

Motif Analysis of Amino-Acid Sequences by a Quantification Method: Application to Phosphorylation Signals of Protein Kinase C and cAMP-Dependent Protein Kinase

Yôichi Iida

Division of Chemistry, Graduate School of Science, Hokkaido University, Sapporo 060-0810

Received June 19, 2003; E-mail: Yoichi.Iida@mb8.seikyou.ne.jp

Prediction of protein function using amino-acid sequence motifs is based on the observation that the functionally important region is strongly conserved in short segments of amino-acid sequences. Although the consensus sequence has been used to describe such a functional signal, the actual sequence differs from it to a greater or lesser degree, and the consensus sequence only describes the signal qualitatively. In the present report, we study phosphorylation signals of protein kinase C (PKC) and cAMP-dependent protein kinase (PKA). PKC phosphorylates solely Ser or Thr residues. The consensus sequence is given by $(R/K_{1-3},X_{2-0})$-S/T-$(X_{2-0},R/K_{1-3})$, where X denotes no particular amino acid. PKA also phosphorylates Ser or Thr, but its consensus sequence is described by R-R/K-X-S/T. We analyzed such signals by a quantification method, and estimated the strength of the signal quantitatively. This approach was applied to several proteins and peptide analogues, and the replacement effect of amino acids upon catalytic activities of phosphorylation was explained in terms of the strength of the signal (sample score of peptide sequence).

Prediction of protein function using amino-acid sequence motifs is based on the observation that functionally important region is strongly conserved in short segments of amino-acid sequences. So far, the consensus sequence has been used to describe the functional signal.[1–3] The usefulness of the consensus sequence lies in its simplicity, and the complexities of the substrate recognition process are summarized as sets of short recognition sequences. However, the assumption that the local primary sequence alone may control recognition is oversimplified. Factors such as secondary/tertiary structures or distant secondary recognition sites may play significant roles in substrate recognition. The second problem with the consensus sequence is that the actual amino-acid sequence differs from the consensus sequence to a greater or lesser degree. Moreover, the consensus sequence gives only favorable amino-acid residues, but does not show any negative influence of unfavorable residues. It is, then, ambiguous to what degree matching between the actual sequence and the consensus sequence is necessary to specify the exact recognition site. The third problem is that mutational studies have demonstrated that the importance of amino-acid matching to the consensus sequence differs from position to position. The consensus sequence does not teach us the relative importance of each amino-acid residue at each position of the sequence. To improve this situation,[1] a frequency matrix of each amino acid at each position in compiled data of known sites has been used. This profiling gives more information than a consensus sequence, but does not tell us about the negative influence of unfavorable residues quantitatively. Several attempts, such as the perceptron algorithm and neural network, have been made to go beyond such a frequency matrix.

In the present paper, a quantification method was developed to analyze phosphorylation signals of protein kinase C (PKC) and cAMP-dependent protein kinase (PKA). PKC phosphorylates solely Ser or Thr residues, and prefers places where some basic residues, such as Arg and Lys (Arg appears to be superior to Lys), lie around the Ser/Thr phosphorylation site. When amino-acid residues are denoted by one-letter codes, the consensus sequence for the motif of PKC phosphorylation is given by $(R/K_{1-3},X_{2-0})$-S/T-$(X_{2-0},R/K_{1-3})$, where the underlined S or T is phosphorylated.[2,3] When two amino acids function interchangeably, both are listed with a stroke (/) separating them. X denotes no particular amino acid, but this does not always guarantee that all possible substitutions at that position will not affect the recognition. On the other hand, PKA prefers basic amino acids in the N-terminal region to the Ser/Thr phosphorylation site, but prefers no particular basic residues in the C-terminal region. Its consensus sequence is given by R-R/K-X-S/T.[2,3]

Earlier, a quantification method was applied to analyze functional signals of DNA, such as the splice signal of mammalian mRNA precursors.[4–6] In DNA sequences, there are only four kinds of nucleotides (A, G, C, and T) at each position, whereas there are as many as twenty kinds of amino-acid residues at each position in protein sequences. Although several difficulties arise in the motif analysis of protein, this quantification method was again applicable to the analysis of the phosphorylation signals of PKC and PKA. The method treats primary amino-acid sequences, but does not improve the previous problem for their secondary/tertiary structures. However, it improves the second and third problems of the consensus sequence. Our method tells us the strength of the phosphorylation signal, which quantitatively explains the replacement effect of amino-acid residues upon the catalytic efficiencies of phosphorylation by PKC and PKA.

Published on the web February 10, 2004; DOI 10.1246/bcsj.77.281

Method of Analysis

The principle of the quantification analysis was essentially the same as that described previously in the DNA sequence analysis.[4–6] Several problems arise in the analysis of protein sequences (they will be discussed later in Concluding Remarks). The first is that 20 kinds of amino acids occur at each position of the sequence. This gives 20 categories to each position, and increases considerably the number of item-categories (see below). Our computer program was revised, and we could analyze as long as 15-amino-acid sequences consisting of a total of 300 item-categories (15 positions × 20 categories). As an example of this analysis, we first show the results of the PKC phosphorylation signal of neuronal protein, neurogranin, as reported by Chen et al.[7]

In the quantification analysis, we have to construct two groups of amino-acid sequence data. The first group ($r = 1$) is composed of sequences which include the signal of phosphorylation by PKC. Kreegipuu et al.[2] summarized more than 180 substrate sequences for the phosphorylation signal. They also gave an amino-acid frequency table over the sequences between −12 and +12 (hereafter, the Ser or Thr phosphorylation site is denoted by position 0). Referring to their data, we constructed sequences of the first group ($r = 1$), and collected 207 samples of 9-amino-acid sequences between the −4 and +4 positions, as given in Table 1. Sequences of the second group ($r = 2$) do not include the signal of phosphorylation. They are taken from the amino-acid sequence of neurogranin. It is a selective substrate for PKC and is composed of 76 amino-acid residues, in which the phosphorylation site is the Ser at the 34-th position from the N-terminal.[2,7] To collect sample sequences of the second group, we first take the 9-amino-acid sequence at the N-terminal. Next, we progress one residue in the C-terminal direction and take the next 9-amino-acid sequence. In this way, we window 9-amino-acid sequences at every position of the whole neurogranin sequence. In those samples, however, one sequence lies at the phosphorylation site, and is brought into the first group ($r = 1$). The remaining 67 sequences are summarized in the second group ($r = 2$), as shown in Table 1. However, such a sampling is found to be insufficient to provide the second group. As will be discussed in the Concluding Remarks, the whole neurogranin sequence is not long enough. Amino-acid residues do not occur equally, and some residues, such as Trp, occur very rarely. To compensate for the length of the sequence and the inequality of amino-acid residues, 500 samples of 9-amino-acid sequences composed of randomly arranged residues were generated by a series of random numbers and were added to the neurogranin sequences in the second group. Some of those random sequences are also shown in Table 1.

Next, the sequence data of Table 1 were transformed into item-category data. For this purpose, we introduced a dummy variable, $x_{i(\alpha)}^{r(\nu)}$, which is defined by item ($i = 1, 2, \ldots, 9$), category ($\alpha = 1, 2, \ldots, 20$), group ($r = 1, 2$) and sample ($\nu = 1, 2, \ldots, n_r$). Nine items correspond to the positions of residues in the 9-amino-acid sequence, i being given by the order from N- to C-terminal of the sequence. Twenty categories denote the kinds of amino-acid residues at each position, where α/residue is given by 1/A, 2/C, 3/D, 4/E, 5/F, 6/G, 7/H, 8/I, 9/K, 10/L, 11/M, 12/N, 13/P, 14/Q, 15/R, 16/S, 17/T, 18/V, 19/W, and 20/Y, respectively. Parameter ν specifies each sample sequence belonging to the group ($r = 1$ or 2). $n_1 = 207$ and $n_2 = 567$ are the total number of sample sequences in each group. The dummy variable, $x_{i(\alpha)}^{r(\nu)}$, takes 1 if the sample sequence (ν) of the group (r) has an amino-acid residue (α) at the position (i), otherwise it takes 0. Using this variable, we transformed the sequence data of Table 1 into the item-category data composed of 0 or 1.

Quantification of each sequence can be done by calculating the sample score value,

$$y^{r(\nu)} = \sum_{i=1}^{9} \sum_{\alpha=1}^{20} x_{i(\alpha)}^{r(\nu)} a_{i(\alpha)}, \tag{1}$$

Table 1. Nine-Amino-Acid Sequence Data of PKC Phosphorylation Signal to be Analyzed by Quantification Method

Group (r)[a)]	No. (ν)	Sequence	Source[b)]
1	1	PKDPSQRRR	A001 S-11
1	2	DRLVSARSV	A009 S-985
:	:	:	
1	20	KIQASFRGH	B009 S-34 (Neurogranin)
:	:	:	
1	207	KRQGSVRRR	B360 S-11
2	1	DICDFWWKV	Random Sequence
2	2	PMLCMCHWQ	
:	:	:	
2	500	KVQRRVANS	
2	501	MCTESACSC	B009 (Neurogranin)
2	502	CTESACSCP	
:	:	:	
2	567	GAGGGPSGG	

a) Group (1) is composed of amino-acid sequences including PKC phosphorylation signal, while group (2) comprises sequences including no such signal. See text for further details. b) See Ref. 2 and text.

where $r = 1, 2$ and $\nu = 1, 2, \ldots, n_r$. The coefficient of $a_{i(\alpha)}$ is a real number and is called the category weight. Our quantification method determines $a_{i(\alpha)}$ and $y^{r(\nu)}$ values in such a way that the two groups of sequences including the PKC phosphorylation signal ($r = 1$) and sequences including no such signal ($r = 2$) can be discriminated most distinctly. This optimization can be achieved by the following procedure. First, we calculate the mean value of sample scores within the group r, $\bar{y}^r$, and the mean value of the total samples, $\bar{y}$. Then, the variance of the total samples, σ^2, and the variance between groups 1 and 2, ${\sigma_B}^2$, are given by

$$\sigma^2 = (1/N)\sum_{r=1}^{2}\sum_{\nu=1}^{n_r}(y^{r(\nu)} - \bar{y})^2 \quad (2)$$

$${\sigma_B}^2 = (1/N)\sum_{r=1}^{2} n_r(\bar{y}^r - \bar{y})^2, \quad (3)$$

where $N = n_1 + n_2$. To discriminate the sequences between groups 1 and 2 most distinctly, we maximize the ${\sigma_B}^2/\sigma^2$ value. Estimation of $a_{i(\alpha)}$ values at this optimum condition can be done by solving the eigen-value problem, as was described previously.[4] The maximum value of ${\sigma_B}^2/\sigma^2$ was calculated to be 0.848, and the estimated $a_{i(\alpha)}$ values are given in Table 2. The sample score of any 9-amino-acid sequence in the neurogranin protein is then calculated by Eq. 1 together with the $a_{i(\alpha)}$ values. Our analysis demonstrates that the higher the score of a sequence, the stronger signal of PKC phosphorylation the sequence has. In the following section, we analyze the phosphorylation signals of PKC and PKA for amino-acid sequences of peptide substrate and its amino-acid replaced derivatives in terms of such sample scores.

PKC Phosphorylation Signal

PKC is a calcium-activated, phospholipid-dependent kinase, which plays important roles in signal transduction regulation of various cellular processes.[8] A number of proteins, such as the regulatory myosin light chain, have been identified as potential PKC substrates. To define the amino-acid sequence specificity of PKC phosphorylation, several attempts have been done to use synthetic peptides derived from these proteins. As was shown in the Introduction, basic amino acids on both sides of the phosphorylated Ser/Thr constitute a recognition determinant for PKC. However, the precise determinants for potency and selectivity of PKC substrates are not well known.

Analysis of Synthetic Peptide Substrates Derived from Neuronal Neurogranin. The neuronal protein, neurogranin, is a selective substrate for PKC, but no significant phosphorylation is detected by either PKA or calcium/calmodulin-dependent protein kinase II. Chen et al.[7] synthesized a peptide, $NG_{(28–43)}$, corresponding to the phosphorylation domain of neurogranin (amino acids from position 28 to 43) and characterized its properties as PKC substrate. As shown in Table 3, they tested $NG_{(28–43)}$ and several other peptide analogues for their potency and specificity as kinase substrates to understand structural determinants involved in the phosphorylation of Ser at position 34. The high affinity and selectivity of $NG_{(28–43)}$ for PKC suggest that the specificity determinants of native neurogranin as a PKC substrate lie within amino acids 28–43 of its sequence. However, the replacement of amino-acid residues around phosphorylation site considerably changed the experimental catalytic efficiency (V_{max}/K_m), where V_{max} and K_m are derived from Michaelis–Menten's equation. For example, Table 3 shows that, in a peptide analogue of [I^{36}] $NG_{(28–43)}$, the single replacement of Arg^{36} with Ile caused a marked reduction in the catalytic efficiency, as compared to that of $NG_{(28–43)}$. On the other hand, in [R^{30}] $NG_{(28–43)}$, the replacement of Lys^{30} with Arg enhanced the efficiency significantly. It was found that the presence of basic amino acids on either side of the phosphorylated Ser/Thr contributed to the potency of the neurogranin-derived peptides, in agreement with the previous consensus sequence for PKC. However, such amino-acid sequence specificity is described only qualitatively, and we have to explain the effect of amino-acid replacement upon catalytic efficiency more quantitatively. This can be done by a quantification analysis.

Although $NG_{(28–43)}$ and its peptide analogues are composed of the neurogranin sequence from position 28 to 43, amino acids at positions 28, 29 and from 37 to 43 are common in all of the peptides. In the quantification analysis, we take the 9-amino-acid sequence from position 30 to 38. The phosphorylated Ser/Thr site is set as position 0, so that positions 30 to 38 in neurogranin change into positions -4 to $+4$, respectively. The procedure and results of the analysis of 9-amino-acid sequences in neurogranin were already described in Method of Analysis. First, we consider the estimated category weight values of $a_{i(\alpha)}$ given in Table 2. Here, the positive values of $a_{i(\alpha)}$ contribute greatly to the phosphorylation signal, whereas the negative values are unfavorable for the signal. For example, $a_{i(\alpha)}$, with item $i = 5$ and $\alpha = 16$, possesses the largest value of 21.208, and the value of 8.079 with $i = 5$ and $\alpha = 17$, the next largest, indicating that Ser^0 or Thr^0 at the phosphorylation site is essential for the signal. The next important amino acids for the signal are $i = 7$ with $\alpha = 9$ and $\alpha = 15$, which are Lys^{+2} and Arg^{+2}, respectively. At this position, category weight values are most negative with $\alpha = 2$, $\alpha = 10$, and $\alpha = 20$, indicating that Cys, Leu, and Tyr are very unfavorable for the signal. In the region from -4 to -1, Arg^{-3}, Gln^{-1}, Lys^{-2}, and Arg^{-2} are important for the signal, while Trp^{-3} and Met^{-2} are very unfavorable. We note that all amino acids with positive category weight values correspond well to amino acids in the consensus sequence described above. Our category weight values indicate the relative importance of each amino acid at each position quantitatively. Since the quantification method discriminates sequences containing the phosphorylation signal from those containing no signal most distinctly, we can tell not only what amino acid at each position is favorable for the signal, but also what amino acid is unfavorable. Considering all favorable and unfavorable contributions quantitatively, a sample score value of the 9-amino-acid sequence gives the strength of the signal contained in the sequence.

Next, we discuss the cases of $NG_{(28–43)}$ and its peptide analogues given in Table 3. A sample score of the parent 9-amino-acid sequence (KIQASFRGH) taken from -4 to $+4$ of $NG_{(28–43)}$ was calculated to be 31.87 by the use of Eq. 1 together with $a_{i(\alpha)}$ values given in Table 2. This value was found to be the greatest among all of the 9-amino-acid sequences in neurogranin, being in good agreement with the experimental finding

Table 2. The Optimum Category Weight Values of $a_{i(\alpha)}$ for PKC Phosphorylation Signal Calculated with Quantification Analysis of the Data of Table 1[a)]

Item (i)	Category (α)/Amino-acid residue				
	1/A	2/C	3/D	4/E	5/F
	6/G	7/H	8/I	9/K	10/L
	11/M	12/N	13/P	14/Q	15/R
	16/S	17/T	18/V	19/W	20/Y
1	−1.065	−3.214	2.128	−3.744	−0.114
	2.361	0.519	−3.665	−0.593	1.236
	−4.559	−1.244	1.358	−0.951	2.052
	1.719	−0.788	2.994	−1.458	−2.165
2	−0.211	−3.682	0.761	0.504	−0.831
	0.198	−2.632	−1.548	0.486	−1.332
	−4.261	−3.657	1.087	−0.477	5.769
	0.950	1.715	1.688	−7.745	−1.129
3	0.853	−2.297	−2.045	−1.674	−2.334
	−2.094	−3.647	−0.288	4.169	−0.777
	−4.852	−4.356	−2.509	2.644	3.901
	3.051	0.502	1.278	−2.757	−1.440
4	3.160	−1.092	−1.480	−3.005	0.225
	0.070	−2.813	0.328	−2.625	2.765
	−1.797	−2.510	3.113	4.795	0.366
	1.789	−0.207	0.255	−4.035	−4.688
5	−11.055	−7.914	−8.108	−9.118	−9.397
	−12.175	−8.594	−10.669	−9.223	−10.060
	−11.319	−9.984	−7.372	−10.349	−10.450
	21.208	8.079	−10.298	−5.488	−7.293
6	−2.086	−1.434	−1.809	−0.796	0.885
	−1.469	−0.879	−4.613	2.666	2.878
	−1.475	−1.058	−1.085	−0.168	1.208
	1.694	−2.109	3.801	−2.958	0.043
7	−0.112	−4.389	−1.250	−3.122	−0.913
	−1.294	−1.519	−2.587	6.985	−4.556
	1.396	−2.848	−2.403	−0.839	6.453
	−0.716	0.353	3.174	−2.386	−5.061
8	0.074	−2.601	−3.545	−0.230	1.708
	0.422	−0.924	−1.917	2.711	2.430
	−0.519	−5.083	−0.775	−2.990	3.060
	1.652	−0.265	0.846	0.115	−3.677
9	−0.239	0.613	−2.768	−2.049	−0.940
	−1.906	−0.761	2.850	1.712	−0.399
	−4.203	−1.255	2.773	−0.601	0.664
	1.909	−1.917	1.075	1.391	−0.445

a) Item number (i) specifies the position of amino-acid residue, while category number (α), the kind of amino-acid residue. See text.

that underlined Ser in KIQASFRGH is the sole phosphorylation site of neurogranin. In the other constructs given in Table 3, we also calculated sample scores of 9-amino-acid sequences, and compared them with experimental catalytic efficiencies (V_{max}/K_m). In the [R^{30}] $NG_{(28–43)}$ construct, the score increased up to 34.51 as compared to 31.87 for $NG_{(28–43)}$. In accordance with this, (V_{max}/K_m) of the [R^{30}] $NG_{(28–43)}$ construct, >30.00, is greater than 17.90 of $NG_{(28–43)}$. Both experimental data and the quantification analysis support that Arg is preferable to Lys at position −4. In the [I^{35}] $NG_{(28–43)}$ construct, however, the score decreased to 26.37, as compared to 31.87 for $NG_{(28–43)}$. In accordance with this, (V_{max}/K_m) of the [I^{35}] $NG_{(28–43)}$ construct, 0.52, is smaller than 17.90 of $NG_{(28–43)}$. Therefore, replacement of Phe^{+1} by Ile decreases catalytic efficiency. Similarly, in the [I^{36}] $NG_{(28–43)}$ construct, the score decreased further to 22.83, which explains why the (V_{max}/K_m) value of [I^{36}] $NG_{(28–43)}$, 0.017, is much smaller than 0.52 of [I^{35}] $NG_{(28–43)}$. This means that the basic residue of Arg^{+2} is very important for the phosphorylation signal. One exception lies in the analogues replaced at the +1 position. The sample

Table 3. Phosphorylation Efficiency of $NG_{(28–43)}$ and Peptide Analogues by PKC and Sample Score Calculated by Quantification Analysis

Peptide		V_{max}/K_m [a)]	Sample score [b)]
$NG_{(28–43)}$	AAKIQASFRGHMARKK	17.90	31.87
[R^{30}] $NG_{(28–43)}$	AARIQASFRGHMARKK	>30.00	34.51
[A^{35}] $NG_{(28–43)}$	AAKIQASARGHMARKK	0.003	28.90
[I^{35}] $NG_{(28–43)}$	AAKIQASIRGHMARKK	0.52	26.37
[I^{36}] $NG_{(28–43)}$	AAKIQASFIGHMARKK	0.017	22.83

a) See Ref. 7. V_{max} (μmol/min/mg) and K_m (μM). b) Score is calculated with nine-amino-acid sequence in the region from position 30 to 38. See text for further details.

Table 4. Phosphorylation Efficiency of Glycogen Synthase Peptide and Its Analogues by PKC and Sample Score Calculated by Quantification Analysis

Peptide	V_{max}/K_m [a)]	Sample score
PLSRTLSVSS	6.2×10^{-2}	26.83
PLSLTLSVSS	3.7×10^{-5}	18.15
PLSRTLSVAA	4.0×10^{-2}	27.68
PLSKTLSVAA	3.6×10^{-4}	25.03
PLRRTLSVAA	1.0×10^{-1}	32.25
PLSRRLSVAA	6.0×10^{-2}	33.18
PLSRTLTVAA	1.1×10^{-2}	20.57

a) See Ref. 9. V_{max} (μmol/min/mg) and K_m (μM).

score of the [A^{35}] $NG_{(28–43)}$ construct was calculated to be 28.90. This value is smaller than the 31.87 of $NG_{(28–43)}$, but is larger than the 26.37 of [I^{35}] $NG_{(28–43)}$. From this, the quantification analysis can explain why the (V_{max}/K_m) value of [A^{35}] $NG_{(28–43)}$, 0.003, is much smaller than 17.90 of $NG_{(28–43)}$, but fails in explaining why (V_{max}/K_m) of [A^{35}] $NG_{(28–43)}$, 0.003, is smaller than 0.52 of [I^{35}] $NG_{(28–43)}$. In view of these results, although there may be an exceptional case, the quantification analysis can explain the degree of catalytic efficiencies of phosphorylation quantitatively.

Analysis of Synthetic Peptide Substrates Derived from Glycogen Synthase. House et al.[9] studied the influence of basic residues on the substrate specificity of PKC with a series of peptide analogues of rabbit glycogen synthase, where the underlined Ser in the N-terminal peptide Pro-Leu-Ser-Arg-Thr-Leu-Ser-Val-Ser-Ser was phosphorylated. They replaced residues at various positions systematically and measured catalytic activities of phosphorylation, (V_{max}/K_m). Some of their experimental data are summarized in Table 4. First, we examine the role of Arg^{-3}. Replacement of Arg^{-3} by Leu or Lys resulted in a dramatic decrease in (V_{max}/K_m), where 6.2×10^{-2} for PLSRTLSVSS is compared to 3.7×10^{-5} for PLSLTLSVSS, and where 4.0×10^{-2} for PLSRTLSVAA is compared to 3.6×10^{-4} for PLSKTLSVAA. These data demonstrate that PKC strongly prefers the basic Arg residue at the −3 position. Concerning the influence of additional basic residues on the N-terminal side, replacement of Ser^{-4} or Thr^{-2} with Arg caused a 1.5–2.5 fold increase in the (V_{max}/K_m) value, where 1.0×10^{-1} for PLRRTLSVAA and 6.0×10^{-2} for PLSRRLSVAA are compared to 4.0×10^{-2} of PLSRTLSVAA. Next, we examined effect of replacing residues at positions +2 and +3. When (V_{max}/K_m) $= 4.0 \times 10^{-2}$ for PLSRTLSVAA is compared with 6.2×10^{-2} for PLSRTLSVSS, replacing Ala^{+2} and Ala^{+3} with Ser, little increase in the value was observed, causing no considerable effect on catalytic efficiency. Usually, PKC phosphorylates substrates at either Ser or Thr, and comparison between Ser and Thr in the glycogen synthase peptide is interesting. House et al. examined the effect of replacement of Ser^0 with Thr in PLSRTLSVAA (Table 4). This replacement does not significantly alter the V_{max}, but increase the apparent K_m, so that the total catalytic efficiency (V_{max}/K_m) decreases considerably in going from Ser to Thr.

These experimental results were again analyzed by our quantification analysis. The procedure of this analysis is similar to the preceding case of neurogranin, but in the present glycogen synthase, 10-amino-acid sequences from position −6 to +3 were analyzed, and 202 and 642 numbers of sequences were summarized into groups $r = 1$ and 2, respectively. Optimum category weight values of $a_{i(\alpha)}$ were again calculated, and a sample score of each 10-amino-acid sequence was estimated by an equation similar to Eq. 1. First, we examine the role of Arg^{-3}. Replacement of Arg^{-3} with Leu resulted in a 2000-fold decrease in (V_{max}/K_m), indicating that PKC strongly prefers the basic Arg residue at the −3 position. This experimental finding is well explained by our quantification analysis. As is shown in Table 4, the sample score of PLSRTLSVSS (26.83) is much larger than 18.15 of PLSLTLSVSS, explaining that replacement of Arg^{-3} by Leu causes a dramatic decrease in the catalytic activity. In a similar way, the score of PLSRTLSVAA (27.68) is a little larger than 25.03 of PLSKTLSVAA. In accordance with this, replacement of Arg^{-3} with Lys resulted in 200-fold decrease in (V_{max}/K_m), suggesting that Arg is the most effective amino acid at position −3, Lys is the next effective, and Leu is the least. As for the residue at the −4 position, if Ser is replaced by Arg, both (V_{max}/K_m) and sample score increase remarkably, indicating that basic residue Arg is very effective at this position. This is shown by comparison of (V_{max}/K_m) $= 4.0 \times 10^{-2}$ and the score 27.68 of PLSRTLSVAA with 1.0×10^{-1} and 32.25 of PLRRTLSVAA, respectively. The basic residue Arg^{-2} is also important for the signal. This is shown by comparing (V_{max}/K_m) $= 4.0 \times 10^{-2}$ and the score 27.68 of PLSRTLSVAA with 6.0×10^{-2} and 33.18 of PLSRRLSVAA, respectively, where replacement of Thr^{-2} with Arg increases both catalytic activity and the sample score. However, there lies a problem which cannot explain the magnitudes of catalytic activity by the sample score. Although both (V_{max}/K_m) and the score of PLRRTLSVAA and PLSRRLSVAA are larger than those of PLSRTLSVAA, (V_{max}/K_m) of PLRRTLSVAA is larger than that of PLSRRLSVAA, while the score of PLRRTLSVAA is a little

smaller than that of PLSRTLSVAA. Primary sequence data alone are not sufficient to explain this, and the secondary/tertiary structure may play a role in the catalytic activity. As for the residues at positions +2 and +3, we compare $(V_{max}/K_m) = 6.2 \times 10^{-2}$ and the sample score 26.83 of PLSRTLSVSS with 4.0×10^{-2} and 27.68 of PLSRTLSVAA, respectively. Although replacing both Ser^{+2} and Ser^{+3} with Ala gave a slight increase in (V_{max}/K_m) and a slight decrease in the sample score, such replacements cause no significant changes in either catalytic activity or sample score.

PKC phosphorylates Ser or Thr at position 0. Comparison of $(V_{max}/K_m) = 4.0 \times 10^{-2}$ of PLSRTLSVAA with 1.1×10^{-2} of PLSRTLTVAA shows that phosphorylation for Ser is more effective than Thr (Table 4). This experimental finding is also supported by the quantification analysis, where the sample score of 27.68 of PLSRTLSVAA is much greater than 20.57 of PLSRTLTVAA.

PKA Phosphorylation Signal

cAMP-dependent protein kinase (PKA) and cGMP-dependent kinase exhibit a number of similar physical and functional properties and may be homologous proteins. In the absence of cAMP, PKA is sequestered as an inactive tetrameric complex composed of both regulatory (R) and catalytic (C) subunits.[10,11] Upon binding of cAMP, the tetrameric complex (R_2C_2) dissociates into an R_2–$cAMP_4$ dimer and two active, monomeric C subunits. The catalytic core of phosphorylation is contained in the C subunit, and it has been shown that the other eukaryotic protein kinases share such a conserved catalytic core. Although their protein substrate specificities are also similar to one another, PKA favors only basic residues in the −2 and −3 regions, and its consensus sequence is given by R-R/K-X-S/T.[3]

Phosphorylation of Synthetic Peptide Analogs of Histone H2B by PKA. Glass and Krebs[12] used a synthetic heptapeptide Arg-Lys-Arg-Ser-Arg-Lys-Glu (RKRSRKE) corresponding to residues 29–35 in histone H2B as a model substrate, and studied its phosphorylation by PKA (phosphorylated Ser is underlined). It is known that the basic residues in the N-terminal region to Ser are important, but that basic residues in the C-terminal region to Ser have a negative influence on this enzyme. To confirm this observation, Glass and Krebs prepared a RKRSRKE peptide together with its analogues by introducing systematic replacement of amino-acid residues at various positions, and examined their catalytic efficiency of phosphorylation by PKA. Some of their experimental data are reproduced in Table 5. Replacement of basic residues in the parent peptide greatly altered the kinetic parameters for PKA. The importance of basic residues in the N-terminal region to Ser was shown by the discovery that replacing Arg^{-3} with Ala and Lys^{-2} with Ala remarkably decreased the (V_{max}/K_m) values of phosphorylation, as compared to that of the parent peptide, while the negative influence of basic residues located in the C-terminal region to Ser was demonstrated by the replacement of Lys^{+2} with Ala, which remarkably increased the (V_{max}/K_m) value. As for the residues at positions −1, +1, and +3, if Arg^{-1} is replaced by Ala, the (V_{max}/K_m) value becomes 4-fold as much as that of the parent peptide. A small decrease in the (V_{max}/K_m) value was found when Arg^{+1} or Glu^{+3} was replaced by Ala. On the other hand, PKA is known to phosphorylate not only Ser but also Thr. In place of RKRSRKE, Glass and Krebs examined RKRTRKE, and its phosphorylation efficiency, (V_{max}/K_m), was much less than that of RKRSRKE.

Table 5. Phosphorylation Efficiency of Histone H2B Peptide and Its Analogues by PKA and Sample Score Calculated by Quantification Analysis

Peptide	V_{max}/K_m [a)]	Sample score
RKRSRKE	0.0048	41.47
AKRSRKE	0.0001	24.51
RARSRKE	0.0001	31.07
RKASRKE	0.0185	37.72
RKRSAKE	0.0029	41.30
RKRSRAE	0.1089	43.20
RKRSAAE	0.0348	43.03
RKRSRKA	0.0038	41.53
RKRTRKE	0.0003	18.75

a) See Ref. 12. V_{max} (μmol/min/mg) and K_m (μM).

These experimental results were also analyzed by our quantification method, which was done in a way similar to that of PKC. Sequence data composed of 7-amino-acid residues (from the −3 to +3 positions with respect to Ser at position 0) were taken to analyze the phosphorylation signal of PKA. For group 1 ($r = 1$), we collected 167 sample sequences including the signal, as compiled by Kreegipuu.[2] For group 2 ($r = 2$), we summarized 616 samples (116 from histone H2B and 500 random sequences), which include no such signal. Quantification analysis was performed for the data set. Calculated category weight values of $a_{i(\alpha)}$ clearly show that the basic residues of Arg and Lys in the N-terminal region have a strong positive effect on the signal, while those in the C-terminal region are unfavorable. Next, we compared calculated sample scores of RKRSRKE and its analogues with the experimental data. The results are summarized in Table 5.

The value of (V_{max}/K_m) and the sample score for PKA phosphorylation in the parent RKRSRKE are 0.0048 and 41.47, respectively. However, if Arg^{-3} is replaced by Ala, both (V_{max}/K_m) and the sample score decrease dramatically to 0.0001 and 24.51, respectively. If Lys^{-2} is replaced by Ala, (V_{max}/K_m) and the sample score decrease to 0.0001 and 31.07, respectively. Therefore, the quantification analysis strongly supports the experimental finding that the basic residues of Arg^{-3} and Lys^{-2} are very favorable for phosphorylation. As for the negative influence of basic residues located in the C-terminal region to Ser, replacing Lys^{+2} with Ala, (V_{max}/K_m) and the sample score increase dramatically to 0.1089 and 43.20, as compared to 0.0048 and 41.47 of the parent peptide, respectively. If both Arg^{+1} and Lys^{+2} are replaced by Ala, (V_{max}/K_m) and the sample score also increase to 0.0348 and 43.03, respectively, which are not greater than those of the single replacement of Lys^{+2} by Ala. Moreover, if only Arg^{+1} is replaced by Ala, (V_{max}/K_m) and the sample score decrease slightly to 0.0029 and 41.30, respectively. Both experiments and the quantification analysis clearly show that a basic residue at position +2 plays definitely negative role in PKA phosphorylation, but that a basic residue at position +1 is slightly preferable. Replacement of Glu^{+3} by Ala gave values of $(V_{max}/K_m) = 0.0038$ and a sample score = 41.53. As compared to the values of the parent peptide, (V_{max}/K_m) decreases slight-

ly, while the sample score increases slightly. Although the change in (V_{max}/K_m) cannot be explained by the sample score, such a difference is a minor one, and the replacement of Glu^{+3} with Ala does not change the catalytic activity significantly. Replacement of Arg^{-1} with Ala gave values of (V_{max}/K_m) = 0.0185 and a sample score = 37.72. As compared to the values of the parent peptide, the replacement causes a 4-fold increase in (V_{max}/K_m), while the sample score decreases. The change in (V_{max}/K_m) was not explained by the sample score. Quantification analysis, assuming an independent site model, may not be enough for elucidation, because catalytic efficiency is sensitive to the −1 position nearest the phosphorylation site of Ser through interaction between sites. However, there lies another possibility that the data of sequences, including phosphorylation signal (group 1), may be insufficient to elucidate. After further compilation of signal sequences, quantification analysis may explain the behavior at the −1 position.

If the phosphorylation site of Ser is replaced by Thr, (V_{max}/K_m) = 0.0003 and the sample score = 18.75 were obtained for RKRTRKE, which can be compared with (V_{max}/K_m) = 0.0048 and sample score = 41.47 for RKRSRKE. Although PKA phosphorylation occurs with Thr, its catalytic efficiency decreases drastically as compared to Ser. This was supported by both experimental data and quantification analysis.

Phosphorylation of Rat 6-Phosphofructo-2-kinase by PKA. The amino-acid sequence surrounding the phosphorylation site in 6-phosphofructo-2-kinase has been shown to be Val-Leu-Gln-Arg-Arg-Arg-Gly-Ser-Ser-Ile-Pro-Gln. It is characteristic of substrates of PKA to have multiple Arg residues in the N-terminal region to a phosphorylated Ser (underlined). Glass et al.[13] synthesized a VLQRRRGSSIPQ peptide and its derivatives, and measured kinetic constants of catalytic phosphorylation efficiency by PKA. Their experimental data are summarized in Table 6.

The parent VLQRRRGSSIPQ peptide has a (V_{max}/K_m) value as high as 3.58. First, we examine cases where Ser is phosphorylated. In an analogue of VLQARRGSSIPQ, where Arg^{-4} is replaced by Ala, a 5-fold decrease of (V_{max}/K_m) was observed as compared to the value of the parent peptide. If Gly^{-1} is replaced by Pro, a 3-fold decrease in (V_{max}/K_m) was observed. If Thr is phosphorylated in place of Ser, more than 10-fold decrease of (V_{max}/K_m) was observed with VLQARRGTSIPQ, as compared to the parent VLQARRGSSIPQ. In the peptides where Thr is phosphorylated, replacement of Arg^{-4} with Ala caused a 3-fold decrease in (V_{max}/K_m) as compared to VLQRRRGTSIPQ. If Gly^{-1} is replaced by Pro, more than a 2-fold decrease of (V_{max}/K_m) is observed.

Table 6. Phosphorylation Efficiency of 6-Phosphofructo-2-kinase Peptide and Its Analogues by PKA and Sample Score Calculated by Quantification Analysis

Peptide	V_{max}/K_m [a)]	Sample score
QRRRGSSIPQ	3.58	49.37
QARRGSSIPQ	0.74	47.15
QRRRPSSIPQ	1.17	50.43
QRRRGTSIPQ	0.14	26.64
QARRGTSIPQ	0.05	24.42
QRRRPTSIPQ	0.06	27.70

a) See Ref. 13. V_{max} (μmol/min/mg) and K_m (μM).

These experimental results are also explained by our quantification analysis. Although Glass et al. studied 12-amino-acid peptides, residues of V and L at positions −7 and −6 are common to all peptide analogues, and they were neglected in our analysis. Sequence data composed of 10-amino-acid residues (from −5 to +4 positions with respect to Ser at position 0) were then taken to analyze phosphorylation signal of PKA. For group 1, we collected 164 sample sequences including the signal, as compiled by Kreegipuu.[2] For group 2, 1460 samples including no signal were summarized. Quantification analysis was done with such a data set. Next, calculated sample scores of QRRRGSSIPQ and its analogues were compared with the experimental data, and the results are summarized in Table 6.

The sample score of QRRRGSSIPQ (49.37) is the largest among the 461 samples of 10-amino-acid sequences composed of rat 6-phosphofructo-2-kinase. Another sequence, RRRGSSIPQF, whose phosphorylated Ser lies one-residue forwards the C-terminal direction, has a score of 30.84. Its value is much smaller than the 49.37 of QRRRGSSIPQ, being in agreement with the experimental finding that only Ser of QRRRGSSIPQ is phosphorylayted. For this QRRRGSSIPQ, if Arg^{-4} is replaced by Ala, the sample score decreases to 47.15. This corresponds to the 5-fold decrease in (V_{max}/K_m) as compared to the value (3.58) of the parent peptide, and the basic residue at the −4 position is preferable for PKA phosphorylation. On the other hand, replacement of Gly^{-1} with Pro increases the sample score to 50.43, but causes a 3-fold decrease in (V_{max}/K_m). In this respect, the sample score does not explain the tendency of (V_{max}/K_m). Such discrepancy was also found in the previous histone H2B, where replacement of Arg^{-1} with Ala caused a 4-fold increase in (V_{max}/K_m), while the sample score decreased. The change in (V_{max}/K_m) is not explained by the sample score, and the catalytic efficiency appears to be sensitive to the −1 position nearest the phosphorylation site of Ser (refer to the discussion on the −1 position in the preceeding histone H2B).

The catalytic efficiency for the phosphorylation of Ser, (V_{max}/K_m) = 3.58, in QRRRGSSIPQ decreases to 0.14 if the phosphorylation site is replaced by Thr. In accordance with this, the sample score (49.37) of QRRRGSSIPQ decreases to 26.64 for QRRRGTSIPQ. As for the phosphrylation of Thr, replacement of Arg^{-4} with Ala decreases both the sample score to 24.42 and (V_{max}/K_m) to 0.05. However, replacement of Gly^{-1} with Pro increases the sample score to 27.70, but decreases (V_{max}/K_m) to 0.06. Clearly, replacement effects at positions −4 and −1 on QRRRGTSIPQ are very similar to those in QRRRGSSIPQ.

Concluding Remarks

So far, the consensus sequence and sequence motif are derived from amino-acid sequences in functionally important regions, which are strongly conserved in short segments of amino-acid sequences. In such cases, only favorable amino acids are collected, while unfavorable amino acids, which have a negative effect on the signal, are not taken into account. Our quantification analysis requires two sequence data sets with

and without functional signal, and discriminates them most distinctly. Therefore, this approach can tell us not only what amino acid at each position is favorable for the signal, but also what amino acid is unfavorable. Considering all favorable and unfavorable contributions quantitatively, the sample score value gives the strength of the signal. To show the advantage of our method, we discuss phosphorylation signals of PKC and PKA. Both phosphorylate Ser or Thr. The consensus sequences show that PKC strongly favors both N- and C-terminal basic residues with respect to the phosphorylation site, while only N-terminal basic residues are required for PKA. However, the consensus sequences cannot explain why PKA does not recognize the PKC recognition site. Our quantification analysis explains this, and supports different phosphorylation sites recognized by PKC and PKA, because C-terminal basic residues are strongly unfavorable for PKA.

Next, we discuss the sampling problem of random sequences to construct the data set of group 2. As discussed in a previous DNA analysis of splice signals of human β-globin pre-mRNA,[4] there are only four kinds of nucleotides at each position of the sequence, and they occur almost equally. Moreover, the β-globin pre-mRNA sequence is long enough, and gave 1596 samples to group 2. These situations provide a reasonable basis for application of the quantification analysis. However, analysis of amino-acid sequences does not fulfill such a condition. For example, the sequence of neurogranin is composed of only 76 amino-acid residues. Residues do not occur equally, and some residues, such as Trp, occur very rarely. To apply our quantification analysis, we have to have a sufficient number of sample sequences in group 2, including no signal. For this purpose, random sequences were constructed and added to group 2, but several problems may arise in this treatment. One is that one or two sequences may happen to coincide with the signal sequence belonging to group 1. However, this mismatch does not have a serious influence on the values of category weights and sample scores, as long as group 2 is provided with a sufficient number of sample sequences. Another problem is that the calculated category weights and sample scores may depend on the sample number of group 2 generated by random sequences. To discuss this problem, we again consider the previous case of neurogranin phosphorylated by PKC, where group 2 consists of 67 samples coming from neurogranin and 500 generated by random sequences. For another case, we further did the analysis on the data of group 2 consisting of 67 samples coming from neurogranin and 300 generated by random sequences, keeping the same data of group 1. Although absolute values of category weights are different, no appreciable difference in relative magnitudes of category weights was found between the two cases. It appears that, if we provide a sufficient number (about 5–8 times the number of neurogranin sequences) of random sequences as group 2, our quantification analysis will reach almost the same conclusion.

Next, we discuss the neural network approach for analysis of protein sequence motifs. The present quantification analysis assumes an independent site model, where optimum category weight is given to each amino-acid residue at each position, and where interaction between residues at different positions is neglected. Rigorously speaking, this assumption is too simple, because there may lie interactions between residues through the secondary/tertiary structure of the protein. The neural network takes such kind of interactions into consideration, but requires too many parameters to be determined.[1] Datasets studied in the present work are not sufficient to describe all of the parameters precisely. In addition, experimental results on the structure analysis of an inhibitor peptide co-crystallized with the C subunit of PKA revealed that the substrate peptide in the consensus recognition region from -3 to $+3$ forms an extended conformation, but that, in the N-terminal region from -16 to -4, the substrate forms an α-helix and turn structure.[10] Therefore, as long as the sequence in the consensus region is concerned, the phosphorylation signal is mostly given by the primary structure of the amino-acid sequence. Moreover, experimental data on the replacement effect of amino-acid residues on the catalytic efficiencies were almost explained by the independent site model. These findings strongly justify the application of quantification analysis to the phosphorylation signals of PKC and PKA, but few of the experimental data still remain to be elucidated. The neural network may be effective for such a problem.

Finally, the present paper reported a quantification analysis to study the sequence motifs of the phosphorylation signals of PKC and PKA. It is to be noted that this approach is generally applicable to analyze other functional signals of proteins, as long as their sequence motifs are mostly expressed in terms of the primary structure of amino-acid sequences. However, if a functional signal is given by structural motif, such as helix–turn–helix, or by secondary/tertiary structure, it will be very difficult for the quantification method to analyze such a signal.

References

1 See, for example; D. W. Mount, "Bioinformatics: Sequence and Genome Analysis," Cold Spring Harbor Laboratory Press, New York (2001).

2 A. Kreegipuu, N. Blom, S. Brunak, and J. Jarv, *FEBS Lett.*, **430**, 45 (1998). See also internet home page (http://www.cbs.dtu.dk/database/PhosphoBase/).

3 P. J. Kennelly and E. G. Krebs, *J. Biol. Chem.*, **266**, 15555 (1991).

4 Y. Iida, *Comput. Appl. Biosci.*, **3**, 93 (1987).

5 Y. Iida and T. Masuda, *Nucleic Acids Res.*, **24**, 3313 (1996).

6 Y. Iida and D. Kanagu, *Bull. Chem. Soc. Jpn.*, **76**, 913 (2003).

7 S.-J. Chen, E. Klann, M. C. Gower, C. M. Powell, J. S. Sessoms, and J. D. Sweatt, *Biochemistry*, **32**, 1032 (1993).

8 Y. Nishizuka, *Nature (London)*, **334**, 661 (1988).

9 C. House, R. E. H. Wettenhall, and B. E. Kemp, *J. Biol. Chem.*, **262**, 772 (1987).

10 D. R. Knighton, J. Zheng, L. F. Ten Eyck, N.-H. Xuong, S. S. Taylor, and J. M. Sowadski, *Science*, **253**, 414 (1991).

11 J. Zheng, D. R. Knighton, L. F. Ten Eyck, R. Karlsson, N.-H. Xuong, S. S. Taylor, and J. M. Sowadski, *Biochemistry*, **32**, 2154 (1993).

12 D. B. Glass and E. G. Krebs, *J. Biol. Chem.*, **257**, 1196 (1982).

13 D. B. Glass, M. R. El-Maghrabi, and S. J. Pilkis, *J. Biol. Chem.*, **261**, 2987 (1986).