Motif Decomposition of the Phosphotyrosine Proteome Reveals a New N-terminal Binding Motif for SHIP2*[S]

Martin Lee Miller‡, Stefan Hanke§, Anders Mørkeberg Hinsby‡, Carsten Friis‡, Søren Brunak‡, Matthias Mann§, and Nikolaj Blom‡¶

Advances in mass spectrometry-based proteomics have yielded a substantial mapping of the tyrosine phosphoproteome and thus provided an important step toward a systematic analysis of intracellular signaling networks in higher eukaryotes. In this study we decomposed an uncharacterized proteomics data set of 481 unique phosphotyrosine (Tyr(P)) peptides by sequence similarity to known ligands of the Src homology 2 (SH2) and the phosphotyrosine binding (PTB) domains. From 20 clusters we extracted 16 known and four new interaction motifs. Using quantitative mass spectrometry we pulled down Tyr(P)-specific binding partners for peptides corresponding to the extracted motifs. We confirmed numerous previously known interaction motifs and found 15 new interactions mediated by phosphosites not previously known to bind SH2 or PTB. Remarkably, a novel hydrophobic N-terminal motif ((L/V/I)(L/V/I)pY) was identified and validated as a binding motif for the SH2 domain-containing inositol phosphatase SHIP2. Our decomposition of the *in vivo* Tyr(P) proteome furthermore suggests that two-thirds of the Tyr(P) sites mediate interaction, whereas the remaining third govern processes such as enzyme activation and nucleic acid binding. ***Molecular & Cellular Proteomics 7:181–192, 2008.***

Phosphorylation-dependent protein-protein interaction is one of the key organizing principles in intracellular signaling events. The phosphotyrosine binding (PTB)[1] domain and the Src homology 2 (SH2) domain are modular domains that typically bind phosphotyrosine (Tyr(P))-containing peptides (1, 2). "Linear motifs" (unstructured sequence recognition patches with conserved residues at specific positions (3)) that direct Tyr(P)-dependent interaction have traditionally been studied using degenerate oriented peptide libraries. Such studies revealed that PTB and SH2 domains have preference for specific amino acids N- and C-terminal to the Tyr(P) residue, respectively (4, 5).

Recent methodological developments in MS-based proteomics have made it possible to identify hundreds to thousands of protein phosphorylation sites in a single project (6–14). Extensive mapping of the phosphoproteome is an important step toward analyzing the regulatory components of the cell. Because the majority of newly identified phosphopeptides are uncharacterized with respect to signaling context, there is now a unique opportunity to mine the phosphoproteome for novel phosphorylation motifs. Methods have been developed that successfully mine for overrepresented motifs from large protein data sets in general (15–17) and more recently also from phosphoproteomics data sets (18). However, these methods do not partition the data set into smaller subsets with high sequence similarity prior to motif extraction. Sequence patches flanking the motif also govern phosphorylation-dependent recognition (19); consequently there is a risk of extracting false positive motifs from functionally unrelated peptides. Furthermore the above mentioned methods are *in silico* approaches and do not combine prediction with experimental validation.

To overcome such limitations in the area of Tyr(P) motif discovery and classification, one may partition the data set into smaller subsets *e.g.* by sequence similarity with known kinase or binding substrates prior to motif extraction. Thus, the risk of retrieving false positive motifs is minimized because overrepresented motifs are extracted from peptides closely related in sequence and function. Besides Tyr(P) recognition motifs for kinases and interaction domains, there may also potentially exist Tyr(P) motifs that mediate other processes than binding such as *e.g.* enzyme activation and nucleic acid binding. Thus, it is essential to validate the extracted motifs both by experimental and bioinformatical means to obtain a functional classification.

With this in mind we developed a motif extraction and validation methodology and classified Tyr(P) motifs on a pro-

From the ‡Center for Biological Sequence Analysis, Technical University of Denmark, Kemitorvet, Building 208, DK-2800 Lyngby, Denmark and §Department of Proteomics and Signal Transduction, Max Planck Institute for Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

Received, May 24, 2007, and in revised form, October 5, 2007
Published, MCP Papers in Press, October 15, 2007, DOI 10.1074/mcp.M700241-MCP200

[1] The abbreviations used are: PTB, phosphotyrosine binding; PAM, partitioning around medoids; ITIM, immunoreceptor tyrosine-based inhibition motif; SH2, Src homology 2; GO, Gene Ontology; IPI, International Protein Index; zf-C_2H_2, Cys_2-His_2 zinc finger protein; N-WASP, neural Wiskott-Aldrich syndrome protein; PI3K, phosphatidylinositol 3-kinase; GAP, GTPase-activating protein; STAT, signal transducers and activators of transcription; PIP_3, phosphatidylinositol 3,4,5-trisphosphate.

teome level. Operating in sequence space, we stretched the MS-mapped Tyr(P) peptides over a backbone of ligands already known to be involved in Tyr(P)-dependent interaction. Using experimentally verified Tyr(P) ligands of the PTB and SH2 domains as both a clustering backbone and as a control for the partitioning, we split a literature-extracted data set of mammalian Tyr(P) peptides into 20 different clusters. We obtained a meaningful clustering because the controls partitioned correctly into separate clusters.

From the 20 clusters we extracted both known and unknown phosphorylation motifs, and peptides matching these motifs were synthesized and assayed for phosphorylation-specific interaction partners using a peptide pulldown assay based on quantitative proteomics (20). In contrast to the oriented peptide library approach that uses artificial degenerate peptides, we used naturally occurring peptides as baits to pull down binding partners from the cell lysate. Moreover because the interaction partners are in competition for binding, mimicking the *in vivo* binding situation, the risk of finding kinetically unfavorable interaction motifs is minimized. Finally this technique can potentially identify new types of domains with modification-specific binding capability.

Using the pulldown assay we identified the expected binding partners for numerous known C-terminally directed SH2 domain motifs. We also found 15 new phosphorylation-dependent interactions mediated by phosphosites not previously shown to direct interaction. Surprisingly we identified a new N-terminal hydrophobic motif ((L/V/I)(L/V/I)pY where pY is phosphotyrosine) for the SH2 domain-containing inositol phosphatase SHIP2. The specificity of the motif was confirmed by mutational analysis. Surprisingly this motif is N-terminally directed, which is in contrast to previous observations showing that binding of prototypical SH2 domains are directed by C-terminal recognition (21). Until now the only other known SH2 domain binding motif that is partly directed by N-terminal recognition is the immunoreceptor tyrosine-based inhibition motif (ITIM) (I/L/V)*X*pY*XX*(I/L/V) (22, 23).

On a proteome level we analyzed which Gene Ontology (GO) categories were overrepresented in proteins matching the extracted motifs. We found that motifs that mediate interaction in the pulldown assay are typically found in proteins involved in signal transduction, whereas non-binding motifs are found in enzymes and ion- and nucleic acid-binding proteins. Thus, we estimate that one-third of the *in vivo* Tyr(P) sites are not directly involved in interaction via domains such as SH2 and PTB but rather are sites that could alter the catalytic activity of enzymes or modulate the DNA binding affinity of *e.g.* transcription factors.

EXPERIMENTAL PROCEDURES

Data Set Preparation—Large scale data sets of tyrosine phosphorylation sites mapped in MS/MS experiments with mammalian cell lines were collected from the literature (8, 10, 12–14, 24, 25) yielding a total of 847 tyrosine phosphorylation sites. To filter out phosphopeptides from closely related homologs and orthologs only unique 13-mer peptide sequences with the Tyr(P) centrally positioned were considered. This reduced the MS-based data set to 481 phosphopeptides distributed in 380 proteins. Furthermore 162 experimentally verified Tyr(P) peptide ligands of one PTB domain and 10 different SH2 domains were extracted from the Phospho.ELM database (26). The 162 peptides were included in the data set as positive controls, resulting in a data set of 643 Tyr(P) peptides (see supplemental Table 3). The criteria for selecting the positive controls were the existence of a consensus binding motif and that a suitable amount of examples could be obtained.

Generation of Weight Matrices—13-mers of the 162 phosphopeptide ligands of the 11 respective PTB and SH2 domains (see Table I) were used to create 11 weight matrices using the weight matrix mode of EasyGibbs 1.0 (27). Default settings were used except motif length was set to 13 fixed around the central Tyr(P) residue. Subsequently all phosphopeptides in the MS-based data set (481) and the positive control data set (162) were scored by each of the 11 weight matrices, and thus each phosphopeptide could be represented as a vector of the 11 weight matrix scores.

Clustering Using Partitioning around Medoids (PAM)—A matrix consisting of the 11 weight matrix scores and the 643 phosphopeptides was generated and subsequently clustered by the PAM method (28) using the cluster package in R. The PAM algorithm is a robust version of *k*-means, and it searches for a specified number of medoids (representatives), *k*, around which clusters are constructed. The clusters are generated by minimizing the sum of the dissimilarities of all observations and assigning them to their closest medoid. Using a hypergeometrical test (see "Statistics") the optimal number of clusters ($k = 20$) was inferred because this resulted in the best partitioning of the positive controls. We use z-scores, *i.e.* multiples of standard deviations from the mean, to account for the different numeric ranges of the measured parameters.

The choice of an appropriate clustering algorithm is a complex one because no given algorithm is universally superior (29, 30). Rather the best choice will depend on the data set and in particular on what constitutes a good distance measure for it. Another relevant concern is the desired outcome and whether a hierarchical or partitional result is preferable. Many sophisticated methods exist that are capable of automatically determining the number of "natural clusters" in the data like the popular density-based clustering algorithms that can describe very complex non-circular relations in the data (31). It is, however, not clear whether the ability to recognize non-circular structures in the data is beneficial in this case. Proteins that share the same features are likely to be related and will form a circular relation in feature space. On the other hand, an elongated cluster in feature space will contain proteins that share only some features but not others, and the biological implications thereof can be quite diverse. Other than being computationally effective and easy to implement, the PAM algorithm was selected because it satisfies the need for a robust clustering algorithm and because its reliance on an Euclidean distance measure ensures that the result can be easily interpreted. The primary weakness of PAM is the need to arbitrarily select a number of clusters for the data, which in this case is overcome by the mentioned application of the hypergeometric test.

Dendrogram and Sequence Logo Plots—Weight matrices of the peptides in the 20 clusters were made using positional weighting of the three residues flanking the central Tyr(P) residue (27) and used to calculate distance matrices as described previously (32). The distance matrices were used as input to the program neighbor from version 3.5 of PHYLIP (Phylogeny Inference Package). To estimate the significance of the neighbor-joining clustering we used the bootstrap method and estimated the consensus tree by bootstrapping for 1000 repetitions as described earlier (32).

The frequencies of amino acids at particular positions in each

cluster were calculated, and subsequently sequence logo plots were used for graphic visualization (33). Each position in the aligned sequences corresponds to a column in the logo plot. The height of the column represents the degree of conservation at that position, whereas the height of the individual letters is proportional to the relative frequency of this amino acid residue. The maximal height of the column for the 20-amino acid alphabet is $\log_2 20 = 4.32$ bits.

Extraction of Motifs and Selection of Peptides to Synthesize—The identified phosphomotifs in each of the 20 clusters were found using the publicly available TEIRESIAS pattern discovery tool from IBM Bioinformatics (17). Parameters were set so the extracted motifs were within a window of 13 residues centered on the phosphoresidue. The minimal number of literals in the motif was set to 4, and the amino acids were grouped according to their chemical nature (Ala/Gly, Asp/Glu, Phe/Tyr, Lys/Arg, Ile/Leu/Met/Val, Gln/Asn, Ser/Thr, Pro, Trp, His, and Cys (17)). For each of the 20 clusters the most abundant motif was selected, and subsequently one peptide matching the motif was chosen from the respective cluster. Because multiple peptides in each cluster matched the extracted motif, peptides from mouse and peptides not previously known to be involved in phosphorylation-dependent interaction were preferred. In the few cases (three) where mouse sequences could not be obtained, peptides from humans with high homology in mouse were chosen.

Gene Ontology Analysis—Gene Ontology categories were obtained from Gene Ontology Annotation mouse database version 29.0. The extracted motifs were matched to proteins in the International Protein Index (IPI) mouse proteome version 3.20. Using a hypergeometrical test (see "Statistics") with the total proteome as background we found the 10 most overrepresented GO terms in retrieved proteins. The hits were inspected manually, and the consensus GO term was assessed for each motif. For the purpose of the hypergeometrical test, each annotated GO category was taken to include all of its ancestral terms to avoid problems with diverging levels of annotation.

Statistics—To determine whether the positive controls were significantly overrepresented in specific clusters compared with the whole data set, hypergeometric sampling without replacement (34) was performed. The hypergeometric test is a statistical test used to describe the arbitrariness of a sampling without replacement from a background of true or false examples. The probability (p) to observe a given or more extreme situation by a pure coincidence is given by the hypergeometric distribution,

$$P(X = x|N,M,K) \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}} \quad \text{(Eq. 1)}$$

where N is the total number of peptides, M is the number of peptides in the given set, K is the number of peptides in a particular cluster, and x is the number of K that belongs to M. A Bonferroni correction was performed to correct for multiple comparisons. In the case of GO analysis, we performed the test once for each GO category present in the data and evaluated the probability of sampling the set of retrieved proteins from the background of the total proteome by mere chance, considering a protein 'true' or 'false' depending on whether it had been assigned the category in question. The end result of this test was one p value for each GO category, describing the degree of overrepresentation of that particular assignment in the retrieved set against the background of the entire proteome.

Cell Culture—Mouse C2C12 muscle cells were grown in arginine- and lysine-deficient Dulbecco's modified Eagle's medium with 10% dialyzed fetal bovine serum for at least five passages and then switched to 2% dialyzed fetal bovine serum to differentiate the cells for 8 days. In accordance with the stable isotope labeling by amino acids in cell culture (SILAC) procedure, one cell population was supplemented with normal isotopic abundance L-arginine (Sigma) and L-lysine, and the other was supplemented with >99% isotopic abundance [$^{13}C_6$,$^{15}N_4$]arginine and [$^{13}C_8$]lysine (Aldrich) as described previously (35). Thereby full labeling of all proteins was achieved.

Peptide Synthesis and Pulldowns—Desthiobiotinylated peptides were synthesized on a solid-phase peptide synthesizer using amide resin (Intavis, Koeln, Germany). All peptides were designed as 15-mers with the Tyr(P) residue placed centrally at position 7 or 8 expect for one peptide from cluster 1 (see Table I) that was 20 amino acids long. The peptides were synthesized with an N-terminal biotin and an SG dipeptide linker. Peptides were synthesized as pairs in phosphorylated and a non-phosphorylated "control" form. The identity and purity of the synthesized peptides was confirmed by mass spectrometric analysis. For pulldowns, 1.5 nmol of immobilized peptide was added to an average of 1.5 mg of cell lysate. Dynabeads MyOne Streptavidin were saturated with biotinylated peptide prior to incubation with cell lysates. Cells were lysed as described previously (36), and equal amounts of protein were incubated overnight with the respective immobilized peptides at 4 °C. After three rounds of washing with lysis buffer, beads of pulldown pairs with the phosphorylated form and control were combined (20), and bound proteins were eluted using 16 mM biotin. Eluted proteins were precipitated and subsequently digested with trypsin for LC-MS/MS analysis.

LC/MS/MS, Database Searching, and Quantitation—After reduction in 1 μg of DTT and alkylation with 5 μg of iodoacetamide the eluted proteins were in-solution digested with 1 μg of endoproteinase Lys-C (Wako) for 3 h at room temperature. Subsequently samples were diluted with 4 volumes of 50 mM NH_4HCO_3 and further digested with 1 μg of trypsin (Promega) overnight at room temperature. Peptide mixtures were desalted on stop and go extraction tips (37) and loaded onto reversed phase analytical columns for liquid chromatography (38). Peptides were eluted from the analytical column by a multistep linear gradient running from 2 to 40% acetonitrile in 100 min and sprayed directly into the orifice of an LTQ-FT or an LTQ-Orbitrap mass spectrometer (Thermo Electron, Bremen, Germany). Proteins were identified by MS/MS by information-dependent acquisition of fragmentation spectra of multiply charged peptides. The peak list was generated using in-house software, raw2msm version 1.2, with default settings. The identified proteins were then searched against the mouse IPI database using the Mascot (version 2.1.0) algorithm (39). The MS/MS ion search parameters were set as follows: enzyme specificity for trypsin, trypsin/Pro + AspPro; maximum number of missed cleavages, 2; fixed modification, carbamidomethylcysteine; variable modifications, oxidation (Met), *N*-acetyl (protein), deamidation (NQ), [$^{13}C_6$,$^{15}N_4$]arginine, [$^{13}C_8$]lysine, and pyro (N-terminal QC); mass tolerance for precursor ions, 5 ppm; fragment mass tolerance, 0.5 Da; database version, IPI_mouse mouse_v314 with 68,655 entries. Common contaminants like human keratins were manually added. No species-specific restrictions were used. MSQuant (SourceForge) was used for quantitation and spectra validation. MSQuant uses peak area and extracted ion chromatogram for quantitation.

Determination of Significant Binding Partners—Intensity ratios of labeled to unlabeled forms of each validated tryptic peptide and the associated average ratio for the whole protein were obtained by MSQuant. We used 'crossover' experiments in which the specific interaction partners were required to have inverse ratios compared with the 'normal' experiment (20). A significant binder was defined as a protein with a log ratio at least three standard deviations over the log average ratio ($>3\log(\sigma) + \log(\mu)$) of all the proteins identified in a pulldown experiment. Furthermore the binder had to be confirmed in at least two pulldown experiments (normal and crossover experiment), and we only report specific binders for the phosphopeptide

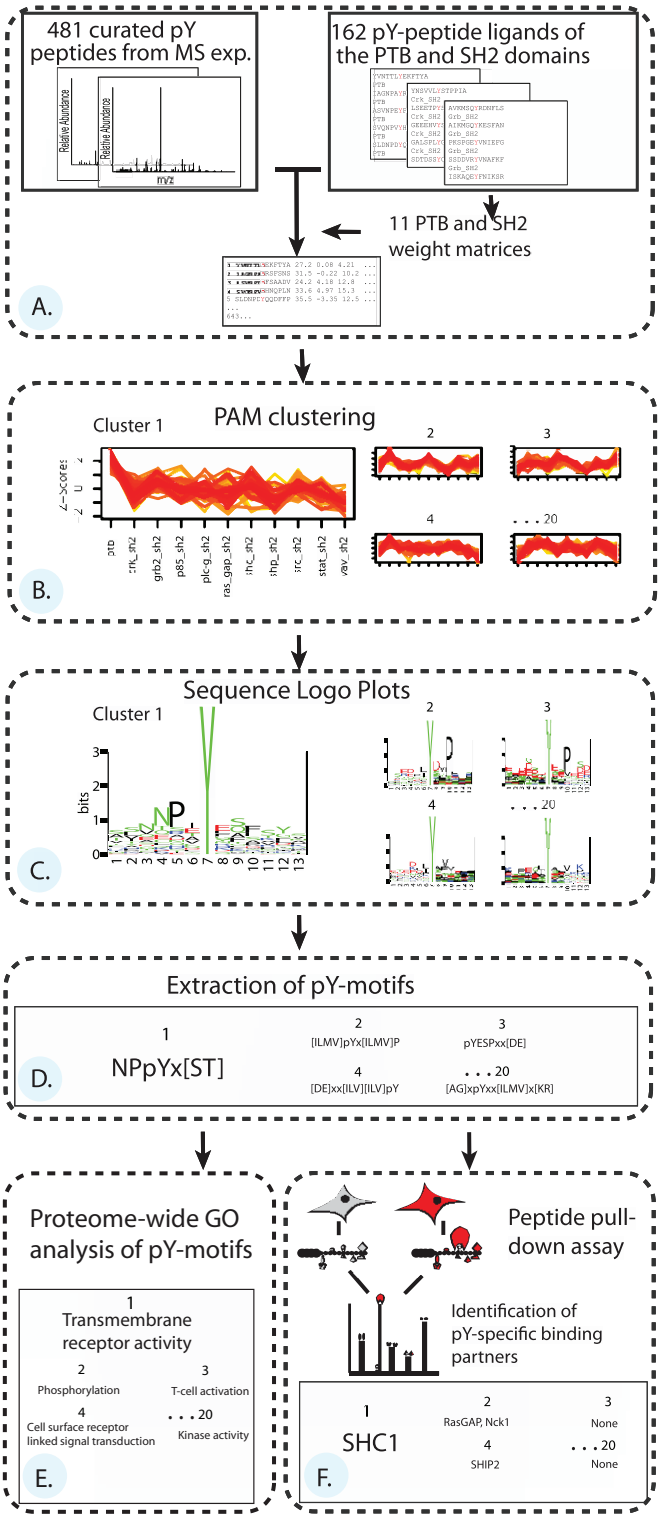

FIG. 1. **Overview of the motif extraction and validation strategy.** *A*, verified Tyr(P) peptide ligands (162) of different PTB and SH2 domains are used to create 11 specific weight matrices that subsequently score both a large scale data set of 481 Tyr(P) peptides identified in MS experiments and the 162 ligands themselves. All peptides are 13-mers with the Tyr(P) (*red*) centrally positioned. Consequently each phosphopeptide can be represented as a vector of the 11 weight matrix scores. *B*, this is used as input for the *k*-means-derived PAM algorithm (28), and the data set is split up into clusters (20, see text for details) depending on sequence similarity of the ligands used in the weight matrix generation. We use z-scores, *i.e.* multiples of standard deviations from the mean, to account for the different numeric ranges of the measured parameters. *C*, sequence logo plots (33) of the peptides in the different clusters are used to visually inspect the result of the partitioning (see Fig. 2). *D*, after partitioning the most conserved motif from each cluster is extracted (17). *E*, a GO analysis of the motifs on a proteome level is conducted to determine in which functional class of proteins the motifs generally are present compared with the whole proteome as background. *F*, a peptide pulldown assay based on quantitative proteomics (20) is used to evaluate the Tyr(P)-specific binding capability of the motifs. Peptides matching the motifs are synthesized in phosphorylated and non-phosphorylated pairs and used to pull down Tyr(P)-specific interaction partners from the lysate of C2C12 muscle cells.

and not the non-phosphorylated peptides. Finally at least one peptide had to have a score above 30, corresponding to $p < 0.05$. In the 64 pulldowns performed we identified a few sequence-unspecific binders with high affinity to either the phosphorylated or non-phosphorylated peptides (staphylococcal nuclease domain-containing protein, eukaryotic translation initiation factor, peptidylprolyl isomerase B, RNA-binding protein SiahBP, and RIKEN cDNA 2410104I19). These proteins were excluded because we consider them as false positive binders, *i.e.* they bind in a sequence-unspecific manner and occasionally bind most strongly to the non-phosphorylated peptide. In all the pulldown experiments an average of 140 ± 41 proteins was quantified with an average ratio of 1.217 ± 0.529.

RESULTS

A New Clustering Approach for Motif Extraction and Classification—To discover new interaction motifs in large aligned peptide data sets, we developed a method that partitions the sequences based on similarity with proteins previously known to be involved in interaction. The method is generally applicable for peptide interaction data sets. In this work we clustered and classified motifs in the Tyr(P) proteome. A flow chart of the method is presented in Fig. 1. In short, we used verified Tyr(P) ligands of the PTB and SH2 domains as a backbone to cluster a data set of uncharacterized phosphopeptides mapped in MS-based proteomics experiments with mammalian cell lines. From each clusters the most conserved motif was extracted. To classify the function of each motifs in an unbiased and systematic manner we conducted both experimental and bioinformatical investigations. Peptides matching the motifs were assayed for binding partners using a peptide-protein interaction screen based on quantitative proteomics (20, 36). On a proteome level we analyzed which GO categories were overrepresented in proteins matching the extracted Tyr(P) motifs. Thus, we established a workflow that extracts, verifies, and classifies motifs in phosphoproteome.

Clustering—In detail, we extracted 162 known 13-mer phosphopeptide ligands of 10 different SH2 domains and a general PTB domain from the Phospho.ELM database (26). In this database, which is the key repository for high quality

TABLE I

Clustering and motif extraction of the Tyr(P) proteome

The PAM algorithm was used to partition the data set by sequence similarity with the known ligands of the Tyr(P) binding domains (positive controls). The first and second columns show the cluster number and the size of the clusters, respectively. The ability of the algorithm to significantly partition ($p < 0.05$) the positive controls into different clusters is shown in the third column. For example, eight out of a total of 10 ligands of the Crk SH2 domain are grouped in cluster 2, corresponding to a significant overrepresentation in a Bonferonni corrected hypergeometric sampling test ($p < 3.67e-08$). The most conserved motif in each cluster was extracted and is stated in the fourth column. The Tyr(P) residue is indicated in bold, and an *"X"* represents any amino acid. Also indicated is the number of occurrences of the motif in the respective cluster and in the total data set. The identified motifs were matched to a library of known Tyr(P) binding motifs, and the expected binding partner is indicated (fifth column). 15–20-mer peptides matching the motifs were synthesized in pairs, one with a Tyr(P) as indicated in bold. Furthermore the parent protein of the peptide is given by Swiss-Prot entry name, and the position of the Tyr(P) is stated. The Tyr(P)-dependent interaction partner(s) identified in a quantitative proteomics peptide-protein screen is shown in the last column (see also Supplemental Table 1). PLC, phospholipase C.

Cluster	Size	Positive controls	Extracted motifs	Matched motifs, expected partner	PubMed ID	Peptides synthesized	Identified partners
1	41	PTB 10 of 10	NP*X***pY***X*(S/T) 7 of 7	SHC PTB (NP*X***pY**)	7542744 7541030	KEVCDGWSLPNPE**pY**YTLRYA ELMO2_MOUSE, 48	SHC
2	36	Crk SH2 8 of 10	(I/L/M/V)**pY***X*(I/L/M/V)P 8 of 14	Crk/RasGAP SH2 (**pY***XX*P)	9233798 11607838	KPSTDPL**pY**DTPDTRG RIN1_MOUSE, 35	RasGAP Nck1
3	29	Vav SH2 4 of 4	**pY**ESP*XX*(D/E) 5 of 5	Vav SH2 (**pY**ESP)	9151714	TETKTIT**pY**ESPQIDG E41L2_MOUSE, 889	None
4	28		(D/E)*XXX*(I/L/V)(I/L/V)**pY** 4 of 6	New motif		RETSKVI**pY**DFIEKTG WASL_MOUSE, 253	SHIP2
5	20		(D/E)(D/E)*XXX***pY***X*N 4 of 6	Grb2 SH2 (**pY***X*N)	11994738	VYDEDSP**pY**QNIKILH SPSY_MOUSE, 147	Grb2 RasGAP
6	41	Grb2 SH2 14 of 31	**pY***X*N(I/L/M/V)*XX*L 5 of 7	Grb2 SH2 (**pY***X*N)	11994738	ELFDDPS**pY**VNIQNLD SHC1_MOUSE, 423	Grb2
7	21	Grb2 SH2 6 of 31	(D/E)**pY***X*N(I/L/M/V) 4 of 11	Grb2 SH2 (**pY***X*N)	11994738	QPASVTD**pY**QNVSFSN ITSN2_HUMAN, 858	Grb2
8	45		**pY**(I/L/M/V)*X*M*X*P 4 of 10	p85-PI3K SH2 (**pY***XX*M)	7511210 11994738	PQRVDPNG**pY**MMMSPS IRS1_MOUSE, 658	p85-PI3Ka p85-PI3Kβ
9	40		**pY**(D/E)*X*(I/L/M/V)*X*(I/L/M/V) 5 of 22	Fps/Fes SH2 (**pY**(E/D)*X*(I/V))	7511210	AGKQKLQ**pY**EGIFIKD SF3A1_MOUSE, 757	None
10	32		(D/E)*XX***pY**(D/E)*X*(I/L/M/V) 7 of 27	Fps/Fes SH2 (**pY**(E/D)*X*(I/V))	7511210	DGGSDQN**pY**DIVTIGA INP4A_HUMAN, 355	None
11	35	PI3K SH2 16 of 24	**pY**(I/L/M/V)PM*X*P 6 of 7	p85 PI3K SH2 (**pY***XX*M)	7511210 11994738	NLHTDDG**pY**MPMSPGV IRS1_MOUSE, 608	p85-PI3Kα
12	23		D**pY**(I/L/M/V)*X*(I/L/M/V) 7 of 18	SHP2 SH2 (**pY**(I/V/L)*X*(I/V/L))	7680959	DLINRMD**pY**VEINIDH VIGLN_MOUSE, 437	SHP2
13	28	SHP2 SH2 7 of 12	(I/L/M/V)*X***pY**(I/L/M/V)*X*(I/L/M/V)D 6 of 7	SHP2 SH2 (**pY**(I/V/L)*X*(I/V/L))	7680959	DIKEKLC**pY**VALDFEQ ACTB_MOUSE, 218	SHP2
14	32	PLCγ SH2 5 of 16	(I/L/M/V)**pY***XX*(I/L/M/V)(I/L/M/V) 5 of 11	General/SHP2 SH2 (**pY**(I/V/L)*X*(I/V/L))	7511210 7680959	GKSKQPL**pY**SSIVTVE O88185_MOUSE, 948	SHP2 SHIP2
15	29	RasGAP SH2 7 of 8	(A/G)(I/L/M/V)**pY***XX*P 6 of 10	Crk/RasGAP SH2 (**pY***XX*P)	9233798 11607838	GVVDSGV**pY**AVPPPAE BCAR1_HUMAN, 410	RasGAP
16	37	SHC SH2 9 of 13	P*X*E**pY***XXXXX*(I/L/M/V) 3 of 3	New motif		TTEAPGEYFFSDGVR IMDH1_MOUSE, 400	None
17	25	Src SH2 7 of 14	**pY**(D/E)*X*(I/L/M/V)H 4 of 6	Fps/Fes SH2 (**pY**(E/D)*X*(I/V))	7511210	ELTAEFL**pY**DEVHPKQ TWF2_MOUSE, 309	RasGAP
18	32	STAT SH2 15 of 19	**pY**(I/L/M/V)PQ 4 of 4	STAT SH2 (**pY***XX*Q)	14966128	SGENFVP**pY**MPQFQTC LEPR_MOUSE, 1138	None
19	33		H(S/T)G*X*KP**pY***XC XX*CG 10 of 10	New motif		RIHTGEKP**pY**ECVQCGK ZNF24_MOUSE, 335	None
20	36		(A/G)*X***pY***XX*(I/L/M/V)*X*(K/R) 8 of 15	New motif		KKNRIAI**pY**ELLFKEG RS10_MOUSE, 12	None

annotated phosphorylation sites, these 11 domains had the highest number of annotated ligands (see Table I for details). Thus, the 162 substrates of the 11 domains represent the broadest available *in vivo* data set of Tyr(P) ligands. These ligands were aligned with the Tyr(P) centrally positioned and used to generate a position-specific weight matrix for each of the 11 domains. The 481 phosphopeptides in the MS-based data set and the 162 ligands of the PTB and SH2 domains ("positive controls") were then scored by each of the 11 weight matrices. Consequently each phosphopeptide was

represented as a vector of the 11 weight matrix scores (Fig. 1*A*). Based on this profile of vectors we clustered the total of 643 peptides using the *k*-means-derived PAM algorithm (40). The PAM algorithm searches for a predefined number of medoids (representatives) around which clusters are constructed. We tried various cluster sizes, and by using a hypergeometrical distribution test the number of clusters was set to 20 because this gave the statistically best partition of the positive controls into different clusters (Fig. 1, *B* and *C*).

We were able to obtain a convenient fit of the model because the positive controls were distributed such that all were statistically overrepresented in separate clusters ($p < 0.05$) (see Table I). For example, 10 of a total of 10 ligands of the PTB domains were grouped in cluster 1 ($p < 7.34e-12$). Only the 31 ligands for the Grb2 SH2 domain were split up in two significant groups (clusters 6 and 7). Furthermore eight of the 20 clusters did not contain a significant overrepresentation of the known ligands used.

Motif Extraction—From each of the 20 clusters the most conserved motif was extracted with the TEIRESIAS pattern discovery tool from IBM Bioinformatics (Fig. 1*D*) (17) as described under 'Experimental Procedures.' The 20 identified motifs are presented in Table I followed by the number of matches to peptides in the particular cluster and in the total data set. For example, a unique motif, NP*X*pY*X*(S/T), was extracted from cluster 1 because all peptides (seven) that matched this motif were in this cluster. To compare the 20 identified motifs with already characterized interaction motifs, we matched each motif to a comprehensive library of Tyr(P) interaction motifs in the Human Protein Reference Database (HPRD) (41). Using the same example again, the motif extracted from cluster 1 could be matched to the NP*X*pY motif described for the SHC and IRS-1 PTB domains (1), although our extended motif contains a Ser or Thr residue at position +2 from the Tyr(P) residue. Considering that the ligands (the positive controls) of the PTB domain were grouped in cluster 1, it is not surprising that we extracted the NP*X*pY*X*(S/T) motif in this cluster; however, only three of the seven matching peptides in cluster 1 were from these PTB ligands (data not shown). Because this was a general trend it shows the ability of the clustering method to gather previously uncharacterized peptides with high sequence similarity to known ligands of the different Tyr(P) binding domains. In total, 16 of the 20 identified motifs could be matched to the motif library, showing an overall consistency between the positive controls and the extracted and matched motifs. In four clusters new Tyr(P) motifs were identified.

Gene Ontology Analysis—We used a GO analysis of the 20 extracted motifs on a proteome level to determine in which type of proteins the extracted motifs generally are present (Fig. 1*E*). We retrieved all proteins in the mouse proteome that matched the motifs and used a hypergeometrical test with the total proteome as background and thereby found the 10 most overrepresented GO terms in retrieved proteins (see "Experimental Procedures" for details). Using the same example again, the NP*XYX*(S/T) motif from cluster 1 was overrepresented in proteins involved in processes like 'receptor activity' and 'intrinsic to membrane' with the consensus parent GO term assessed to be 'transmembrane receptor activity' (see supplemental Table 2). Thus, analyzing the motif on a proteome level indicates that proteins containing the motif are involved in early signaling transduction. This makes sense in that this is a motif for the PTB domain, which is found in proteins that function as molecular scaffolds and adaptors in signaling pathways (1).

Tyr(P)-specific Interaction Partners—To experimentally verify the 20 extracted motifs we used a phosphorylation-specific peptide-protein interaction screen (Fig. 1*F*) (20, 36). This assay is based on differential labeling of proteins using stable isotope labeling by amino acids in cell culture (SILAC) making it possible to distinguish specific binders from background binders by their isotope ratios determined by quantitative mass spectrometry (35, 42). The peptides are synthesized in a phosphorylated form and a non-phosphorylated form and used as baits to pull down competing binding partners from cell lysate, thus mimicking the *in vivo* binding situation.

We synthesized peptide pairs matching the 20 extracted motifs. If there were multiple matches we chose peptides with Tyr(P) sites not previously known to mediate interaction. Using this experimental approach we could test our clustering and motif extraction method and investigate the relevance of known motifs in a near *in vivo* situation, *i.e.* endogenous proteins competing for binding, and potentially discover binding partners for novel motifs. Again using cluster 1 as an example, we synthesized a peptide pair from the engulfment and cell motility protein 2 in which the Tyr-48 residue was either phosphorylated or non-phosphorylated. This is an uncharacterized phosphosite identified in a large scale phosphoproteomics study (12), and it has not previously been shown to direct Tyr(P)-dependent interaction. Using this peptide pair as bait we retrieved one specific binder with a ratio more than three standard deviations over the log mean of a total of 162 background binders. This protein, SHC-transforming protein 1 (SHC), which contains both a PTB and an SH2 domain, had a total Mascot score of 1654 with 15 identified peptides of which nine were quantifiable (see supplemental Table 1). Theoretically it could be either the SH2 or the PTB domain that binds the bait phosphopeptide; however, because the peptide matches the consensus NP*X*pY motif known to direct PTB domain binding, it is most likely that SHC binds to the peptide through its PTB domain.

Assaying the Tyr(P) Sequence Space for Interaction Partners—The identified phosphospecific binding partners of the representative peptides from each cluster can be seen in Table I. Besides the aforementioned SHC protein, these proteins all contain SH2 domains, making it very likely that this domain governed the phosphospecific interaction. The majority (13 of 20) of the peptides retrieved one or more interaction

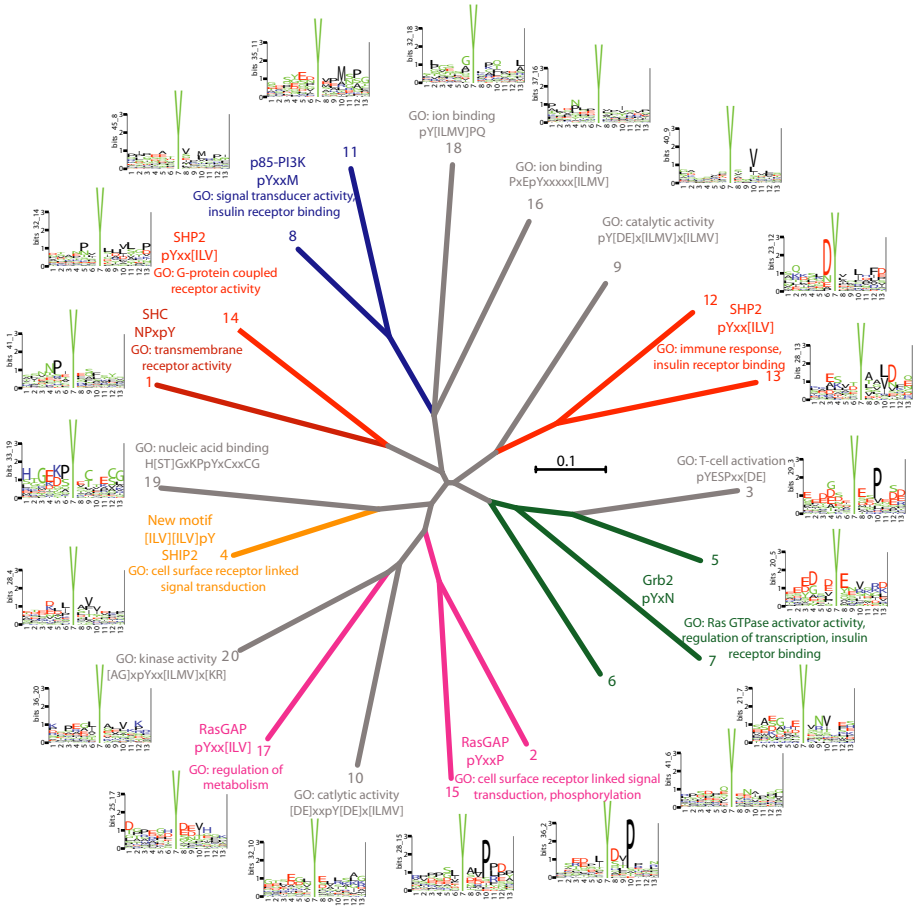

FIG. 2. **Dendrogram representing the *in vivo* Tyr(P) sequence space, motifs, and binding partners.** Peptides in the 20 different clusters are used to generate weight matrices that subsequently are used as input in a phylogenetic alignment. The tree is a consensus tree of 1000 bootstrap trees (32). The tree represents the distance between the clusters in sequence space, which is also visually illustrated by the sequence logo plots (33) of the peptides in each cluster. The *color* of the *branches* is based on the Tyr(P)-dependent interaction partners that were experimentally identified using a peptide-protein interaction screen. Both novel and previously known consensus motifs (see Table I, fifth column) that govern these specific interactions are indicated in the same color. Branches and extracted motifs are *gray* if no interaction partners were retrieved via the motifs. Furthermore overrepresented GO terms from proteins in the whole mouse proteome containing the motifs are stated (see text for details). Note that the motifs that do not retrieve a specific binding partner (*gray*) are typically found in proteins mediating processes such as ion binding, 'catalytic activity,' and nucleic acid binding.

proteins. Of these proteins seven were unique because some proteins were identified several times. This is not surprising because some of the clusters were close to each other in sequence space resulting in extraction of similar motifs and ultimately retrieving the same interaction partners.

To get an overview of the results of the pulldown experiments, the GO analysis, and the sequence similarity between the different clusters, we generated weight matrices of the peptides in the individual clusters and constructed a dendrogram based on an alignment of these matrices (32). The tree can be seen in Fig. 2 together with sequence logo plots where the height of each position represents the degree of conservation (33). The logo plots visually illustrate a successful partitioning because each cluster has a distinct pattern where particular amino acids are highly abundant at specific positions flanking the central Tyr(P) residue. The tree is colored according to the identified interaction partners retrieved by peptides matching the motifs in the different clusters. For example, clusters 8 and 11 are close in sequence space with an overrepresentation of hydrophobic residues, especially methionine, at position +3 from the central tyrosine residue. Rather than being distinct clusters, these are more likely to be subsets of the same motif. Thus, from these two clusters the motifs pY(I/L/M/V)*X*M*X*P and pY(I/L/M/V)PM*X*P were extracted and matched to the consensus pY*XX*M in the library of motifs. Accordingly the two peptide pairs synthesized from these clusters both retrieve the PI3K-p85α protein. In the same manner the majority of the expected interaction partners were identified using peptides matching to the well characterized C-terminally directed Tyr(P) motifs, such as pY*X*N, pY*XX*P, and pY(I/V/L)*X*(I/V/L), that retrieved Grb2, RasGAP, and SHP2, respectively. This illustrates the clear consistency between the sequence similarity of the clusters, the conserved residues in the motifs, and the interaction partners identified.

There were four clusters (clusters 3, 9, 10, and 18) where we did not identify the partners (Vav, Fps/Fes, and STAT) as expected from the signature of motifs alone (see Table I). For instance, we did not pull down the SH2 domain protein Fps/Fes when using peptides matching a pY(E/D)*X*(I/V/L) motif (clusters 9 and 10), which has previously been shown to direct binding (4). This motif was defined using *in vitro* oriented peptide library experiments, which have the inherent risk of defining motifs that are not relevant *in vivo*. Whether this is the case, the Fps/Fes protein was not present in the cell lysate, or because of technical limitations remains unclear. In total, here

we report 15 phosphorylation-dependent interactions mediated by phosphosites not previously known to direct protein interaction (see supplemental Table 1).

Non-binding Tyr(P) Motifs—We observed that motifs that do not mediate interaction in the pulldown assay are typically found in proteins involved in processes other than signal transduction (see Fig. 2). It is particularly interesting that we extracted a highly conserved H(S/T)G*X*KPpY*XC**XXX**CG motif from a number of closely related peptides from Cys_2-His_2 zinc finger proteins (zf-C_2H_2) concentrated in cluster 19 that did not pull down any phosphorylation-specific interaction partner. The phosphosite of the first tyrosine residue in the zf-C_2H_2 domain was also described recently in a study that mined for novel motifs in the phosphoproteome (18), although the identified motif in this work (E*XX*pY) was different from our top scoring motif (H(S/T)G*X*KPY*XC**XXX**CG), which does not contain an acidic residue in position −3. However, our second best motif in cluster 19 (H*X*GE*XX*pY) closely resembles that reported by Schwartz and Gygi (18).

The H(S/T)G*X*KPY*XC**XXX**CG motif is extremely specific for proteins containing the zf-C_2H_2 domain: of 33,758 proteins we retrieved 656 matches, all of which had the GO term 'nucleic acid binding' (GO:0003676) (see supplemental Table 2), whereas 647 of the 656 matches had the term 'zinc ion binding' (GO:0008270) ($p < 1e-100$).

Recently a role for phosphorylation of zf-C_2H_2 domains in inhibition of transcription has been suggested (43, 44), supposedly as a consequence of the negatively charged phosphomoiety that reduces DNA affinity (45). Indirectly our results support this; because we did not retrieve any interaction partner for the synthesized phosphopeptide matching the zinc finger motif, it is unlikely that this motif directs protein-protein interaction, but rather phosphorylation of this motif modulates nucleic acid binding.

Similarly the novel motifs from clusters 16 and 20 could mediate mechanisms other than protein-protein interactions, for example, a kinase motif that mediates enzyme activation, nucleic acid binding, protein folding, etc. Supporting this, we found these motifs to be present in proteins overrepresented in proteins with GO terms 'ion binding' and 'kinase activity,' respectively. Likewise the motifs from clusters where we did not find the expected partners (clusters 3, 9, 10 and 18) are all except motif 3 overrepresented in proteins involved in enzymatic processes and ion binding (see Fig. 2 or supplemental Table 2).

This indicates that the motifs not mediating protein binding could govern processes such as phosphorylation-dependent enzyme activation and nucleic acid binding. Taken together with the results from the pulldown experiment where 13 of 20 motifs mediated interaction, we estimate that one-third of the Tyr(P) motifs in the proteome mediate processes other than interaction through prototypic SH2 and PTB domains.

Identification of a New N-terminal Hydrophobic Tyr(P) Binding Motif for SHIP2—From cluster 4 we extracted a new N-terminal hydrophobic motif, (D/E)*XXX*(I/L/V)(I/L/V)pY. We synthesized a peptide pair from the neural Wiskott-Aldrich syndrome protein (N-WASP) matching the motif where Tyr-253 was either phosphorylated or non-phosphorylated. This phospho-Tyr-253 was identified in a large scale phosphoproteomics experiment (12). It has previously been shown that phosphorylation of Tyr-253 modulates localization of N-WASP from nucleus to cytoplasm, thereby possibly stimulating cell migration (46). Using the N-WASP peptide pair in our quantitative proteomics experiment, we found the SH2 domain-bearing inositol 5-phosphatase 2 (SHIP2) as a significant binding partner with 12 peptides (ratio, 9.5 ± 1.5) of 131 protein background partners (ratio, 1.2 ± 0.9) (see supplemental Table 1). To address the specificity of this N-terminal hydrophobic phosphomotif we repeated the experiment with the only difference being that the two hydrophobic residues were mutated to alanines (AApY) and paired this peptide with the wild type N-WASP phosphopeptide (VIpY) (Fig. 3). The SHIP2 protein bound specifically to the wild type phosphopeptide (ratio, 10.2 ± 3) but not the mutated phosphopeptide, confirming the specificity of the hydrophobic motif. In a third interaction experiment we obtained similar results by using a phosphopeptide in which all three residues in the extracted motif were mutated (see supplemental Table 1). This supports the notion that recognition is based on two hydrophobic amino acids adjacent to the Tyr(P), and we could confine the motif to (I/L/V)(I/L/V)pY.

We next extracted three additional peptides from the same cluster matching the (I/L/V)(I/L/V)pY motif. Two of three of these phosphopeptide pairs also specifically pulled down SHIP2 (see supplemental Table 1). Together these experiments show that this indeed is a generic binding motif for SHIP2.

Similarly to previous studies using this phosphospecific pulldown method (20, 36) in this work we only retrieved specific interaction partners containing either an SH2 or a PTB domain. Because SHIP2 contains an SH2 domain it is highly unlikely that interaction between SHIP2 and the phosphopeptides could be mediated by domains other than the SH2 domain.

Because the observed interaction seems to be one of the first examples where N-terminal residues of the ligands guide SH2 domain binding, we wanted to exclude the possibility that the peptide was bound in a reverse fashion. We synthesized the N-WASP peptide pair with the reverse sequence (GTKEIFDpYIVKSTER) and found that SHIP2 was not retrieved using this scrambled peptide pair as bait (see supplemental Table 1). Thus, the assay has directional specificity, and only the two hydrophobic residues N-terminal and not C-terminal from the Tyr(P) direct SH2 domain-mediated SHIP2 binding.

To investigate the significance of the motif on a proteome level, we used the aforementioned GO analysis and found that proteins containing the motif indeed are involved in signal transduction. The (I/L/V)(I/L/V)Y is not particularly specific in

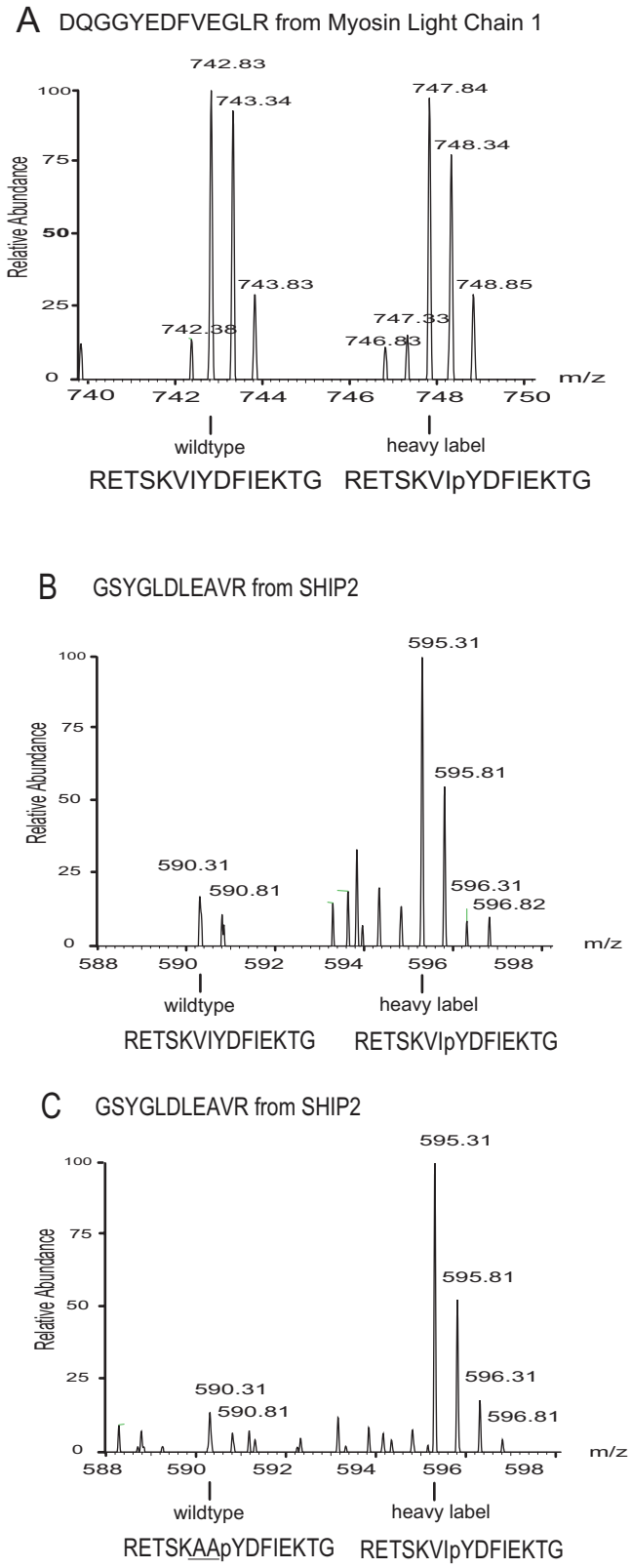

FIG. 3. **Identification of a new hydrophobic N-terminal motif for SHIP2.** *A*, mass spectrum of a doubly charged [$^{13}C_6$,$^{15}N_4$]Arg-labeled tryptic peptide from myosin light chain 1 pulled down with the RETSKVIYDFIEKTG peptide from N-WASP. The myosin light chain 1 protein is an unspecific binding protein because the ratio of the ion intensities of the two differentially labeled peptides is 1:1. *B*, mass spectrum of a doubly charged [$^{13}C_6$,$^{15}N_4$]Arg-labeled tryptic peptide from SHIP2. SHIP2 specifically binds to the phosphorylated peptide from N-WASP as indicated by the larger peak intensity of the labeled peptides. *C*, when mutating the VIpY motif to AApY and comparing these two 15-mer phosphopeptides directly in a pulldown assay, SHIP2 is still retrieved as a specific binding partner to the VIpY motif, confirming the specificity of the hydrophobic motif immediately N-terminal to the Tyr(P).

itself because it matches 3618 times, which is about 10% of all proteins in the mouse proteome. However, of the 3618 proteins, 2495 can be backtracked to the term 'cell surface receptor-linked signal transduction' (GO:0007166) ($p < 1e-100$). In summary, the N-terminal hydrophobic motif we characterized mediates SHIP2 interaction and is generally overrepresented in proteins involved in signal transduction.

DISCUSSION

Currently there is a unique opportunity to mine and classify Tyr(P) motifs in the proteome. 1) A significant portion of the Tyr(P) sites have been mapped by MS-based proteomics experiments. 2) Although not exhaustively, classical biochemical studies have pinpointed interaction between SH2 and PTB domain-containing proteins and specific Tyr(P) ligands. 3) Advances of high throughput validation methods, like the peptide pulldown assay used in this study, make it possible to validate dozens of interaction motifs. Together these developments constitute the basis for discovering and validating Tyr(P) motifs on a global level.

Methodological Considerations—Compared with previously published methods that mine for motifs in large scale phosphoproteomics data sets we partitioned by similarity with functionally characterized peptides prior to motif extraction. By this clustering approach we obtained high resolution and can extract meaningful patterns from functionally and physically related groups of peptides. We used the binding ligands of PTB and SH2 domains both as a clustering backbone and as positive controls in the clustering and could consequently obtain a satisfactory fit of the model. Ideally as more interaction data become available one would use independent controls; however, this was not possible due to data limitations. For all the 20 extracted motifs we could either detect a match to an existing Tyr(P) motif, retrieve a binding partner, or obtain a meaningful GO term for all proteins containing the motif, and thus, we estimate that we have an insignificantly low false positive motif extraction rate in our method.

In the few cases where we did not retrieve the expected binding partner from the Tyr(P) motif alone, *i.e.* Fps/Fes, STAT, and Vav, it can be speculated that 1) the motifs, originally defined *in vitro*, could be low affinity motifs *in vivo*, 2) the proteins were not expressed in the C2C12 muscle cell line used in this study, or 3) it could be due to a technical limitation in our assay.

Implications for a New Motif for SHIP2—One of our findings was that the SH2 domain-containing inositol 5-phosphatase SHIP2 binds to a novel N-terminal hydrophobic motif, (I/L/V)(I/L/V)pY. The specificity of this motif was confirmed by mutational analysis. A scrambled peptide pair with the reverse sequence and thus with the motif C-terminal of the Tyr(P) did not retrieve SHIP2, confirming that the recognition lies on the N-terminal side of the Tyr(P). SHIP2 is also retrieved by two other peptide pairs matching the motifs, indicating this is a generic motif for SHIP2 binding.

In general, the C-terminal residues Tyr(P) +1 and +3 are considered as the most important for the binding specificity of prototypic SH2 domains (1, 2). Interestingly the SH2 binding motif of SHIP2 that we describe is N-terminal, indicating that the peptide binding groove of some SH2 domains also accommodates residues N-terminal of the Tyr(P). In agreement with this, the binding properties of the tandem SH2 domains of the protein-tyrosine phosphatase SHP-2 are governed by residues Tyr(P) −2 to +5 (47). SHP-1 and SHP-2 both bind the (I/L/V)*X*pY*XX*(I/L/V) ITIM motif in the cytoplasmic part of Fc receptors, and the Tyr(P) −2 hydrophobic residues have specifically been shown to mediate binding (23, 48, 49). In contrast to the prototypic SH2 domains that have an Arg in the αA2 binding pocket groove, the tandem SH2 domains of the SHP-1 and SHP-2 phosphatases instead have Gly (50). Supposedly this creates a gap that is filled by the side chain of the Tyr(P) −2 residue of the bound peptide. Supporting this hypothesis, it has been shown that a single point mutation in αA2 Gly → Arg disrupts the Tyr(P) −2-mediated binding specificity of SHP-2 (48). Furthermore it is known that signaling lymphocytic activation molecule-associated protein (SAP) and Eat2 SH2 domains in part are directed by N-terminal binding to ITIM motifs (51).

The binding motif of the SH2 domain of SHIP2 has not previously been investigated using degenerate peptide binding experiments; however, the ITIM motif has also been reported as a docking point for the SH2 domain of SHIP2 (52, 53). Combined with our observations, this indicates that the binding specificity of SHIP2 may be conferred by hydrophobic residues immediately upstream of the Tyr(P) with contributions from downstream hydrophobic patches. To our knowledge, other than that of SHP, this is the only other reported N-terminally directed Tyr(P) binding motif.

It could be speculated that the SHIP inositol phosphatases could have a binding mechanism similar to that of SHP protein-tyrosine phosphatases. However, in contrast to SHP the SH2 domains of the SHIP phosphatases resemble the prototypic SH2 domain because they have the highly conserved αA2 Arg (50), indicating that the N-terminally directed binding mechanisms differ between SHIP and SHP phosphatases. Because the crystal structure of the SHIP2 phosphatase with a bound ligand has not been resolved, the specific binding mechanisms have yet to be described.

SHIP2 is involved in membrane signaling by dephosphorylating the 5′-phosphate group of the key secondary messenger phosphatidylinositol 3,4,5-trisphosphate (PIP_3). Through this action SHIP2 inhibits PI3K-mediated receptor tyrosine kinase signaling because PIP_3 is generated by PI3K (54). The new motif we describe for the SH2 domain of SHIP2 fits into this overall function because a GO term analysis of all proteins containing the (I/L/V)(I/L/V)pY motif revealed that these proteins are involved in cell surface receptor-linked signal transduction. Presumably SHIP2 could bind to a number of yet unidentified membrane signaling proteins through its SH2 domain and in this way be translocated to the membrane or act cooperatively with these proteins. Negative regulators of PI3K and PIP_3 are attractive as antiobesity and diabetes drug targets because PI3K is the main effector of insulin signaling. Recently a role for SHIP2 as a candidate for such therapeutic intervention has been proposed by studies of SHIP2 knock-out mice (55, 56). Thus, the new motif described in this report may not only be involved in regulation of SHIP2-mediated signal transduction but could also be relevant for medical use.

Conclusions—This work presents the first system-wide approach to mine the proteome for Tyr(P) interaction motifs using both bioinformatics methods and experimental validation. Strikingly 16 of the 20 motifs extracted could be matched to previously described interaction motifs. Our experimental validation shows that the majority of the Tyr(P) interaction motifs that previously have been defined *in vitro* are also able to pull down interaction partners from complex lysate mixtures. The GO analysis revealed that motifs that mediate interaction in the pulldown assay are found in proteins involved in signal transduction, whereas remaining non-binding motifs are found in enzymes and ion- and nucleic acid-binding proteins. This raises the intriguing possibility that about one-third of the *in vivo* Tyr(P) sites are not directly involved in interaction via domains such as SH2 and PTB but rather are sites that could alter the catalytic activity of enzymes or modulate the DNA binding affinity of *e.g.* transcription factors.

Perspectives—The developed clustering method is applicable to other types of complex large scale data sets of post-translational modification where a substantial amount of peptide-protein interactions have been identified. As MS-based methods map more modifications such as acetylation and methylation sites, such interactomes could be mined for conserved motifs and assayed for binding partners in a similar manner. Combining proteomics and bioinformatics enables one to do large scale screens in an unbiased way and thus allows one to reconfirm previous knowledge and discover new mechanisms at the same time. As mass spectrometric streamlining and automation advances, entire post-translational modification proteomes can be mapped, binding motifs can identified, and thus ultimately the specificity of signaling networks can be unraveled.

Acknowledgments—We thank members of the Department of Proteomics and Signal Transduction and the Center for Biological Sequence Analysis for comments on the manuscript.

* This work was supported by Interaction Proteome, FP6, Contract LSHG-CT-2003-505520, a grant from the European Commission in the 6th framework program; by the Danish Platform for Integrative Biology, a grant from the Danish National Research Foundation (Danmarks Grundforskningsfond); and by the Danish Research Agency (Forskningsstyrelsen). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

¶ To whom correspondence should be addressed. Tel.: 45-4525-2477; Fax: 45-4593-1585; E-mail: nikob@cbs.dtu.dk or nblm@novozymes.com.

REFERENCES

1. Yaffe, M. B. (2002) Phosphotyrosine-binding domains in signal transduction. *Nat. Rev. Mol. Cell Biol.* **3,** 177–186
2. Pawson, T. (2004) Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell* **116,** 191–203
3. Bork, P., and Koonin, E. V. (1996) Protein sequence motifs. *Curr. Opin. Struct. Biol.* **6,** 366–376
4. Songyang, Z., Shoelson, S. E., McGlade, J., Olivier, P., Pawson, T., Bustelo, X. R., Barbacid, M., Sabe, H., Hanafusa, H., and Yi, T. (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
5. Songyang, Z., Margolis, B., Chaudhuri, M., Shoelson, S. E., and Cantley, L. C. (1995) The phosphotyrosine interaction domain of SHC recognizes tyrosine-phosphorylated NPXY motif. *J. Biol. Chem.* **270,** 14863–14866
6. Ballif, B. A., Villen, J., Beausoleil, S. A., Schwartz, D., and Gygi, S. P. (2004) Phosphoproteomic analysis of the developing mouse brain. *Mol. Cell. Proteomics* **3,** 1093–1101
7. Beausoleil, S. A., Jedrychowski, M., Schwartz, D., Elias, J. E., Villen, J., Li, J., Cohn, M. A., Cantley, L. C., and Gygi, S. P. (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 12130–12135
8. Brill, L. M., Salomon, A. R., Ficarro, S. B., Mukherji, M., Stettler-Gill, M., and Peters, E. C. (2004) Robust phosphoproteomic profiling of tyrosine phosphorylation sites from human T cells using immobilized metal affinity chromatography and tandem mass spectrometry. *Anal. Chem.* **76,** 2763–2772
9. Ficarro, S., Chertihin, O., Westbrook, V. A., White, F., Jayes, F., Kalab, P., Marto, J. A., Shabanowitz, J., Herr, J. C., Hunt, D. F., and Visconti, P. E. (2003) Phosphoproteome analysis of capacitated human sperm. Evidence of tyrosine phosphorylation of a kinase-anchoring protein 3 and valosin-containing protein/p97 during capacitation. *J. Biol. Chem.* **278,** 11579–11589
10. Ficarro, S. B., Salomon, A. R., Brill, L. M., Mason, D. E., Stettler-Gill, M., Brock, A., and Peters, E. C. (2005) Automated immobilized metal affinity chromatography/nano-liquid chromatography/electrospray ionization mass spectrometry platform for profiling protein phosphorylation sites. *Rapid Commun. Mass Spectrom.* **19,** 57–71
11. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127,** 635–648
12. Rush, J., Moritz, A., Lee, K. A., Guo, A., Goss, V. L., Spek, E. J., Zhang, H., Zha, X. M., Polakiewicz, R. D., and Comb, M. J. (2005) Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.* **23,** 94–101
13. Salomon, A. R., Ficarro, S. B., Brill, L. M., Brinker, A., Phung, Q. T., Ericson, C., Sauer, K., Brock, A., Horn, D. M., Schultz, P. G., and Peters, E. C. (2003) Profiling of tyrosine phosphorylation pathways in human cells using mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 443–448
14. Zheng, H., Hu, P., Quinn, D. F., and Wang, Y. K. (2005) Phosphotyrosine proteomic study of interferon α signaling pathway using a combination of immunoprecipitation and immobilized metal affinity chromatography. *Mol. Cell. Proteomics* **4,** 721–730
15. Jonassen, I., Collins, J. F., and Higgins, D. G. (1995) Finding flexible patterns in unaligned protein sequences. *Protein Sci.* **4,** 1587–1595
16. Nevill-Manning, C. G., Wu, T. D., and Brutlag, D. L. (1998) Highly specific protein sequence motifs for genome analysis. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 5865–5871
17. Rigoutsos, I., and Floratos, A. (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics* **14,** 55–67
18. Schwartz, D., and Gygi, S. P. (2005) An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol.* **23,** 1391–1398
19. Forman-Kay, J. D., and Pawson, T. (1999) Diversity in protein recognition by PTB domains. *Curr. Opin. Struct. Biol.* **9,** 690–695
20. Schulze, W. X., and Mann, M. (2004) A novel proteomic screen for peptide-protein interactions. *J. Biol. Chem.* **279,** 10756–10764
21. Songyang, Z., and Cantley, L. C. (1995) Recognition and specificity in protein tyrosine kinase-mediated signalling. *Trends Biochem. Sci.* **20,** 470–475
22. Vely, F., and Vivier, E. (1997) Conservation of structural features reveals the existence of a large family of inhibitory cell surface receptors and non-inhibitory/activatory counterparts. *J. Immunol.* **159,** 2075–2077
23. Burshtyn, D. N., Yang, W., Yi, T., and Long, E. O. (1997) A novel phosphotyrosine motif with a critical amino acid at position −2 for the SH2 domain-mediated activation of the tyrosine phosphatase SHP-1. *J. Biol. Chem.* **272,** 13066–13072
24. Hinsby, A. M., Olsen, J. V., Bennett, K. L., and Mann, M. (2003) Signaling initiated by overexpression of the fibroblast growth factor receptor-1 investigated by mass spectrometry. *Mol. Cell. Proteomics* **2,** 29–36
25. Hinsby, A. M., Olsen, J. V., and Mann, M. (2004) Tyrosine phosphoproteomics of fibroblast growth factor signaling: a role for insulin receptor substrate-4. *J. Biol. Chem.* **279,** 46438–46447
26. Diella, F., Cameron, S., Gemund, C., Linding, R., Via, A., Kuster, B., Sicheritz-Ponten, T., Blom, N., and Gibson, T. J. (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* **5**, 79
27. Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S., Lamberth, K., Buus, S., Brunak, S., and Lund, O. (2004) Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* **20,** 1388–1397
28. Kaufman, L., and Rousseeuw, P. J. (1990) *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley, New York
29. Fraley, C., and Raftery, A. E. (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput. J.* **41,** 578–588
30. Jain, A. K., Murth, M. N., and Flynn, P. J. (1999) Data clustering: a review. *ACM Comput. Surv.* **31,** 264–323
31. Ester, M., Kriegel, H. P., Sander, S., and Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise, in *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, Portland, August 2–4, 1996* (Simoudis, E., Han, J., and Fayyad, U., eds) Vol. 1, pp. 226–231, Association for the Advancement of Artificial Intelligence (AAAI) Press, Menlo Park, CA
32. Lund, O., Nielsen, M., Kesmir, C., Petersen, A. G., Lundegaard, C., Worning, P., Sylvester-Hvid, C., Lamberth, K., Roder, G., Justesen, S., and Buus, S., and Brunak, S. (2004) Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics* **55,** 797–810
33. Schneider, T. D., and Stephens, R. M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18,** 6097–6100
34. Johnson, L. N., Kotz, S., and Kemp, A. W. (1992) *Univariate Discrete Distributions*, 2nd Ed., Wiley-Interscience, New York
35. Ong, S. E., Kratchmarova, I., and Mann, M. (2003) Properties of ^{13}C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC). *J. Proteome Res.* **2,** 173–181
36. Schulze, W. X., Deng, L., and Mann, M. (2005) Phosphotyrosine interactome of the ErbB-receptor kinase family. *Mol. Syst. Biol.* **1,** 2005 0008
37. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75,** 663–670
38. Ishihama, Y., Rappsilber, J., Andersen, J. S., and Mann, M. (2002) Microcolumns with self-assembled particle frits for proteomics. *J. Chromatogr. A* **979,** 233–239
39. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999)

Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
40. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. (2002) The Pfam protein families database. *Nucleic Acids Res.* **30,** 276–280
41. Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T. K., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H. N., Rashmi, B. P., Ramya, M. A., Zhao, Z., Chandrika, K. N., Padma, N., Harsha, H. C., Yatish, A. J., Kavitha, M. P., Menezes, M., Choudhury, D. R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S. K., Madavan, V., Joseph, A., Wong, G. W., Schiemann, W. P., Constantinescu, S. N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G. C., Dang, C. V., Garcia, J. G., Pevsner, J., Jensen, O. N., Roepstorff, P., Deshpande, K. S., Chinnaiyan, A. M., Hamosh, A., Chakravarti, A., and Pandey, A. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13,** 2363–2371
42. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386
43. Dovat, S., Ronni, T., Russell, D., Ferrini, R., Cobb, B. S., and Smale, S. T. (2002) A common mechanism for mitotic inactivation of C2H2 zinc finger DNA-binding domains. *Genes Dev.* **16,** 2985–2990
44. Saydam, N., Adams, T. K., Steiner, F., Schaffner, W., and Freedman, J. H. (2002) Regulation of metallothionein transcription by the metal-responsive transcription factor MTF-1: identification of signal transduction cascades that control metal-inducible transcription. *J. Biol. Chem.* **277,** 20438–20445
45. Jantz, D., and Berg, J. M. (2004) Reduction in DNA-binding affinity of Cys2His2 zinc finger proteins by linker phosphorylation. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 7589–7593
46. Wu, X., Suetsugu, S., Cooper, L. A., Takenawa, T., and Guan, J. L. (2004) Focal adhesion kinase regulation of N-WASP subcellular localization and function. *J. Biol. Chem.* **279,** 9565–9576
47. Eck, M. J., Pluskey, S., Trub, T., Harrison, S. C., and Shoelson, S. E. (1996) Spatial constraints on the recognition of phosphoproteins by the tandem SH2 domains of the phosphatase SH-PTP2. *Nature* **379,** 277–280
48. Huyer, G., and Ramachandran, C. (1998) The specificity of the N-terminal SH2 domain of SHP-2 is modified by a single point mutation. *Biochemistry* **37,** 2741–2747
49. Beebe, K. D., Wang, P., Arabaci, G., and Pei, D. (2000) Determination of the binding specificity of the SH2 domains of protein tyrosine phosphatase SHP-1 through the screening of a combinatorial phosphotyrosyl peptide library. *Biochemistry* **39,** 13251–13260
50. Liu, B. A., Jablonowski, K., Raina, M., Arce, M., Pawson, T., and Nash, P. D. (2006) The human and mouse complement of SH2 domain proteins-establishing the boundaries of phosphotyrosine signaling. *Mol. Cell* **22,** 851–868
51. Poy, F., Yaffe, M. B., Sayos, J., Saxena, K., Morra, M., Sumegi, J., Cantley, L. C., Terhorst, C., and Eck, M. J. (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
52. Muraille, E., Bruhns, P., Pesesse, X., Daeron, M., and Erneux, C. (2000) The SH2 domain containing inositol 5-phosphatase SHIP2 associates to the immunoreceptor tyrosine-based inhibition motif of Fc γRIIB in B cells under negative signaling. *Immunol. Lett.* **72,** 7–15
53. Bruhns, P., Vely, F., Malbec, O., Fridman, W. H., Vivier, E., and Daeron, M. (2000) Molecular basis of the recruitment of the SH2 domain-containing inositol 5-phosphatases SHIP1 and SHIP2 by fcγ RIIB. *J. Biol. Chem.* **275,** 37357–37364
54. Damen, J. E., Liu, L., Rosten, P., Humphries, R. K., Jefferson, A. B., Majerus, P. W., and Krystal, G. (1996) The 145-kDa protein induced to associate with Shc by multiple cytokines is an inositol tetraphosphate and phosphatidylinositol 3,4,5-triphosphate 5-phosphatase. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 1689–1693
55. Marion, E., Kaisaki, P. J., Pouillon, V., Gueydan, C., Levy, J. C., Bodson, A., Krzentowski, G., Daubresse, J. C., Mockel, J., Behrends, J., Servais, G., Szpirer, C., Kruys, V., Gauguier, D., and Schurmans, S. (2002) The gene INPPL1, encoding the lipid phosphatase SHIP2, is a candidate for type 2 diabetes in rat and man. *Diabetes* **51,** 2012–2017
56. Sleeman, M. W., Wortley, K. E., Lai, K. M., Gowen, L. C., Kintner, J., Kline, W. O., Garcia, K., Stitt, T. N., Yancopoulos, G. D., Wiegand, S. J., and Glass, D. J. (2005) Absence of the lipid phosphatase SHIP2 confers resistance to dietary obesity. *Nat. Med.* **11,** 199–205