Exceptional Disfavor for Proline at the P+1 Position among AGC and CAMK Kinases Establishes Reciprocal Specificity between Them and the Proline-directed Kinases*

Guozhi Zhu‡§, Koichi Fujii‡§, Natalya Belkina‡, Yin Liu‡, Michael James‡, Juan Herrero‡, and Stephen Shaw‡§**
From the ‡Experimental Immunology Branch, NCI, National Institutes of Health, Bethesda, Maryland 20892 and §AxCell Biosciences, Newtown, Pennsylvania 18940

To precisely regulate critical signaling pathways, two kinases that phosphorylate distinct sites on the same protein substrate must have mutually exclusive specificity. Evolution could assure this by designing families of kinase such as basophilic kinases and proline-directed kinase with distinct peptide specificity; their reciprocal peptide specificity would have to be very complete, since recruitment of substrate allows phosphorylation of even rather poor phosphorylation sites in a protein. Here we report a powerful evolutionary strategy that assures distinct substrates for basophilic kinases (PKA, PKG and PKC (AGC) and calmodulin-dependent protein kinase (CAMK)) and proline-directed kinase, namely by the presence or absence of proline at the P+1 position in substrates. Analysis of degenerate and non-degenerate peptides by in vitro kinase assays reveals that proline at the P+1 position in substrates functions as a “veto” residue in substrate recognition by AGC and CAMK kinases. Furthermore, analysis of reported substrates of two typical basophilic kinases, protein kinase C and protein kinase A, shows the lowest occurrence of proline at the P+1 position. Analysis of crystal structures and sequence conservation provides a molecular basis for this disfavor and illustrate its generality.

Phosphorylation is a prevalent modification in cells that controls many functions such as signaling transduction, proliferation and apoptosis. It is estimated that at least one-third of all proteins in eukaryotic cell are phosphorylated at any given time (1, 2). More than 500 human protein kinases have been identified so far (3). A high degree of selectivity in substrate phosphorylation is necessary to maintain functional integrity of this very complicated signaling environment. Precision in phosphorylation is particularly critical when a substrate protein is phosphorylated at two (or more) phosphorylation sites, and those sites 1) are phosphorylated by distinct upstream kinases and 2) confer distinct properties on that substrate. To assure fidelity of signaling in this common situation, each upstream kinase must show high specificity by phosphorylating only the relevant phosphorylation site and not the inappropriate site(s). This requirement poses a major challenge in evolutionary design of kinase peptide specificity, since those upstream kinases are usually recruited to the substrate, and such recruitment can overcome much of the barrier provided by peptide specificity (4).

These considerations raise the important issue as to what elements in kinase peptide specificity confer the strongest reciprocal specificity between kinases, i.e. which prevent one kinase from phosphorylating substrates phosphorylated by another. For Ser/Thr kinases, we propose that much of this reciprocal specificity is provided by the evolution of three broad classes (5): basophilic kinases that phosphorylate sites with clustered positive charges, acidophilic kinases that phosphorylate sites with clustered negative charges, and proline-directed kinases that phosphorylate sites in which Ser/Thr is followed immediately by a proline (i.e. proline at the P+1 position). This classification by peptide specificity corresponds generally with classification of Ser/Thr kinases based on sequence similarity. Based on studies of a sampling of family members, two large families of kinases appear to be largely included in the basophilic group: the 61 gene AGC (PKA, PKG and PKC) family (6–8) and the 83 gene CAMK (calmodulin-dependent protein kinase) family (5, 9, 10). Proline-directed kinases such as GSK, CDK, and Erk belong to CMGC (CDK, MAPK, GSK3, CLK kinases) superfamily of kinases, which consists of 61 kinases (6, 11–13). The acidophilic kinase group is much smaller, including, for example, casein kinase I and casein kinase II. The basophilic kinases and acidophilic kinases would be expected to be largely non-overlapping in peptide specificity; the electrostatic interactions that promote phosphorylations of basic sites by basophilic kinases will create an aversion for acidic sites (and vice versa). Moreover, proline-directed kinases will eschew substrates of the other two which lack proline at P+1.

However, a missing element in the foregoing paradigm is one that prevents basophilic kinases or acidophilic kinases from phosphorylating sites preferred by proline-directed kinases. Such an element may be greater importance for basophilic kinases, since their preference for basic residues is shared in a limited fashion by some proline-directed kinases (e.g. Cdk1, Cdk2, Cdk5, Dyrk, GSK3) (5, 6, 13, 14). We have identified an evolutionary strategy widely used by basophilic kinases that addresses that problem. We find a disfavor for proline at the P+1 position among AGC and CAMK kinases. That disfavor is

* 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.

‡ These authors contributed equally to this work.
¶ Present address: First Dept. of Internal Medicine, University of Occupational and Environmental Health, Kitakyushu, Fukuoka 807-8555, Japan.
** To whom correspondence should be addressed: Experimental Immunology Branch, NCI, National Institutes of Health, Bethesda, MD 20892. Tel.: 301-496-0887; E-mail: sshaw@nih.gov.

1 The abbreviations used are: AGC, PKA, PKG and PKC; CAMK, calmodulin-dependent protein kinase; PK, protein kinase; PDB, protein data base.

Printed in U.S.A.
Materials and Methods

Degenerate Peptides and Kinases—Biotinylated single sequence peptides and degenerate peptides were synthesized as C-terminal amides on Mimotopes (Clayton, Australia) SynPhase Rink amide acryl-grafted polypropylene solid support (loading 7.5 μmol) using conventional Fmoc (N-[9-Fluorenylethoxycarbonyl) chemistry as described previously (15). Seven catalytically active kinases were used. PKC-δ, PKC-ζ, PKA, and PKG were purchased from Calbiochem, and PKD2 and Cdk2 were purchased from Upstate Biotechnology. A 6 × Histagged construct corresponding to residues 615–874 of human PRK1 (GenBanktm accession number BC040061) was expressed in 293T cell by calcium phosphate transfection and the corresponding protein purified by nickel affinity chromatography.

In Vitro Kinase Assay—Peptides were phosphorylated by in vitro kinase assay as described previously (15). In brief in vitro phosphorylation in the presence of γ-32P-labeled ATP (in 100 μM cold ATP) was performed in 50 μl of solution under standardized conditions resulting in stoichiometry of phosphorylation less than 10%. After reaction termination, 50 pmol of substrate was transferred to streptavidin-coated plates and emissions counted after extensive washing. Kinase buffer for PKC-δ, PKC-ζ, PKD2, PKG, and PKA, and PKR1 is 100 mM HEPES, 0.05% Triton X-100, 1 mM CaCl2, 20 mM MgCl2. For kinase PKC-δ, PKC-ζ, and PKD2, 0.2 mg/ml phosphatidylserine (Avanti Polar Lipids) and 100 ng/ml phorbol 12-myristate 13-acetate were added to the kinase buffer immediately before the assay. Kinase buffer for CAMK II is 25 mM HEPES, 10 mM MgCl2, 1 mM CaCl2, 8 mg/ml calmodulin, and 1 mM dithiothreitol, which was minimally modified from Madhavan (34).

Data Analysis—Data were analyzed as described previously (15). All results shown represent at least duplicate assays, and usually quadruplicate assays. Data analysis was largely done in Microsoft Excel spreadsheets. Functionality of embedded spreadsheet formulas was augmented by automation using Microsoft Visual Basic and data storage using Microsoft SQL Server. The PSSM (position-specific scoring matrix) Logo was generated using postscript files generated by Visual Basic code; some postscript code was adapted from Tom Schneider’s “makelogo” 8.69 (www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi).

Results

During studies of PKC isoform specificity (15) a strong disfavor for proline at the P+1 position was noted. Given the potential biological importance of P+1 proline disfavor, analysis was expanded to other AGC family members such as PKG (Fig. 1A). Markedly different levels of phosphorylation by PKG are observed between peptides that differ only by a single amino acid at their P+1 position. Since each peptide is a pool of about 1 billion individual peptides, the preference (or disfavor) observed must reflect a preference (or disfavor) of that kinase for many peptides in that huge pool. When Pro is present at the P+1 position, the amount of phosphorylation is much lower than with any other residue. Specifically, phosphorylation of peptide #13 (Pro at P+1) is about 1/6 of the average for all peptide pools. Thus, Pro must be unfavorable at the P+1 position for most of the phosphorylatable peptides in this degenerate peptide pool #13.

We investigated whether the strong disfavor for Pro at P+1 was a characteristic shared with other kinases. Four additional members of the AGC kinase family (PKA, PKC-δ, PKC-ζ, and
PRK1) were tested and each showed strong disfavor for Pro at P+1. One way to visualize this comparison of kinases is with a PSSM Logo showing preferences of those kinases at the P+1 position (Fig. 1B). Consider first the results for PKG (Fig. 1B, left column, which corresponds to data shown in Fig. 1A). The parameter represented in the PSSM Logo is log2 of the ratio to mean for each residue; this mathematical conversion gives favorable residue positive scores and disfavored residues negative scores. The PSSM Logo shows the residue most favored by PKG at the top (1) and the most disfavored residue (P) at the bottom. The columns corresponding to other kinases show that the strong disfavor for proline at the P+1 position is observed for each of the four other AGC kinases (PKA, PKC-δ, PKC-ζ, and PRK1). To broaden the analysis, we also examined two members of the CAMK family (CAMK II and PKD2); since they share the preference for Arg at P−3 (and P−2) (5, 6) they can be analyzed with the same sets of degenerate peptides. The results show that CAMK II and PKD2 also have strong disfavor for Pro at P+1 (Fig. 1B).

One useful way to further characterize the proline disfavor at P+1 is to compare it with the preferences for residues in other positions near the phosphorylation site. Such a comparison is shown for the five basophilic kinases studied in further detail (Fig. 2). Consider first the results for PKG (Fig. 2, left panel). The distribution of scores is approximately centered on zero, and most residue preferences are close to zero. The outlier far to the left corresponds to Pro at P+1; its separation from the rest of the distribution demonstrates that this is a singular disfavor, i.e. distinctly stronger than the disfavor for any residue at any other position studied. At the opposite extreme, there are several outliers in the favored direction; these correspond to Arg and Lys at P−2 and P−3, as would be expected based on the known very strong PKG preference for those residues in those positions (5). Thus, disfavor for Pro at P+1 position is by far the single strongest disfavored residue for PKG. The frequency distribution of scores for these other kinases revealed that for each of them Pro at P+1 is the most disfavored residue among all residues at all positions tested. Scores for Pro at position P+1 range from −2.2 to −3.1; these scores are both more extreme than any other disfavored residue we have studied and more consistent between AGC/CAMK kinases than any other disfavored residue.

The foregoing studies of in vitro peptide phosphorylation indicate that Pro at the P+1 position is a disfavored residue for many kinases of the AGC and CAMK family. If this characteristic is relevant not only to peptides but also to proteins, then the frequency of Pro at the P+1 position in reported substrates should be much less than the frequency of Pro in human proteins generally (6.5%). Such a comparison can be convincing only when a large number of substrates have been reported, as is the case for PKA and PKC. We have accumulated a list of 124 PKC substrates from review of the literature (15); the list was not biased by our hypothesis of Pro disfavor, since it was accumulated before that hypothesis was formulated. Analysis of the number and position of Pro residues in these PKC substrates strongly supports the prediction. The overall frequency of Pro residues in these substrates (from position P−8 to P+6) was 4.8% and thus not very different from the proteome at large. Notably the frequency of Pro at P+1 was very low; only 1 of 124 reported substrates had Pro at P+1, which represents a frequency of only 0.8%. It is useful to convert results of residue frequency analysis in substrates into a log score by determining the log base 2 of (observed frequency/expected frequency); the log2 score for Pro at P+1 from reported PKC substrate analysis is −3.1, which is comparable with our score of −2.6 derived from degenerate peptide analysis. Analysis was expanded to determine proline favor/disfavor at residue position P−7 to P+6 (Fig. 3A). There is generally good agreement between the scores derived from residue frequency in reported substrates and scores derived from degenerate peptides. We performed a similar analysis on PKA sites compiled by Shabb (8); of those 136 were mammalian sites plausibly assigned to PKA based on the references cited. The frequency of Pro at P+1 in these substrates was 1.3%, which is again much lower than expected based on overall frequency in those substrates (5.5%) or human proteins generally (6.5%). As was seen with PKC, the position-specific scores for Pro frequency in substrates are in general agreement with scores from degenerate peptides (Fig. 3B). Thus, there is a very low Pro frequency at the P+1 position in reported PKC and PKA substrates and the P+1 position is the only one at which Pro disfavor occurs in both PKC and PKA.

The potential importance of this disfavor is magnified when understood in the context of the virtual requirement for Pro at P+1 among proline-directed kinases (5, 11, 12), consisting primarily of the CMGC family of kinases, such as Cdk2. As noted in the introduction, at least five proline-directed kinases have a modest basophilic preference, which creates the potential risk

**FIG. 2. Singular disfavor for Pro at P+1 is shared by three AGC kinases and two CAMK kinases.** Frequency distributions of scores from position P−7 to P+6 for PKA, PKC-δ, and from P−4 to P+3 for PKG, CAMK II, and PKD2. Log scores for 13 residues in 13 or 7 positions (169 or 91 residue-position combinations) were determined as described above and previously (15) and shown as a frequency histogram. Each bar represents the number of residue-positions having a score in a particular range. The principle is illustrated on the left panel by: 1) indicating where the residues for PKG P+1 position from Fig. 1 are located in the frequency distribution and 2) showing as open bars the residues most strongly favored by PKG (Arg (R) at P−2, P−3, P−4, and Lys (K) at P−2). The arrow indicates the bar representing Pro at P+1. The vertical dashed lines at scores −2 and +2 emphasize outliers.
of poor discrimination between basophilic AGC/CAMK kinases and proline-directed kinases. We therefore tested whether the disfavor for Pro at P\text{+1} is an important element in discrimination of AGC/CAMK sites from sites recognized by proline-directed kinases. As a proof of principle we chose "proteomic" peptides corresponding to basic sites in the proteome that our algorithms suggested could be good substrates for basophilic kinases in all respects except for their Pro at P\text{+1}. We then synthesized peptides with wildtype sequence (Pro at P\text{+1}) or with substitutions to two alternative residues at the P\text{+1} position. One proteomic peptide was particularly informative for comparison of PKC-\(\delta\) versus Cdk2 (Fig. 4A). PKC-\(\delta\) was efficient in phosphorylating the peptide whose P\text{+1} residue was Phe (PKC-\(\delta\) FSSM score 1.0, also see Fig. 1B); less efficient with an His at P\text{+1} (FSSM score 0.1, also see Fig. 1B) and almost inactive with a Pro at P\text{+1} (FSSM score \(-3.1\), also see Fig. 1B). In contrast Cdk2 efficiently phosphorylated only the peptide with Pro at P\text{+1} and neither of the others. Thus, the specificity differences between Cdk2 and PKC-\(\delta\) are as much a result of the aversion of PKC-\(\delta\) for Pro as the preference of Cdk2 for Pro. Analysis of the second proteomic peptide (and P\text{+1} variations thereof) was particularly informative for comparison of CAMK II with Cdk2 (Fig. 4B). CAMK II can efficiently phosphorylate the peptide with Phe at P\text{+1} but is very inefficient in phosphorylating the peptide with Pro at the P\text{+1} position. Note that phosphorylation of the peptide with Pro at P\text{+1} (the most disfavored residue at P\text{+1} for CAMK II, Fig. 1B) is much less than the peptide having Gln at P\text{+1} (the second most disfavored residue, Fig. 1B); thus Pro has an exceptional capacity to veto phosphorylation by CAMK II. Based on the foregoing studies, we introduce the term "proline-aversive" kinase to describe those with a strong disfavor for Pro at P\text{+1} and thereby highlight the reciprocity of their specificity to that of proline-directed kinases.

The structural basis for the difference between the proline-directed kinases and proline-aversive kinase can be understood based on differences between the solved structures of proline-directed kinases with bound peptide (e.g. PDB code 1QMZ) and AGC/CAMK kinases with bound peptide (e.g. PDB code 1ATP). One of the regions of the kinase domain that is most important for substrate recognition region is the activation loop, a flexible region of the kinase that 1) lies along the catalytic cleft, 2) is involved in binding C-terminal residues of substrate, and 3) is subject to extensive conformational regulation (16–18). In the activation loop there is a critical residue which we refer to as the toggle residue whose conformation differs dramatically between proline-aversive and proline-directed kinases. In the two proline-aversive kinases with solved structure (AGK family kinase PKA and the CAMK member PKH) the toggle residue is a glycine, whose backbone carbonyl is oriented toward the catalytic cleft (Gly\textsuperscript{390} (G200, Fig. 5, A and B); that carbonyl facilitates binding of substrate by the formation of a H-bond with the backbone amide of the substrate P\text{+1} residue (19, 20). Substrates having Pro at P\text{+1} cannot form that H-bond because the proline amide is not an H-bond donor. In contrast, in proline-directed kinases (Erk2, Cdk2, GSK3) the carbonyl group of their toggle residue is oriented away from the catalytic cleft (21–25) (Val\textsuperscript{164} (V164), Fig. 5, C and D5). So binding of all residues except Pro at P\text{+1} position will be disfavored because of an uncompensated H-bond from the main-chain P\text{+1} amide of the substrate (23, 24).

The orientation of the toggle residue in proline-directed kinases is enforced by a strategically placed arginine (Arg\textsuperscript{169} (R169), Fig. 5D), which we refer to as a toggle-regulating residue. That arginine is located in the P\text{+1} loop, which is a loop adjacent to the activation loop (that is firmly attached to the stable C-lobe of the kinase domain). That arginine forms an H-bond with the carbonyl of the toggle residue, thereby maintaining the usual conformation of that toggle residue (23, 24) (Fig. 5D). In the foregoing model, the single residue most pivotal to creating the two distinct conformations of the toggle residue is this toggle-regulating arginine. Sequence analysis of human kinases reveals a striking pattern of conservation of the toggle-regulating residue (Table I). All 61 kinases of the CMGC family have arginine at that position. In contrast, none of the
324 kinases in the AGC, CAMK, STE, TK, and TKL families have arginine at that position; instead they have acyclic hydrophobic residues, which cannot provide an H-bond to the carbonyl of the toggle residue. Therefore, this amino acid position has the property of absolute conservation within the CMGC family and complete discordance from AGC/CAMK and other kinases. Recently Kannan and Neuwald (26) noted that this arginine residue, which they refer to as the “CMGC-arginine,” is the most characteristic feature of the CMGC kinase family.

Analysis of sequence conservation of the toggle-regulating residue and the toggle residue among human kinases

| Positions       | Residue | Kinase family |
|-----------------|---------|---------------|
|                 |         | CMGC | AGC | CAMK | STE | TK + TKL |
| Toggle-regulating | ILMV    | 0    | 95  | 59  | 98  | 65      |
|                 | Ala     | 0    | 2   | 30  | 0   | 3       |
|                 | Thr     | 0    | 0   | 1   | 0   | 3       |
|                 | Arg     | 100  | 0   | 0   | 0   | 0       |
| Toggle          | AG      | 21   | 95  | 87  | 94  | 26      |
|                 | Gln     | 20   | 0   | 2   | 0   | 0       |
|                 | FILMV   | 46   | 3   | 4   | 0   | 61      |
| No. of kinases  |         | 61   | 63  | 81  | 48  | 132     |
occurs at a much reduced frequency at the P1+1 position among reported substrates for PKA and PKC. A structural explanation for this disfavor for proline is evident in crystal structures of AGC/CAMK in which a carbonyl group is strategically oriented toward the catalytic cleft. The ability of all substrate P1+1 residues except proline to provide a backbone amide for H-bonding to that carbonyl gives such kinases a singular disfavor for proline.

The foregoing analysis focuses on substrate binding but does not explicitly consider substrate positioning. A cardinal feature identified by structure/function analysis of protein kinases is that proper orientation of the activation loop and peptide is essential for catalysis (16–18). For proline-directed kinases substrate positioning is provided by docking of proline into a proline-selective pocket (Fig. 5C). Remarkably those residues (Val183 and Glu162) form this proline-selective pocket only when the activation loop is phosphorylated at Thr160 and the kinase domain is bound to cyclin A; in the unphosphorylated inactive complex the conformations of these two residues are very different, and there is no such tight and specific docking site for proline (27). For proline-aversive kinases positioning is provided in a different manner; the H-bond between activation loop and the P1+1 backbone amide positions the substrate into an orientation suitable for phosphorylation (Fig. 5A). This H-bond also stabilizes the P1+1 residue in a preferred orientation in which its side chain is oriented into the hydrophobic P1+1 pocket of PKA; this orientation is observed with isoleucine at P1+1 in 1ATP and is consistent with the preference for hydrophobic residues at the P1+1 position for PKA (6). But in the special case of peptides with proline at P1+1, this positioning mechanism cannot operate for PKA (or other CAMK/AGC kinase). We hypothesize that proline at P1+1, since it lacks the orientation/ constraint of a H-bond, can find another binding site(s) in the wide hydrophobic pocket formed by Leu205, Leu198, or even Phe (Figs. 5A). Improper binding of the proline residue can be quite stable but fails to result in catalysis because of incorrect positioning of the substrate (and phosphorylatable residue) in the catalytic cleft or conformation changes resulting from ectopic binding of proline. Extensive in silico and experimental analysis will be required to properly test this hypothesis.

Proline is unique among natural amino acids in having a substituted amide and resulting conformational rigidity. Those properties make it uniquely suited for particular structural roles. For example, it is an important element in the peptide sequences recognized by proline-specific binding domain such as SH3, WW, and EVH1 domains (28–30). The role of proline in sequences recognized by proline-specific binding domain such as SH3, WW, and EVH1 domains (28–30). The role of proline in sequences recognized by proline-specific binding domain such as SH3, WW, and EVH1 domains (28–30). The role of proline in sequences recognized by proline-specific binding domain such as SH3, WW, and EVH1 domains (28–30). The role of proline in sequences recognized by proline-specific binding domain such as SH3, WW, and EVH1 domains (28–30). The role of proline in sequences recognized by proline-specific binding domain such as SH3, WW, and EVH1 domains (28–30). The role of proline in sequences recognized by proline-specific binding domain such as SH3, WW, and EVH1 domains (28–30). The role of proline in sequences recognized by proline-specific binding domain such as SH3, WW, and EVH1 domains (28–30). The role of proline in sequences recognized by proline-specific binding domain such as SH3, WW, and EVH1 domains (28–30).