., Howie, D., Grande, M. S., Sayos, J., Wang, N., Wu, C. *et al*. (2001). X-linked lymphoproliferative disease: a progressive immunodeficiency. *Annu. Rev. Immunol.* **19**, 657–682.
13. Chan, B., Lanyi, A., Song, H. K., Griesbach, J., Simarro-Grande, M., Poy, F. *et al*. (2003). SAP couples Fyn to SLAM immune receptors. *Nature Cell Biol.* **5**, 155–160.
14. Coffey, A. J., Brooksbank, R. A., Brandau, O., Oohashi, T., Howell, G. R., Bye, J. M. *et al*. (1998). Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene. *Nature Genet.* **20**, 129–135.
15. Mayer, B. J. & Baltimore, D. (1993). Signalling through SH2 and SH3 domains. *Trends Cell Biol.* **3**, 8–13.
16. Nantel, A., Mohammad-Ali, K., Sherk, J., Posner, B. I. & Thomas, D. Y. (1998). Interaction of the Grb10 adapter protein with the Raf1 and MEK1 kinases. *J. Biol. Chem.* **273**, 10475–10484.
17. Songyang, Z., Shoelson, S. E., Chaudhuri, M., Gish, G., Pawson, T., Haser, W. G. *et al*. (1993). SH2 domains recognize specific phosphopeptide sequences. *Cell*, **72**, 767–778.
18. Songyang, Z., Shoelson, S. E., McGlade, J., Olivier, P., Pawson, T., Bustelo, X. R. *et al*. (1994). Specific motifs recognized by the SH2 domains of Csk, 3BP2, fps/fes, GRB-2, HCP, SHC, Syk, and Vav. *Mol. Cell. Biol.* **14**, 2777–2785.
19. Morra, M., Lu, J., Poy, F., Martin, M., Sayos, J., Calpe, S. *et al*. (2001). Structural basis for the interaction of the free SH2 domain EAT-2 with SLAM receptors in hematopoietic cells. *EMBO J.* **20**, 5840–5852.
20. Poy, F., Yaffe, M. B., Sayos, J., Saxena, K., Morra, M., Sumegi, J. *et al*. (1999). Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition. *Mol. Cell*, **4**, 555–561.
21. Obenauer, J. C., Cantley, L. C. & Yaffe, M. B. (2003). Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucl. Acids Res.* **31**, 3635–3641.
22. Yaffe, M. B., Leparc, G. G., Lai, J., Obata, T., Volinia, S. & Cantley, L. C. (2001). A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nature Biotechnol.* **19**, 348–353.
23. Wollacott, A. M. & Desjarlais, J. R. (2001). Virtual interaction profiles of proteins. *J. Mol. Biol.* **313**, 317–342.
24. Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L. *et al*. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accts Chem. Res.* **33**, 889–897.
25. Wang, W., Donini, O., Reyes, C. M. & Kollman, P. A. (2001). Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein–ligand, protein–protein, and protein–nucleic acid noncovalent interactions. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 211–243.
26. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK.
27. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E. *et al*. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.* **31**, 365–370.
28. Schueler-Furman, O., Altuvia, Y., Sette, A. & Margalit, H. (2000). Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci.* **9**, 1838–1846.
29. Hwang, P. M., Li, C., Morra, M., Lillywhite, J., Muhandiram, D. R., Gertler, F. *et al*. (2002). A "three-pronged" binding mechanism for the SAP/SH2D1A SH2 domain: structural basis and relevance to the XLP syndrome. *EMBO J.* **21**, 314–323.
30. Bader, G. D., Betel, D. & Hogue, C. W. (2003). BIND: the Biomolecular Interaction Network Database. *Nucl. Acids Res.* **31**, 248–250.
31. Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M. & Cesareni, G. (2002). MINT: a Molecular INTeraction database. *FEBS Letters*, **513**, 135–140.
32. Braunger, J., Schleithoff, L., Schulz, A. S., Kessler, H., Lammers, R., Ullrich, A. *et al*. (1997). Intracellular signaling of the Ufo/Axl receptor tyrosine kinase is mediated mainly by a multi-substrate docking-site. *Oncogene*, **14**, 2619–2631.
33. Diella, F., Cameron, S., Gemund, C., Linding, R., Via, A., Kuster, B. *et al*. (2004). Phospho. ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, **5**, 79.
34. Janssen, J. W., Schulz, A. S., Steenvoorden, A. C., Schmidberger, M., Strehl, S., Ambros, P. F. & Bartram, C. R. (1991). A novel putative tyrosine kinase receptor with oncogenic potential. *Oncogene*, **6**, 2113–2120.
35. Rochlitz, C., Lohri, A., Bacchi, M., Schmidt, M., Nagel, S., Fopp, M. *et al*. (1999). Axl expression is associated with adverse prognosis and with expression of Bcl-2 and CD34 in *de novo* acute myeloid leukemia (AML): results from a multicenter trial of the Swiss Group for Clinical Cancer Research (SAKK). *Leukemia*, **13**, 1352–1358.
36. Stitt, T. N., Conn, G., Gore, M., Lai, C., Bruno, J., Radziejewski, C. *et al*. (1995). The anticoagulation factor protein S and its relative, Gas6, are ligands for the Tyro 3/Axl family of receptor tyrosine kinases. *Cell*, **80**, 661–670.
37. Rohrschneider, L. R., Bourette, R. P., Lioubin, M. N., Algate, P. A., Myles, G. M. & Carlberg, K. (1997). Growth and differentiation signals regulated by the M-CSF receptor. *Mol. Reprod. Dev.* **46**, 96–103.
38. Lioubin, M. N., Myles, G. M., Carlberg, K., Bowtell, D. & Rohrschneider, L. R. (1994). Shc, Grb2, Sos1, and a 150-kilodalton tyrosine-phosphorylated protein form complexes with Fms in hematopoietic cells. *Mol. Cell. Biol.* **14**, 5682–5691.
39. Yin, T. & Green, K. J. (2004). Regulation of desmosome assembly and adhesion. *Semin. Cell Dev. Biol.* **15**, 665–677.
40. Batzer, A. G., Rotin, D., Urena, J. M., Skolnik, E. Y. & Schlessinger, J. (1994). Hierarchy of binding sites for Grb2 and Shc on the epidermal growth factor receptor. *Mol. Cell. Biol.* **14**, 5192–5201.
41. Lorch, J. H., Klessner, J., Park, J. K., Getsios, S., Wu, Y. L., Stack, M. S. & Green, K. J. (2004). Epidermal growth factor receptor inhibition promotes desmosome assembly and strengthens intercellular adhesion in squamous cell carcinoma cells. *J. Biol. Chem.* **279**, 37191–37200.
42. Bottino, C., Falco, M., Parolini, S., Marcenaro, E., Augugliaro, R., Sivori, S. *et al*. (2001). NTB-A [correction of GNTB-A], a novel SH2D1A-associated surface molecule contributing to the inability of

natural killer cells to kill Epstein–Barr virus-infected B cells in X-linked lymphoproliferative disease. *J. Expt. Med.* **194**, 235–246.
43. Pearce, A. C., Senis, Y. A., Billadeau, D. D., Turner, M., Watson, S. P. & Vigorito, E. (2004). Vav1 and vav3 have critical but redundant roles in mediating platelet activation by collagen. *J. Biol. Chem.* **279**, 53955–53962.
44. Aoukaty, A. & Tan, R. (2005). Role for glycogen synthase kinase-3 in NK cell cytotoxicity and X-linked lymphoproliferative disease. *J. Immunol.* **174**, 4551–4558.
45. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al*. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
46. Nioche, P., Liu, W. Q., Broutin, I., Charbonnier, F., Latreille, M. T., Vidal, M. *et al*. (2002). Crystal structures of the SH2 domain of Grb2: highlight on the binding of a new high-affinity inhibitor. *J. Mol. Biol.* **315**, 1167–1177.
47. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
48. Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S. *et al*. (2004). UniProt: the Universal Protein knowledgebase. *Nucl. Acids Res.* **32**, D115–D119.
49. Li, C., Iosef, C., Jia, C. Y., Han, V. K. & Li, S. S. (2003). Dual functional roles for the X-linked lymphoproliferative syndrome gene product SAP/SH2D1A in signaling through the signaling lymphocyte activation molecule (SLAM) family of immune receptors. *J. Biol. Chem.* **278**, 3852–3859.
50. Canutescu, A. A., Shelenkov, A. A. & Dunbrack, R. L., Jr (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**, 2001–2014.
51. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935.
52. Wang, J. M., Cieplak, P. & Kollman, P. A. (2000). How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **21**, 1049–1074.
53. Suenaga, A., Hatakeyama, M., Ichikawa, M., Yu, X., Futatsugi, N., Narumi, T. *et al*. (2003). Molecular dynamics, free energy, and SPR analyses of the interactions between the SH2 domain of Grb2 and ErbB phosphotyrosyl peptides. *Biochemistry,* **42**, 5195–5200.
54. Hartigan, J. (1975). *Clustering Algorithms*, Wiley, New York.
55. Witten, I.H. & Frank, F. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations,* Morgan Kaufmann, San Francisco, CA.
56. Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data *via* the em algorithm. *J. Roy. Stat. Soc.* **39**, 1–38.
57. Li, W., Jaroszewski, L. & Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
58. Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Church, D. M. *et al*. (2005). Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.* **33**, D39–D45.
59. IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). (1984). Nomenclature and Symbolism for Amino Acids and Peptides. *Eur. J. Biochem.* **138**, 9–37.

Edited by F. E. Cohen

(Received 20 July 2005; received in revised form 29 November 2005; accepted 5 January 2006)
Available online 26 January 2006