Kinase peptide specificity: Improved determination and relevance to protein phosphorylation

Koichi Fujii*†‡, Guozhi Zhu*‡, Yin Liu*, John Hallam*, Lin Chen§, Juan Herrero§, and Stephen Shaw*¶

*Experimental Immunology Branch, National Cancer Institute, Bethesda, MD 20892; §AxCell Biosciences, Newtown, PA 18940; and †First Department of Internal Medicine, University of Occupational and Environmental Health, Kitakyushu, Fukuoka 807-8555, Japan

Edited by Tony Hunter, The Salk Institute for Biological Studies, La Jolla, CA, and approved August 4, 2004 (received for review March 17, 2004)

Specificity of phosphorylation is critical to signal transduction. Recent emphasis on colocalization of substrate and kinase has eclipsed emphasis on peptide specificity, i.e., kinase preference for particular amino acids surrounding the phosphorylation site. We describe an approach to determining peptide specificity by using positional scanning of biotinylated oriented peptide libraries and insights emerging from those determinations. We accurately determine preference (or disfavor) for residues at a given substrate position (such as P+2) by comparison of *in vitro* phosphorylation of peptide libraries differing by a single residue at that position. By analysis of all positions near the phosphorylation site, position-specific scoring matrices are generated and used both to understand the basis of specificity and to predict phosphorylation. PKC-δ and -ζ predictions have been validated rigorously by comparisons with measured phosphorylation. The results demonstrate specificity and sensitivity (80–90%) much better than the previous predictive method. These predictions can be accessed at http://mpr.nci.nih.gov. The accuracy of the specificity determination allows identification of an important difference in peptide specificity between these closely related kinases; Ile/Leu at the P−1 position is disfavored by PKC-ζ but not PKC-δ. Our findings and visual representation of peptide specificity highlight the importance of disfavored residues. Finally, analysis of 124 experimentally determined PKC sites from the literature demonstrates a very strong role of peptide specificity in many of those sites. Thus, position-specific scoring matrices generated by this method provide a foundation for quantitative analyses of kinase specificity and improved predictions of previously determined physiologically relevant phosphorylation sites.

Protein phosphorylation is one of the most fundamental mechanisms for signal transduction. Consequently, protein kinases are among the largest classes of human genes, encompassing ≈600 members (1). Specificity of signal transduction depends on the ability of each kinase to precisely phosphorylate particular sites on specific substrate proteins. Although many phosphorylation sites have been identified, for few of them is it certain which kinase phosphorylates them. Analysis of kinase specificity therefore continues to be an important area of investigation. The specificity of phosphorylation is determined by at least two major elements: peptide specificity and regulated colocalization of substrate/kinase (2–6).

Peptide specificity is the preference of a kinase for particular amino acids surrounding the phosphorylation site. Considerable information exists about peptide specificity, including detailed characterization of selected substrates; however, this information has not been shown to allow reliable prediction of phosphorylation of unstudied peptides. PKC is one of the best-studied families of basophilic kinases (2, 3), whose specificity has been studied for two decades (6–11). A major step toward quantitative understanding of its peptide specificity was provided by Cantley and coworkers (12), who described a powerful strategy based on phosphorylation of a single degenerate peptide to quantitatively characterize peptide specificity of a kinase; we refer to this method as amino acid sequencing of phosphorylated oriented peptide library (ASP-OPL). Yaffe and coworkers (13, 14) then created a conceptual framework for probabilistic predictions of phosphorylation sites based on those approaches and provided access to such predictions on the World Wide Web (http://scansite.mit.edu). The accuracy of such predictions must be highly dependent on the accuracy of the experimentally determined parameters. A critical missing element has been systematic comparison of predictions with experimental results. We have developed an approach to accurately determine such parameters, which is applicable to analysis of diverse kinases. Herein we describe the method, rigorously validate the results for PKC-δ and -ζ, and demonstrate the relevance of that information to predicting sites of PKC phosphorylation of intact proteins.

Materials and Methods

Peptide Synthesis, Analysis, and Quantitation. Peptides were synthesized as C-terminal amides on Mimotopes SynPhase Rink (Clayton, Australia) amide acrylic-grafted polypropylene solid support by using fluorenylmethoxycarbonyl (Fmoc) chemistry. For degenerate positions, a solution containing a mixture of 19 amino acids was used; the proportions in the mix were based on approximations of residue abundance in regions around published basophilic phosphorylation sites (eight parts each of A, G, S, and T; six parts each of D, E, K, L, P, Q, and R; four parts each of F, H, I, N, V, and Y; and two parts each of M and W). To minimize biases in residue addition from the degenerate mix, a molar ratio of 1× was used rather than the 2× that was used for nondegenerate peptides. Peptides were synthesized with an N-terminal biotin-Dansyl(Lys)-Gly spacer. Purity of peptides was validated by means of high-throughput liquid chromatography-MS, and yield was quantified by measurement of absorbance at 330 nm.

Kinase Assay. PKC-δ, PKC-ζ, and protein kinase A (PKA) were purchased from Calbiochem. PKC-θ tagged at its N terminus with hexahistidine was expressed in baculovirus and purified by nickel affinity chromatography. Ingredients in the kinase assay were typically 5–20 ng of kinase, 10 μM peptide, 1 μCi of [γ-^{32}P]ATP (1 Ci = 37 GBq), 100 μM unlabeled ATP, and 100 ng/ml phorbol 12-myristate 13-acetate in a total volume of 50 μl of kinase buffer [100 mM Hepes/0.05% Triton X-100/1 mM $CaCl_2$/20 mM $MgCl_2$/0.2 mg/ml phosphatidylserine, pH 7.0 (Avanti Polar Lipids)]. Kinase concentration and assay times (typically 10 min) were chosen so that stoichiometry of peptide phosphorylation was <10%. Fifty microliters of stop solution (0.1 M ATP/0.1 M EDTA in water, pH 8) was added to terminate the reaction. Ten microliters was transferred to Reacti-Bind Streptavidin High Binding Capacity Coated Plates (Pierce); this volume corresponds to 50 pmol of peptide, which

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: OPL, oriented peptide library; ASP-OPL, amino acid sequencing of phosphorylated OPL; PS-OPL, positional scanning of OPL; PSSM, position-specific scoring matrix; PKA, protein kinase A.

‡K.F. and G.Z. contributed equally to this work.

¶To whom correspondence should be addressed. E-mail: sshaw@nih.gov.

a

PKC-Delta Analysis at Substrate Position P+2				
Symbolic Representation of peptides: ddddRdd-S-F?d				
Peptide Number	Peptide Sequence	Counts	Ratio to Mean	Log Score
#1	ddddRdd-S-F R d	695 ± 10	2.7	1.4 ± 0.03
#2	ddddRdd-S-F K d	575 ± 45	2.2	1.2 ± 0.2
#3	ddddRdd-S-F H d	483 ± 67	1.9	0.9 ± 0.3
#4	ddddRdd-S-F F d	312 ± 2	1.2	0.3 ± 0.01
#5	ddddRdd-S-F L d	273 ± 8	1.1	0.1 ± 0.1
#6	ddddRdd-S-F N d	425 ± 20	1.6	0.7 ± 0.1
#7	ddddRdd-S-F S d	263 ± 7	1.0	0.0 ± 0.1
#8	ddddRdd-S-F G d	121 ± 4	0.5	-1.1 ± 0.1
#9	ddddRdd-S-F A d	140 ± 5	0.5	-0.9 ± 0.1
#10	ddddRdd-S-F Q d	163 ± 7	0.6	-0.7 ± 0.1
#11	ddddRdd-S-F W d	347 ± 22	1.3	0.4 ± 0.1
#12	ddddRdd-S-F P d	94 ± 11	0.4	-1.5 ± 0.2
#13	ddddRdd-S-F D d	137 ± 2	0.5	-0.9 ± 0.03

b

Residue at Query Pos.	Residue Extrapolated	Log Score for Residue at Indicated Position in Substrate												
		P-7	P-6	P-5	P-4	P-3	P-2	P-1	P0	P+1	P+2	P+3	P+4	P+5	P+6
R		1.0	1.2	0.8	1.9	1.5	1.2	0.7		1.1	1.4	0.8	1.0	0.0	0.7
K		0.7	0.7	0.6	0.8	1.0	0.6	0.7		1.0	1.2	0.6	0.8	0.4	0.5
H		0.2	0.3	0.3	0.3	0.8	1.3	0.0		0.2	0.9	0.2	0.6	0.4	0.0
F	Y	0.0	-0.1	0.3	0.3	0.0	-0.7	-0.1		1.0	0.3	0.2	-0.3	-0.1	0.2
L	IMV	-0.1	-0.5	0.7	0.0	-0.6	-0.4	0.0		1.2	0.1	0.6	-0.7	0.3	-0.7
N		0.1	-0.2	-0.5	-0.4	0.0	-0.1	0.4		-0.8	0.7	0.2	-0.1	0.5	0.0
Q		-0.4	-0.2	-0.3	-0.4	-0.2	0.0	-0.2		0.2	-0.7	-0.6	-0.2	0.1	-0.5
S	T	0.0	0.0	0.1	-0.6	0.3	-0.3	-0.4	0.0	0.3	0.0	-0.5	0.3	-0.1	-0.2
G		0.4	0.0	-0.3	0.0	-0.4	-0.1	0.2		-0.4	-1.1	-0.3	-0.1	-0.2	0.0
A		-0.1	-0.4	-0.3	-0.4	-0.4	-0.5	0.3		-0.4	-0.9	0.6	-0.2	-0.2	0.1
W		-0.7	-0.4	-0.3	0.2	0.0	-0.4	-0.6		0.0	0.4	-0.1	-0.7	-0.5	-0.9
P		0.4	0.2	-0.1	-0.5	-0.7	0.1	-0.2		-2.4	-1.5	-0.8	0.1	0.0	-0.1
D	E	-1.1	-0.5	-1.0	-1.2	-1.0	-0.5	-0.9		-1.3	-0.9	-0.7	-0.3	-0.4	-0.7
T		0.0													

Fig. 1. PS-OPL method for determining the PSSM for peptide phosphorylation by PKC-δ, with details shown for the P+2 position. (*a*) Preferences at the P+2 position. Biotinylated peptides were synthesized corresponding to the 13 sequences shown. All peptides share three fixed residues: an R at P−3, an S at P0 (shown as -S-), and an F at P+1. The amino acid fixed in the query (?) position (in this example, P+2, boxed in red) systematically varies between peptides. The remaining degenerate positions (d) are synthesized with a mixture of amino acids. Peptides (10 μM) were phosphorylated *in vitro* by PKC-δ, captured on streptavidin-coated plates, and washed, and ^{32}P incorporation was determined. Results for each peptide are shown as raw counts; ratio-to-mean calculated as (counts)/(geometric mean counts for all 13 peptides); and log score calculated as $\log_2$(ratio-to-mean). (*b*) PSSM for PKC-δ from positions P−7 to P+6. Each value corresponds to a log score from experimentally determined phosphorylation of a different degenerate peptide; data for the P+2 column come from *a*. Data for other columns (such as P−1) were derived in the same fashion, except that the query position for the corresponding degenerate peptides was at a different position (e.g., at P−1). The anchor residues are always as shown in *a* (i.e., Rdd-S-F), except where the query position would coincide with an anchor position, i.e., P−3 (?Rd-S-F); P0 (Rdd-?-F) and P+1 (RRd-S-?). For efficiencies in synthesis and testing, scores for seven residues were not determined experimentally. The six such residues in column ''Residue Extrapolated'' were assigned scores equal to a tested residue with similar physicochemical properties (and considered to be a conservative mutation during protein evolution). A score of zero was assigned for cysteine. Red fill is used for scores ≥ 0.5 and blue fill for scores less than or equal to −0.5.

is reliably and completely captured by the wells. After exhaustive rinses with PBS/0.05% Tween 20, the wells were air-dried, 50 μl of microScint-20 (Packard) was added, and emissions were measured in a TopCount NXT Microplate (Packard).

Data Analysis. Results shown represent at least duplicate (and usually quadruplicate) determinations. The position-specific scoring matrix (PSSM) logo was generated by using POSTSCRIPT files generated by Visual Basic code; some POSTSCRIPT code was adapted from Tom Schneider's MAKELOGO (version 8.69, www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi).

Results

Positional Scanning of Oriented Peptide Library (PS-OPL) and Graphical Representation of Specificity. Peptide specificity of PKC-δ was determined by a method we call PS-OPL. The method ascertains specificity by comparing the amount of phosphorylation of degenerate peptides that differ by single amino acids. Consider the set of peptides used to analyze the P+2 position and the results with that set (Fig. 1*a*). All peptides in the set have a fixed S at P0 (the phosphorylation site) and ''anchor residues'' (R at P−3 and F at P+1) previously determined to be strongly preferred residues for PKC. The only difference between them is the residue fixed at the P+2 position (boxed in red, Fig. 1*a*). Other positions are degenerate (d), i.e., synthesized with a mixture of amino acids at those positions; therefore, each ''peptide'' is a mix of ≈1 billion peptide sequences, each of which has four residues in common (e.g., Rdd-S-FR for peptide #1). The results demonstrate differences among the peptides in phosphorylation by PKC-δ under standardized assay conditions. For example, peptide #1 having an R at the P+2 position is phosphorylated 7-fold more than peptide #12, which is identical except for a P at the P+2 position. This differential phosphorylation estimates PKC-δ's average preference at P+2 because peptides #1 and #12 are large pools of peptides that differ only at P+2. The experimental data then are converted into two useful derived parameters. ''Ratio to mean'' is >1.0 when the residue is favored (such as 2.7 for R) and <1.0 when the residue is disfavored (such as 0.4 for P). ''Log score'' (calculated as the $\log_2$ of ratio to mean) has the useful property that the score of a strongly disfavored residue (such as −1.5 for P) is roughly equal and opposite to that of a strongly favored residue (such as +1.4 for R). Inspection of the results reveals the following: (*i*) the method can provide highly quantitative values for residue preference; (*ii*) R, K, H, and N are residues strongly favored by PKC-δ at P+2; and (*iii*) P, D, Q, G, and A are strongly disfavored at P+2. The log scores for residues at the P+2 position (Fig. 1*a*) are part of a comprehensive set of log scores determined for all substrate positions between P−7 and P+6 (Fig. 1*b*); that set is referred to as a PSSM (15).

To facilitate understanding of the data in a PSSM, we created a compact graphical representation (''PSSM logo''); a key difference between a PSSM logo and the previous analogous approach (16) is that the PSSM logo visually emphasizes the disfavored residues as much as the favored residues. The PSSM logo for PKC-δ (Fig. 2*a*) reflects the data in the PSSM (Fig. 1*b*). Each stack of letters represents residues at a position in the substrate. In each stack, strongly favored residues are at the top and strongly disfavored residues are at the bottom; residues that are neither strongly favored nor disfavored (i.e., have a score near 0) are therefore minimized in visual importance. The tall columns correspond to substrate positions at which the kinase shows large discrimination between different substrate amino acids. Inspection of the PSSM logo (Fig. 2*a*) and PSSM (Fig. 1*b*) reveals key features of PKC-δ's specificity. PKC favors basic residues at all positions; the strongest basophilic preferences are between P−4 and P+4; hydrophobic preference is most pronounced at P+1. The PSSM also highlights disfavored residues not previously emphasized. Of particular note, the acidic residue tested (D) is disfavored at all substrate positions. Moreover, there is a singular disfavor for proline at the P+1 position (G.Z., K.F., Y.L., M. James, J.H., and S.S., unpublished data).

To explore differences among PKC isoforms, PKC-ζ was analyzed by using the PS-OPL approach and the same degenerate peptides used for PKC-δ. Although atypical PKCs (ζ and ι) are considered members of the PKC family, they are the most divergent in sequence from the other nine PKC isoforms (≈50% identity and 60% similarity in the kinase domain). The PSSM derived by PS-OPL for PKC-ζ (Fig. 2*b*) is similar to that of PKC-δ (Fig. 2*a*): like δ, ζ is characterized by generalized

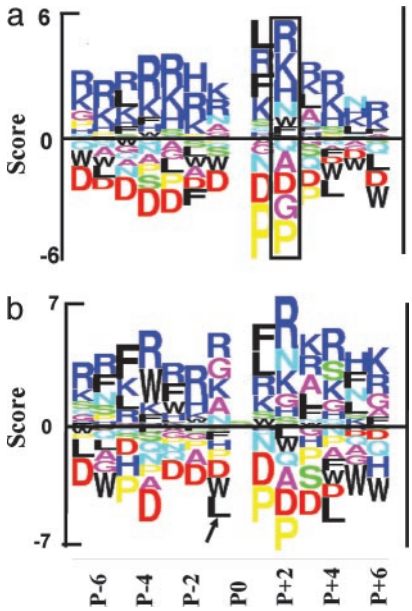

Fig. 2. PSSM logo representation of results for PKC-δ and PKC-ζ. (*a*) PSSM logo for PKC-δ. This PSSM logo is a ''graph'' of the scores in the PKC-δ PSSM in Fig. 1*b*. Each stack of letters represents log scores for residues at a single substrate position; for example, the P+2 column (boxed) is derived from scores shown in Fig. 1*a*. The height of each letter is proportional to the absolute value of the residue's log score, and the positions of the letters in the stack are sorted from bottom to top in ascending value by the log score. Thus, strongly favored residues are at the top and strongly disfavored residues are at the bottom; residues that are neither strongly favored nor disfavored are minimized in visual importance. Each residue code is colored to indicate its residue's physicochemical properties: blue, basic; red, acidic; black, hydrophobic; and yellow, P. (*b*) PSSM logo for PKC-ζ. Black arrow, disfavored residue L analyzed in Fig. 4.

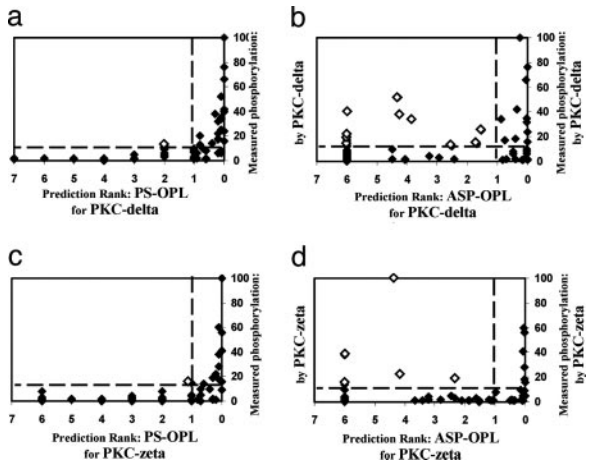

Fig. 3. Validation of predictions for PKC-δ and -ζ by comparison of predicted vs. measured *in vitro* phosphorylation of peptides. A panel was synthesized of 75 nondegenerate peptides that are candidate or reported PKC sites (see also Fig. 8). Prediction ranks were determined from PSSMs derived from the PS-OPL method for PKC-δ (*a*), PKC-ζ (*c*), or the ASP-OPL method (http://scansite.mit.edu) for PKC-δ (*b*) or PKC-ζ (*d*). Results are expressed as relative phosphorylation, normalized to 100% for the most highly phosphorylated peptide. Dashed lines indicate a cutoff of first percentile for positive prediction and >10% for positive phosphorylation. Each diamond indicates results for a single peptide; open diamonds indicate false negatives (i.e., positive phosphorylation but weak or negative prediction).

preference for basic residues as well as generalized disfavor for acidic residues from P−7 to P+6. Close inspection also reveals differences; for example, PKC-δ has a disfavor for hydrophobic residues at P−2, whereas PKC-ζ has a marked disfavor for hydrophobic residues at P−1 (see below).

Rigorous Validation. Yaffe *et al*. (13) developed a simple approach to rank phosphorylation sites for a kinase based on its PSSM (see Fig. 7, which is published as supporting information on the PNAS web site). Ranking of phosphorylation sites represents a prediction from the PSSM that is subject to experimental validation. We have tested the PKC-δ and PKC-ζ PSSMs by comparing predicted and measured phosphorylation on a panel of peptides. We used a panel of 75 nondegenerate peptides that are candidate or reported PKC sites; they are ''proteomic peptides'' because their sequences correspond to amino acid sequences encoded in the human genome (see Fig. 8, which is published as supporting information on the PNAS web site). The PKC-δ PSSM is remarkably accurate in its predictions on phosphorylation of most of the peptides (Fig. 3*a*). For practical purposes, cutoffs chosen were >10% for positive phosphorylation and first percentile for prediction rank; however, the general conclusions are unaffected by variations in choice of these cutoff values. Fig. 3*b* shows an identical comparison, except that the prediction used is the Scansite prediction based on a PKC-δ PSSM derived by the previously described ASP-OPL method (11, 12). Limitations in the previous prediction are apparent because many of the best substrates are not in the top percentile, and 6 of 25 of the positive peptides have a rank lower than the fifth percentile. The PS-OPL-based prediction is better than the previous ASP-OPL-based prediction both in specificity (91% vs. 48%) and sensitivity (81% vs. 70%). The accuracy of the PKC-ζ predictions was tested by using the same approach and the same panel of proteomic peptides; for this isoform also, the PS-OPL-based predictions (Fig. 3*c*) surpassed those of the previous method (Fig. 3*d*). Predictions based on the PS-OPL method are also more accurate than the *ab initio* predictions of PREDIKIN (17) (for these peptides, 46% specificity for PKC-δ and 0% specificity for PKC-ζ). Thus, PS-OPL-based predictions perform well when subjected to rigorous experimental validation.

Accuracy of PSSMs in Distinguishing Between PKC Isoforms. PKC phosphorylation *in vivo* is sometimes mediated by one isoform of PKC much more than others. It would be useful to know whether differences in peptide specificity between isoforms contribute to such discrimination of substrates and, if so, what residues in that substrate are particularly important for that specificity. To investigate similarities/differences in peptide specificity between PKC-ζ and -δ, we compared their phosphorylation of the proteomic peptides; the results (Fig. 4*a*) indicate that although there is a general correlation between the phosphorylation by ζ and δ, there are also differences. If that difference in specificity is correctly reflected in the PSSMs, then the PSSM for PKC-ζ should not correctly predict phosphorylation by PKC-δ (and vice versa). Results confirm that the PSSM for PKC-ζ is uniquely relevant to PKC-ζ because it predicts phosphorylation by PKC-ζ (Fig. 3*c*) but not phosphorylation by PKC-δ (Fig. 4*b*). So, although the PSSMs for PKC-δ and -ζ are closely related, the differences between the two PSSMs are informative.

Which residues in the peptides impose the specificity difference between PKC-ζ and -δ (Fig. 4*a*)? Interpretation would be simplest in peptides (*i*) that have a single phosphorylation site and (*ii*) whose scoring by the two PSSMs indicates differences predominantly at a single residue. These conditions were met for a peptide from cytohesin-2 (18), whose phosphorylation is sensitive to the difference between ζ and δ (arrow, Fig. 4 *a* and *b*; see also Fig. 8). Scoring of that cytohesin-2 peptide by the two PSSMs (Fig. 4*c*) reveals predominantly a large difference at the P−1 position: for ζ the I is disfavored (see L at P−1 position in

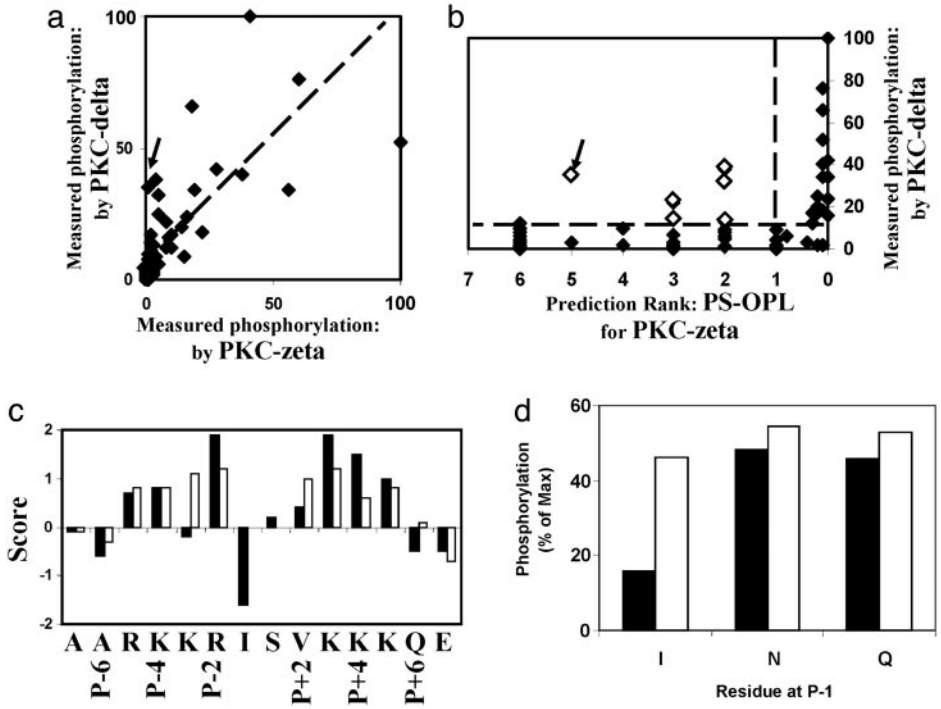

Fig. 4. Differences between PKC-δ and PKC-ζ. (*a*) PKC-ζ differs from PKC-δ in phosphorylation of proteomic peptides. The 75 proteomic peptides (Figs. 2 and 8) were subjected to *in vitro* phosphorylation at 10 μM by PKC-ζ and PKC-δ; data are expressed as relative phosphorylation, normalized to 100% for the best peptide. The dashed diagonal line represents a possible outcome if there were equal phosphorylation by both kinases. Arrow, cytohesin-2 outlier peptide. (*b*) The PKC-ζ PSSM fails to correctly predict phosphorylation by PKC-δ. This result contrasts with its good prediction of PKC-ζ phosphorylation (Fig. 3*c*). Open diamonds, false negatives; arrow, cytohesin-2 outlier peptide. (*c*) The P−1 position is the location of a major difference between PKC-ζ and PKC-δ in PSSM-based scoring of the cytohesin-2 peptide (see Fig. 8, peptide P11); arrows mark this peptide in *a* and *b*. The graph indicates scoring of each residue in its sequence by PSSMs for PKC-ζ (filled bars) or PKC-δ (open bars). (*d*) Replacement of the I residue at P−1 in the cytohesin-2 peptide with amino acids N or Q abolishes the differential phosphorylation by PKC-ζ (filled bars) vs. PKC-δ (open bars).

Fig. 2*b*), whereas for δ I is relatively neutral (see L at P−1 position in Fig. 2*a*). To test the prediction that I at P−1 is an important contributor to differential phosphorylation by ζ vs. δ, peptides were synthesized in which I was replaced with residues that are neutral for both ζ and δ (N and Q). The results (Fig. 4*d*) confirm the prediction; peptides in which that I is replaced (by N or Q) are now similarly phosphorylated by PKC-ζ and -δ. Thus, these PSSMs can accurately localize residues that contribute to specificity differences between PKC isoforms, and a disfavored residue is important in conferring such differences.

The foregoing observations regarding similarities/differences between kinases in peptide specificity are best understood in a broader context. For this purpose, we developed a "PSSM scatterplot" that gives a compact visual representation of the comparison between two different PSSMs. The scatterplot comparison of PKC-δ to PKC-θ (Fig. 5*a*) shows a high correlation between the 169 PSSM values for those two kinases, which must reflect virtual identity of specificity of those most closely related isoforms. The comparison of PKC-δ to PKC-ζ (Fig. 5*b*) demonstrates a correlation that is still strong but clearly lower than with PKC-θ. The distance of any point from the diagonal line indicates the amount of difference between the two kinases for that residue-position. Of the 10 residue-positions farthest from the line in the δ–ζ comparison, 8 instances are for hydrophobic residues and only 3 are for charged residues. For purposes of comparison, a PSSM logo for PKA also is shown (Fig. 5*c*); its comparison with PKC-δ (Fig. 5*d*) shows a much lower correlation. In this comparison between kinase subfamilies, many of the biggest differences come from charged residues, and only four are hydrophobic. The hierarchy of relatedness makes sense from an evolutionary viewpoint; θ is closest in sequence to δ (87% amino acid similarity in the kinase domain), ζ is the most distant amongst PKCs (65% similarity), and PKA is even further removed (58% similarity).

Relevance of Peptide Specificity to Phosphorylation Sites in Intact Proteins. Are such findings regarding peptide specificity relevant to an understanding of phosphorylation of intact proteins? If they are relevant, then many of the sites at which PKC has been shown to phosphorylate intact proteins should have good prediction ranks; if they are not relevant, then prediction ranks of those sites may be essentially random. We examined this issue on myristoylated alanine-rich C kinase substrate (MARCKS), one of the best known PKC substrates (19). When prediction ranks are calculated for the 39 Ser and Thr residues in MARCKS (Fig. 6*a*), the best ranks correspond precisely to the experimentally validated PKC phosphorylation sites on MARCKS (S159, S163, and S170) (see also tabulation in Fig. 9*a*, which is published as supporting information on the PNAS web site). We expanded the approach to a systematic analysis of experimentally determined PKC phosphorylation sites; ranks were determined for 124 experimentally determined PKC sites in the literature by using the conservative strategy of scoring only the sites in the proteome of a single species (i.e., humans) (Fig. 9*b*). The results (Fig. 6*b*) provide a powerful confirmation that peptide specificity is critical in determining many of PKC's physiological phosphorylation sites. There is a 38-fold enrichment of reported PKC phosphorylation sites in the top 1% of sites in the proteome predicted by the method. Moreover, the enrichment in the top 1% by our PS-OPL-based prediction is substantially more than the enrichment observed by using the previously described ASP-OPL-based method (Fig. 6*c*). Only 7% of these sites score

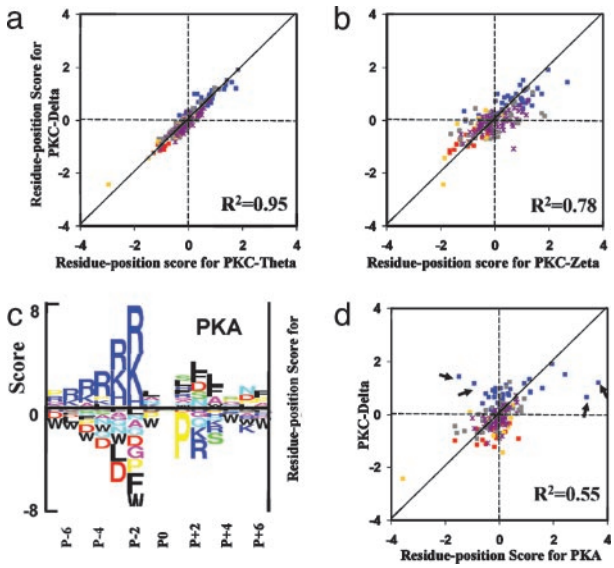

Fig. 5. Systematic analysis of differences between PKC-δ and other PKC isoforms or PKA. PSSM scatterplots compare PKC-δ to the following: PKC-θ (*a*), PKC-ζ (*b*), and PKA (*d*). Each dot represents a single residue at a single position; its *y* value is the score for PKC-δ, and its *x* value is a score for the other kinase. In this comparison, 171 residue-positions are compared (13 residues at 13 positions and 2 residues at P0). Four black arrows in *d* highlight the four positive residue-positions with the biggest differences between PKC-δ and PKA. (*c*) PSSM logo for PKA, derived by analysis with PKA using the same peptides used for PKC-ζ and PKC-δ.

positive by the PREDIKIN predictions for PKC-δ (Fig. 9*b*). These findings indicate greater accuracy of our method in determining PKC peptide specificity.

A further test of usefulness is whether the PKC prediction gives a more favorable distribution of scores to PKC sites than to reported PKA sites, the other best-characterized basophilic kinase. Our analysis (Fig. 6*d*) shows the following. (*i*) PKA sites are similar to PKC sites, reflected in overrepresentation of reported PKA sites in the top five percentiles of the PKC prediction. This finding is reasonable, based on their partially shared basophilic preferences; indeed, many of the PKA sites that fall in the top percentile of PKC predictions have been experimentally determined to be PKA substrates also (data not shown). (*ii*) However, PKC and PKA sites also are distinguished substantially by the PKC prediction. Specifically, the modal rank for PKA sites is the second percentile, whereas the modal rank for PKC sites is the first percentile.

Discussion

The power of the present approach in determining peptide specificity lies in its distinctive combination of four complementary strategies: First, positional scanning (20–24) provides systematic analysis of the preference for diverse residues at every position in the substrate. Second, the template used for positional scanning is an oriented peptide library (OPL, alternatively referred to as an oriented degenerate peptide) (12, 20, 25), consisting of two anchor residues and many degenerate residues. The two anchor residues (typically R at P−3 and F at P+1) provide sufficient favorable interaction of the investigated kinases to all peptide substrates to permit measurable phosphorylation of each peptide pool; the remaining degenerate positions provide diversity, which ensures that the residue preference measured pertains to diverse sequences (rather than to a single peptide, as is the case for most positional scanning studies of peptide specificity). Third, the systematic quantitative raw data have been transformed into an informative PSSM (Fig. 1*b*), which becomes the basis for an ensemble of quantitative approaches highlighted in the current studies. Fourth, the method of Yaffe *et al.* (13) has been used to predict peptide substrates for a kinase from the PSSM derived for that kinase.

Prediction and validation are the foundation of the scientific method. Consequently, we have systematically compared predictions with experimental results for peptide phosphorylation. The results provide powerful validation that PSSMs can accurately describe the kinase specificity of PKC-δ and -ζ (Fig. 3) and that PSSMs derived by our PS-OPL method are more accurate than the corresponding PSSMs derived by other available methods (12, 17). Because the difference between our predictions and the ASP-OPL-based predictions is the experimentally determined data in the corresponding PSSM, the superior predictions for PKC-δ and -ζ must reflect greater accuracy of the PS-OPL-based PSSMs. Potential limitations unique to the ASP-OPL method include difficulty in enriching phosphopeptides while excluding other negatively charged peptides and limited sensitivity in estimating abundance of disfavored residues (see supplementary methods to ref. 13). Having validated a PSSM, it serves as a powerful description of that kinase, which at best can

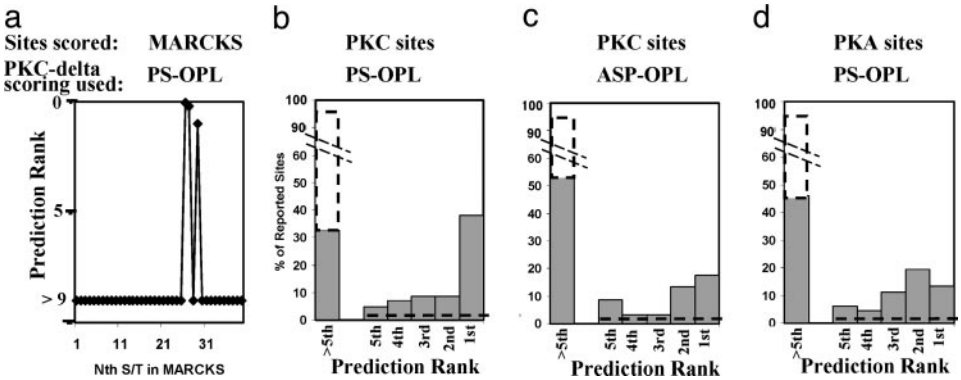

Fig. 6. Peptide specificity is relevant to phosphorylation of intact proteins. (*a*) Prediction rank of all 39 S/T sites in myristoylated alanine-rich C kinase substrate (MARCKS) for PKC-δ (see Figs. 7 and 9*a*). (*b–d*) Assessment of PKC predictions by analysis of scores they assign to reported phosphorylation sites for PKC or PKA. Each panel is a frequency distribution of scores assigned by either the PS-OPL described herein (*b* and *d*) or the ASP-OPL method (*c*) (11) (on which the Scansite prediction is based). Data are shown as histograms of site-abundance vs. prediction rank. Sites scored are either 124 reported mammalian PKC sites (*b* and *c*) or 135 reported mammalian PKA sites (*d*) (32). The PKC sites have been compiled from the peer-reviewed literature (see Fig. 9*b*); sites attributed to atypical PKCs were excluded from the analysis. Predictions shown with dashed lines are those of the null hypothesis that reported that phosphorylation PKC sites are independent of peptide specificity as ascertained by the current methods.

become the basis of whole range of quantitative analyses of that kinase and its relationship to other kinases (e.g., Figs. 4 and 5).

One key concept emerging from such quantitative studies is the importance of disfavored residues in kinase specificity, which previously has not been the subject of much discussion. Our PSSM logo graphic provides balanced visual emphasis to disfavored as well as favored residues (Fig. 2). Three distinct classes of disfavored residues are evident in our studies of PKC. First, negatively charged residues are disfavored throughout the substrate (Fig. 2). This disfavor is a logical consequence of the negatively charged catalytic cleft of PKCs, which would tend to repel substrates with negatively charged residues (11, 26–28). Second, a hydrophobic residue at the P−1 position is disfavored by PKC-ζ and confers substantial selectivity compared with PKC-δ (Fig. 4). Third, an exceptional disfavor for proline at the P+1 position is important for discrimination between kinases such as PKCs and proline-directed kinases (G.Z., K.F., Y.L., M. James, J.H., and S.S., unpublished data).

In contrast to recent emphasis on the importance of recruitment in substrate selection, the present study affirms and extends understanding of the contribution of peptide specificity in substrate phosphorylation. Specifically, there is an ≈40-fold enrichment of reported PKC phosphorylation sites in the first percentile of sites predicted by our PKC-δ PSSM (Fig. 6). This outcome would be predicted only if the following two assertions are true: (*i*) peptide specificity contributes prominently to overall PKC specificity (i.e., is not swamped out by colocalization); and (*ii*) our prediction accurately assesses PKC specificity. Multiple factors probably contribute to the distribution of sites below the 1% prediction rank (Fig. 6*b*), as follows: (*i*) sites in which efficient colocalization of kinase and substrate facilitates phosphorylation despite relatively unfavorable peptide sequence; (*ii*) sites in "folded" domains, which are not recognized as linear peptides; (*iii*) sites recognized as linear peptides in the intact protein but whose phosphorylation is underestimated by the predictive algorithm due to, for example, favorable peptide conformation in the intact protein; and (*iv*) sites erroneously ascribed to PKC. Further study will be required to define the relative contributions of each of these factors.

Predictive methods such as ours (available at http://mpr.nci.nih.gov) facilitate identification of phosphorylation sites in the following three contexts. (*i*) For a given protein, which of its residues are most likely to be phosphorylated by PKC (Fig. 6*a*)? This context is especially useful for choosing residues for site-directed mutagenesis of a known PKC substrate protein or to assess a previously unstudied protein and infer how likely it is to be a PKC substrate. We view sites scoring in the top percentile as especially good PKC candidates and sites below the fifth percentile as unlikely substrates. (*ii*) For a given phosphorylation site, is it likely to be a site for phosphorylation by PKC? This task is increasingly common given the efficiency of MS in site identification. (*iii*) For a given proteome, what sites are particularly likely to be PKC substrates? Because >1 million S/T are encoded in humans, even a very stringent cutoff (top 0.2 percentile) results in ≈2,000 predicted sites for δ and ≈2,000 for ζ. We infer that only some of those sites are physiological PKC phosphorylation sites; for example, sites in "folded domains" often will not be available in an extended conformation suitable for phosphorylation. Moreover, only a subset of those sites will be phosphorylated in a given cell type in response to any particular signal. Therefore, such predictions are most powerful when used in conjunction with other knowledge. For example, Ser-679 of erbb4 is a particularly appealing predicted site (top 0.2% for δ and ζ) because PKC influences erbb4 (29), and the related protein epidermal growth factor receptor has a PKC site in a similar location close to the membrane (30). Similarly, Ser-174 (and Ser-132) of ρ-GDI is an excellent predicted PKC site on a protein reported to be phosphorylated by PKC (31)

The PS-OPL method also will be useful in characterizing the specificity of the many kinases whose peptide specificity is poorly understood. For example, a variant approach is a one-anchor set of degenerate peptides in which the only fixed residue is an R at peptide position P−3 (and an S or T at P0). Such a set can be used with many kinases because of the common preference for R at P−3. Use of such a set allows determination of additional strongly favored positions for a kinase that become the basis for design of a two-anchor degenerate peptide set for that kinase. Because peptide specificity plays a key role (together with recruitment) in determining phosphorylation specificity *in vivo* (Fig. 6), approaches analogous to the one described will be important in building a comprehensive understanding of the "phosphorylation proteome."

We thank Philip Cole, Oskar Fernandez-Capetillo, Larry Samelson, Jean-Baptiste Vincourt, and Ron Wange for discussion and critique of the manuscript, and Holly Clark, Michael James, Lubing Lian, and John Rodwell for facilitating these studies.

1. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. (2002) *Science* **298,** 1912–1934.
2. Newton, A. C. (2001) *Chem. Rev.* **101,** 2353–2364.
3. Parker, P. J. & Parkinson, S. J. (2001) *Biochem. Soc. Trans.* **29,** 860–863.
4. Oancea, E. & Meyer, T. (1998) *Cell* **95,** 307–318.
5. Faux, M. C. & Scott, J. D. (1996) *Trends Biochem. Sci.* **21,** 312–315.
6. Kreegipuu, A., Blom, N., Brunak, S. & Jarv, J. (1998) *FEBS Lett.* **430,** 45–50.
7. Schaap, D., Parker, P. J., Bristol, A., Kriz, R. & Knopf, J. (1989) *FEBS Lett.* **243,** 351–357.
8. House, C., Wettenhall, R. E. & Kemp, B. E. (1987) *J. Biol. Chem.* **262,** 772–777.
9. Pearson, R. B. & Kemp, B. E. (1991) *Methods Enzymol.* **200,** 62–81.
10. Marais, R. M., Nguyen, O., Woodgett, J. R. & Parker, P. J. (1990) *FEBS Lett.* **277,** 151–155.
11. Nishikawa, K., Toker, A., Johannes, F. J., Songyang, Z. & Cantley, L. C. (1997) *J. Biol. Chem.* **272,** 952–960.
12. Songyang, Z., Blechner, S., Hoagland, N., Hoekstra, M. F., Piwnica-Worms, H. & Cantley, L. C. (1994) *Curr. Biol.* **4,** 973–982.
13. Yaffe, M. B., Leparc, G. G., Lai, J., Obata, T., Volinia, S. & Cantley, L. C. (2001) *Nat. Biotechnol.* **19,** 348–353.
14. Obenauer, J. C., Cantley, L. C. & Yaffe, M. B. (2003) *Nucleic Acids Res.* **31,** 3635–3641.
15. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 4355–4358.
16. Schneider, T. D. & Stephens, R. M. (1990) *Nucleic Acids Res.* **18,** 6097–6100.
17. Brinkworth, R. I., Breinl, R. A. & Kobe, B. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 74–79.
18. Santy, L. C., Frank, S. R., Hatfield, J. C. & Casanova, J. E. (1999) *Curr. Biol.* **9,** 1173–1176.
19. Stumpo, D. J., Graff, J. M., Albert, K. A., Greengard, P. & Blackshear, P. J. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 4012–4016.
20. Turk, B. E. & Cantley, L. C. (2003) *Curr. Opin. Chem. Biol.* **7,** 84–90.
21. Pinilla, C., Appel, J. R., Blanc, P. & Houghten, R. A. (1992) *BioTechniques* **13,** 901–905.
22. Uttamchandani, M., Chan, E. W., Chen, G. Y. & Yao, S. Q. (2003) *Bioorg. Med. Chem. Lett.* **13,** 2997–3000.
23. Himpel, S., Tegge, W., Frank, R., Leder, S., Joost, H. G. & Becker, W. (2000) *J. Biol. Chem.* **275,** 2431–2438.
24. Velentza, A. V., Schumacher, A. M., Weiss, C., Egli, M. & Watterson, D. M. (2001) *J. Biol. Chem.* **276,** 38956–38965.
25. Rodriguez, M., Li, S. S., Harper, J. W. & Songyang, Z. (2004) *J. Biol. Chem.* **279,** 8802–8807.
26. Wade, R. C., Gabdoulline, R. R., Ludemann, S. K. & Lounnas, V. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5942–5949.
27. Livesay, D. R., Jambeck, P., Rojnuckarin, A. & Subramaniam, S. (2003) *Biochemistry* **42,** 3464–3473.
28. Grant, B. D., Tsigelny, I., Adams, J. A. & Taylor, S. S. (1996) *Protein Sci.* **5,** 1316–1324.
29. Vecchi, M., Baulida, J. & Carpenter, G. (1996) *J. Biol. Chem.* **271,** 18989–18995.
30. Hunter, T., Ling, N. & Cooper, J. A. (1984) *Nature* **311,** 480–483.
31. Mehta, D., Rahman, A. & Malik, A. B. (2001) *J. Biol. Chem.* **276,** 22614–22620.
32. Shabb, J. B. (2001) *Chem. Rev.* **101,** 2381–2411.